jipitec

2| 2026

Volume 17(2026)
Issue 2 ISSN 2190-3387

Issue 17 (2)

Editorial
by Orla Lynskey

Articles

Regulating AI: A Matrix for Gauging Impact and Its Legal Implementation
by Maurice Chiodo, Henning Grosse Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera and Herbert Zech

The Metaverse as a Privacy Mosaic: Framing Data Protection in Hyper-Granular, Data-Drenched Digital Ecosystems
by Pablo Trigo Kramcsák and Vagelis Papakonstantinou

The Mythical Code-Authorship
by Ewa Laskowska-Litak

The Code as an Element of Trust in Smart Contract Arbitration Clauses
by Eduardo Andrés Calderón-Marenco, Tania Rodríguez Palacios, Jaime Elias Torres Buelvas and Piedad Barreto-Granada

Social Media Bans and the Digital Services Act: A Critical Analysis of Statement of Reasons and Online Dispute Resolution
by A.R. Lodder and J. Zeleznikow

The Concept of Lawful Access and Its Implications for User Freedoms in the EU Copyright Acquis: Opinion of the European Copyright Society
by Thomas Margoni, Martin Kretschmer, Axel Metzger, João Pedro Quintais and Tatiana Synodinou

Editors:

Thomas Dreier
Séverine Dusollier
Lucie Guibault
Orla Lynskey
Thomas Margoni
Axel Metzger
Miquel Peguera Poch
Karin Sein
Gerald Spindler (†)

Journal of Intellectual Property, Information Technology, and Electronic Commerce Law

www.jipitec.eu

2026

jipitec

Journal of Intellectual Property, Information Technology and Electronic Commerce Law

Volume 17 Issue 2 October 2026
www.jipitec.eu
contact@jipitec.eu

A joint publication of:

Prof. Dr. Thomas Dreier, M. C. J. (NYU)
KIT - Karlsruher Institut für Technologie, Zentrum für Angewandte Rechtswissenschaft (ZAR), Vincenz-Prießnitz-Str. 3, 76131 Karlsruhe Germany

Prof. Dr. Axel Metzger, LL. M. (Harvard)
Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin

Prof. Dr. Gerald Spindler (†)
Dipl.-Ökonom, Georg-August-Universität Göttingen, Platz der Göttinger Sieben 6, 37073 Göttingen

Karlsruhe Institute of Technology, Humboldt-Universität zu Berlin and Georg-August-Universität Göttingen are corporations under public law, and represented by their respective presidents.

Editors:
Thomas Dreier
Séverine Dusollier
Lucie Guibault
Orla Lynskey
Thomas Margoni
Axel Metzger
Miquel Peguera Poch
Karin Sein
Gerald Spindler (†)

Board of Correspondents:
Graeme Dinwoodie
Christophe Geiger
Ejan Mackaay
Rita Matulionyte
Giovanni M. Riccio
Cyrill P. Rigamonti
Olav Torvund
Mikko Välimäki
Rolf H. Weber
Andreas Wiebe
Raquel Xalabarder

Editor-in-charge for this issue:
Orla Lynskey

Editorial Coordinator:
Lars Flamme

ISSN 2190-3387

Funded by
DGRI Deutsche Gesellschaft für Recht und Informatik e.V.

Table Of Contents

Editorial

by Orla Lynskey 95

Articles

Regulating AI: A Matrix for Gauging Impact and Its Legal Implementation
by Maurice Chiodo, Henning Grosse Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera and Herbert Zech 97

The Metaverse as a Privacy Mosaic: Framing Data Protection in Hyper-Granular, Data-Drenched Digital Ecosystems
by Pablo Trigo Kramcsák and Vagelis Papakonstantinou 141

The Mythical Code-Authorship
by Ewa Laskowska-Litak 161

The Code as an Element of Trust in Smart Contract Arbitration Clauses
by Eduardo Andrés Calderón-Marenco, Tania Rodríguez Palacios, Jaime Elias Torres Buelvas and Piedad Barreto-Granada 176

Social Media Bans and the Digital Services Act: A Critical Analysis of Statement of Reasons and Online Dispute Resolution
by A.R. Lodder and J. Zeleznikow 192

The Concept of Lawful Access and Its Implications for User Freedoms in the EU Copyright Acquis: Opinion of the European Copyright Society
by Thomas Margoni, Martin Kretschmer, Axel Metzger, João Pedro Quintais and Tatiana Synodinou 203

Editorial

by **Orla Lynskey ***

© 2026 Orla Lynskey

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: Orla Lynskey, Editorial, 17 (2026) JIPITEC 95 para 1.

1 This JIPITEC volume has something for everyone interested in digital regulation, with the featured articles covering key aspects of the legal agenda including new approaches to AI regulation, data protection in data-saturated environments, copyright in contested contexts, platform and content regulation, solutions to smart contracting dilemmas and reflections on user rights in relation to content moderation.

2 Kicking off the issue with a thought-provoking cross-cutting framework for AI regulation is an article entitled 'Regulating AI: A Matrix for Gauging Impact and its Legal Implementation' authored by Maurice Chiodo, Henning Grosse Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera and Herbert Zech. The piece urges us to think about AI regulation from a distinct and complementary vantage point to the existing regulatory provisions in legislation like the AI Act and the Digital Services Act. To this end, it identifies a new avenue for AI regulation centred around a matrix with two key parameters: the impact of the AI system (understood loosely as the number of people affected by it) and the size of the team involved in its development. Key observations follow from this: that team and AI size are often correlated but can sometimes differ and that, generally, the most risky AI potentially comes from 'Small Team-Big AI' while the safest comes from 'Big Team-Small AI'. Armed with this matrix, the authors propose a two-tiered approach to AI regulation based on a form of strict liability for the harms caused by Small AI and a more intensive form of regulatory oversight through constant monitoring and cumulative penalty points for Big AI. We are encouraged to think beyond existential AI risk and to conceive of it also as 'death by 1000 cuts'. In this vein, legal regulatory frameworks that overlook the risks in the 2x2 matrix proposed risk gaps through which AI harms will slip.

3 The second article featured brings us to data protection law. In their article entitled 'The Metaverse as a Privacy Mosaic: Framing Data Protection in Hyper-Granular, Data-Drenched Digital Ecosystems' Pablo Trigo Kramcsák and Vagelis Papakonstantinou conduct a forensic examination of the role of data protection in 'the Metaverse'. The Metaverse, the authors acknowledge, 'resists straightforward definition'. They therefore sketch its contours by identifying key features and technical attributes, and distinguish between centralised and decentralised metaverse platforms and distinct forms of data processed. The authors argue that the normative commitments found in the GDPR are difficult to translate into operational safeguards in the context of such granular data processing at scale. The frictions that emerge when data protection law is applied in a data-intensive networked system such as the Metaverse are highlighted, supporting the claim the authors advance: that we must critically reappraise how to preserve privacy and autonomy in such contexts. While not without its challenges, the solution, the authors suggest, lies in the legal requirement of data protection-by-design.

4 The third article of the issue is authored by Ewa Laskowska-Litak. In 'The Mythical Code-Authorship' she revisits the legal treatment of authorship for copyright purposes in the context of software. It takes aim at what Laskowska-Litak considers to be the growing mismatch between theory and practice: the foundational assumptions underpinning copyright protection of software and the nature of modern software production. Drawing on interdisciplinary insights, the author argues that software

production is inherently collective, collaborative and iterative which has been overlooked as it has been incorporated into copyright law. The paper proposes that rather than approaching copyright in this context through the solitary-author model, it should be understood more technically through 'programmatic attribution'. This refocus would, from the author's perspective, better align legal doctrine with socio-technical realities.

5 The fourth article of the issue by Jaime Torres Buelvas and his co-authors is entitled 'The Code as an Element of Trust in Smart Contract Arbitration Clauses'. In this piece, they examine the doctrinal tensions and structural limitations in the interactions between smart contracts and arbitration. They highlight that the features of smart contracts do not immunise them from contractual disputes. In the event of such disputes concerning the contractual terms encoded in smart contracts, arbitration clauses play a critical role. They argue that given the significance of such arbitration clauses, including in digital contexts, a standardized framework codifying arbitration clauses must be adopted in order to provide confidence to contracting parties and to enhance the validity of such clauses in the event of conflict. Their work proposes such a codified arbitration clause model.

6 Last but not least, in 'Social media bans and the Digital Services Act: A critical analysis of Statement of Reasons and Online Dispute Resolution' by A.R. Lodder and J. Zeleznikow, the authors consider how the EU's Digital Services Act (DSA) regulates the 'banning' of individuals from social media platforms which they define as the suspension or termination of user accounts. They examine whether the DSA provides users with the necessary tools to challenge such bans. Their focus is on user *transparency* (through the 'Statement of Reasons' provided for by Article 17 DSA) and user *redress* (through Article 20 DSA which requires regulated entities to put in place internal complaint-handling systems and Article 21 DSA which establishes out-of-court dispute resolution requirements). They suggest that while the statement of reasons might provide some valuable information to enable individuals to challenge such 'bans', the internal complaint-handling system may add friction to such challenges. They therefore advocate for the use of out-of-court dispute settlement as a more effective alternative. As they hint, time – and accurate empirical insights into the functioning of these mechanisms – are needed to test this hypothesis.

7 Happy reading to one and all!

Regulating AI: A Matrix for Gauging Impact and its Legal Implementation

by **Maurice Chiodo, Henning Grosse Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera, Herbert Zech ***

Abstract: This paper presents new and alternative avenues for regulating AI, understood in the broad sense of state intervention that directly or indirectly affects how AI is developed and deployed. In doing so, it also identifies two new regulatory axes to be addressed - Big/Small AI Impact, and Big/Small AI Teams - and the challenges associated with designing tools that are effective and tailored to these axes. Against this background, our proposals in this paper thus fall into two broad categories:

1. Forms of strict liability, as a means to incentivise operators of Small AI systems (where direct regulation is unlikely to be effective, and often unnecessary) to take steps preventing or at least minimising harm resulting from their AI.

2. A mechanism of constant monitoring through random sampling with human testing and determination, as well as associated cumulative "penalty points", to control the impact of Big AI systems on individuals and society before they cause large-scale, and in particular systematic, harm to society.

Our proposals aim to have the maximal possible impact on the behaviour of AI operators and on where and how AI is used, so as to prevent harm and encourage responsible development and use of AI. In our view, these proposals are implementable, policeable, enforceable, cost-effective, and technologically-neutral; they do not focus on what the AI is, but rather what it does. We consider different legal tools and mechanisms for implementing our proposals, and conclude that while elements of existing tort liability regimes, platform regulation as well as the EU AI Act can partially be relied on, law makers should consider adopting more tailored rules on both liability and constant monitoring.

Keywords: AI regulation, tort liability, EU AI Act, AI Oversight, Digital Services Act

© 2026 Maurice Chiodo, Henning Grosse Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera, Herbert Zech

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: Maurice Chiodo, Henning Grosse, Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera, Herbert Zech, Regulating AI: A Matrix for Gauging Impact and its Legal Implementation, 17 (2026) JIPITEC 97 para 1.

A. Overview of the Problem

I. The Challenge of Regulating AI

1 It is well understood that the control and regulation of artificial intelligence[1] (AI) is both an important

* Authors are listed alphabetically by surname.

Maurice Chiodo: Centre for the Study of Existential Risk, University of Cambridge, UK. mcc56@cam.ac.uk

Henning Grosse Ruse-Khan: Faculty of Law, University of Cambridge, UK. hmg35@cam.ac.uk

Dennis Müller: Institute of Mathematics Education, University of Cologne, Cologne, Germany. dennis.mueller@uni-koeln.de

Lea Ossmann-Magiera: YPOG, Weizenbaum Institute for the Networked Society, Berlin, Germany. lea.ossmann-magiera@ypog.law

Herbert Zech: Humboldt-University of Berlin, Weizenbaum Institute for the Networked Society, Berlin, Germany. herbert.zech@hu-berlin.de

The authors wish to thank the attendees of the Weizenbaum Institute event in November 2023 and those of the International Law and Technology Governance class at NYU in March 2024 for their comments and suggestions, in particular Rochelle Dreyfuss. All errors remain the sole responsibility of the authors.

1 Understood as the turn from pre-programmed software

matter to address, and one for which no satisfactory mechanism exists yet. This is unsurprising, given how complicated such systems are. Indeed, an early 2024 editorial in Nature[2] mentions the EU AI Act has "gaps" that to resolve requires input from the scientific community. Due to "implementation challenges" (like a lack of harmonised standards for the AI Act's high-risk requirements) the European Commission also proposed to postpone the application of the AI Act (which was originally scheduled for 2 August 2026), especially for high-risk AI systems.[3]

2 Lawmaking is often reactive, occurring after a problem(s) arises in society - with an implicit expectation that new problems only manifest a few times before the law is adapted and catches up. However, current AI development is happening at an increasingly rapid pace and scale, such that ongoing attempts to adjust existing laws cannot keep up. Given the AI risks we discuss below, we suggest regulatory approaches are needed that can curtail future harms posed by AI *before their manifestations, often in the form of irreversible damage to individuals, groups and societies.*

3 The following examples illustrate these difficulties, showing the ever-increasing diversity of application domains, types of AI algorithms, and the multitude of ethical and legal challenges involved. These examples also foreshadow our suggestions to regulate AI by *team size* and *impact size* (in both cases counting the number of people involved/affected).

An example of an AI with a big impact, run by a small team:

4 Stability.ai is an AI text-to-image generating platform that produces images according to prompts from the user. It has been accused by Getty Images of significant copyright violations, allegedly using millions of Getty's images for AI training without adequate permission, and has been sued by Getty for over $1.8 trillion in damages.[4] While the English High Court rejected most of Getty's claims in its first instance decision in November 2025,[5] Getty subsequently won the right to appeal in the UK.[6] And as well as potentially infringing on the IP rights of countless artists,[7] Stability.ai (and similar generative AI tools) also run the risk of sending many creators and businesses in the creative industries out of business, by fundamentally shifting the process of visual content creation. Stability.ai impacted a "big" number of people in a relatively short timespan,[8] and

to machine learning. For a more rigorous definition, see the OECD definition in Russell et al., *Updates to the OECD's definition of an AI system explained*, OECD (Nov. 29, 2023), https://oecd.ai/en/wonk/ai-system-definition-update (last accessed Feb. 29, 2024).

2 *There are holes in Europe's AI Act — and researchers can help to fill them,* Editorial, Nature **625**, 216 (Jan. 10, 2024), https://www.nature.com/articles/d41586-024-00029-4 (last accessed Feb. 29, 2024).

3 Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI), COM(2025) 836 final, art. 1 sec. 31 lit. a. The reasons for the proposal are described in the explanatory memorandum, p. 1-3 of the proposal.

4 See Growcoot, *Getty Images is Suing Stable Diffusion for a Staggering $1.8 Trillion,* PetaPixel (Feb. 7, 2023), https://petapixel.com/2023/02/07/getty-images-are-suing-stable-diffusion-for-a-staggering-1-8-trillion/ (last accessed Feb. 29, 2024). Next to the lawsuit in the UK discussed below, Getty is also pursuing claims against Stability.ai in U.S. courts which, at the time of writing, are ongoing - see Getty's filings at the District Court of Northern California at https://www.courtlistener.com/docket/71112094/getty-images-us-inc-v-stability-ai-ltd/. See also the legal filing at GMG Editorial, *Getty Lawsuit Against Stable Diffusion*, Scribd (Mar. 2, 2023), https://www.scribd.com/document/624165761/Getty-Lawsuit-Against-Stable-Diffusion (last accessed Jan. 4, 2026).

5 *Getty Images v Stability AI,* [2025] EWHC 2863 (Ch), 4 November 2025. Already during the proceedings, Getty had to drop its principal claims for copyright infringement as none of the relevant copying activities related to training Stability.ai's Stable Diffusion model could be shown to have taken place within the jurisdiction of English courts. And while Getty succeeded with some elements of its trademark infringement claims (based on the memorisation and reproduction of the Getty Image logo by output generated when using Stability.ai's model), the most relevant remaining copyright claim (based on the argument that the image generation service offered by Stability.ai's Stable Diffusion model constituted an infringing copy imported into the UK under section 22 & 23 CDPA) has been rejected.

6 The Court of Appeal allowed Getty to challenge the High Court's finding on whether the Stable Diffusion AI model can be considered an infringing copy - see Barker, *High Court grants Getty permission to appeal copyright aspects of GenAI claim* (Dec. 19, 2025) - online at: https://thelens.slaughterandmay.com/post/102lydj/high-court-grants-getty-permission-to-appeal-copyright-aspects-of-gen-ai-claim (last accessed Jan. 5, 2026).

7 For further discussion about generative AI and copyright law, see for example Zirpoli, *Generative Artificial Intelligence and Copyright Law*, Congressional Research Service (Feb. 24, 2023), https://digitalcommons.unl.edu/scholcom/243 (last accessed Feb. 29, 2024).

8 Stability.ai, https://stability.ai/about (last accessed Feb. 29, 2024).

moreover was a small company, that at the time had few employees.[9]

An example of an AI with a small impact, run by a small team:

5 Clearview AI is a company that produces AI-powered facial recognition software. They claim to have scraped the Internet, including social networking sites, for over 10 billion photos of individuals,[10] which was then used in the production of their software. Clearview AI have faced numerous lawsuits and fines relating to privacy violations from countries around the world.[11] The company carried out its work in relative obscurity for 3 years, until the New York Times wrote about them in 2020.[12] As they were a small organisation, and effectively impacted only a small number of people, they were able to evade public scrutiny, and even public *awareness*, for a long time.

An example of an AI with a big impact, run by a big team:

6 Large Recommendation Systems, such as those run by Amazon, are (often AI-powered) tools that recommend items to consumers on large platforms.[13] These can be items for purchase (in the case of Amazon), media to view (in the case of Netflix, YouTube), etc. Such systems face a myriad of ethical issues and potential harms, such as inappropriate content, privacy, autonomy and personal identity, opacity, fairness, and wide scale social effects.[14] Many of these platforms employ tens of thousands of employees,[15] or even millions.[16] And the impact of these recommenders is massive: YouTube and Amazon have billions of users,[17] and billions in sales,[18] respectively. The sheer scale of these AI-based systems being deployed can have a significant impact on societies.

An example of an AI with a small impact, run by a big team:

7 In 2022, a team of researchers at a drug discovery company were developing AI to postulate new chemical compounds with the minimum possible toxicity, which could then be synthesised, tested, and eventually applied to medical uses. However, by simply "reversing" the learning objectives to instead postulate new chemical compounds with the *maximum* possible toxicity, the team developed an AI that identified "many other known chemical warfare agents", as well as "new molecules [that] were predicted to be more toxic … than publicly known chemical warfare agents".[19] The impact space of such developments would, though severe for affected individuals, potentially only be on a small proportion of the population. But it would not be an easy process. Those who produced the research had extensive experience and expertise, having "spent decades using computers and AI to improve human health".[20] Moreover, the testing, verification, and production of such postulated compounds requires significant understanding of chemistry, as well as elaborate and expensive production facilities; to do all this requires a large, well-resourced organisation.

8 Given the myriad of possible use cases and scenarios, how can lawmakers foresee all possible uses and harms of AI, so as to legislate, regulate, or otherwise intervene to prevent or address them? This might not be necessary. In this paper, we present a combination of disincentives for AI operators, and

9 195 according to their LinkedIn page: Stability.ai, LinkedIn, https://uk.linkedin.com/company/stability-ai (last accessed Dec. 6, 2023).

10 See Knight, *Clearview AI Has New Tools to Identify You in Photos*, Wired (Oct. 4, 2021), https://www.wired.com/story/clearview-ai-new-tools-identify-you-photos/ (last accessed Feb. 29, 2024).

11 See Mac, Hill, *Clearview AI settles suit and agrees to limit sales of facial recognition database,* New York Times (May 9, 2022), https://www.nytimes.com/2022/05/09/technology/clearview-ai-suit.html (last accessed Feb. 29, 2024).

12 See Hill, *The Secretive Company That Might End Privacy as We Know It*, The New York Times (Jan. 18, 2020), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html (last accessed Feb. 29, 2024).

13 See Milano et al., *Recommender systems and their ethical challenges*, 35 AI & SOCIETY 957 (2020), https://link.springer.com/article/10.1007/s00146-020-00950-y (last accessed Feb. 29, 2024).

14 ibid.

15 See Andreeva, *Netflix Hit By Layoffs; About 150 Mostly U.S.-Based Employees Affected*, Deadline (May 17, 2022), https://deadline.com/2022/05/netflix-layoffs-us-employees-1235026085/ (last accessed Feb. 29, 2024).

16 See Yahoo Finance, Amazon Inc., https://finance.yahoo.com/quote/AMZN/profile (last accessed Feb. 29, 2024).

17 See Wong, Bottorff, *Top Social Media Statistics And Trends Of 2024*, Forbes, https://www.forbes.com/advisor/business/social-media-statistics/ (last accessed Feb. 29, 2024).

18 See Statista, *Net e-commerce sales of Amazon from 2018 to 2027, by business model*, https://www.statista.com/statistics/1309709/amazon-e-commerce-retail-sales-business-models/ (last accessed Feb. 29, 2024).

19 See Urbina et al., *Dual use of artificial-intelligence-powered drug discovery*, 4 Nature Machine Intelligence 189 (2022), https://www.nature.com/articles/s42256-022-00465-9 (last accessed Feb. 29, 2024).

20 ibid.

monitoring of AI behaviour, which we believe may significantly prevent harm. After all, we believe that the primary reason for using legal tools to regulate AI lies in the prevention of harm.

II. Difficulties with Existing AI Regulatory Proposals and Ideas

9 Many suggestions for ways to regulate and control AI can be found in the academic literature,[21] draft laws, and regulatory and legal practice,[22] even though the current narrative of a US-China "race" to Artificial General Intelligence (AGI) has "been used to justify reduced regulatory oversight".[23] However, as we outline below, most of them are infeasible or leave substantial gaps in the detection or mitigation of harm. For example:

1) Proposals to introduce algorithms to monitor algorithms,[24] or go further and challenge algorithms such as a "Public AI",[25] might seem tempting; deploying a technological solution to an apparently technological problem. But these suggestions contain hurdles. Firstly, verifying that the good AI, which checks (or challenges) the potentially bad AI, is doing an effective job is very difficult; do we not also need an AI to check the checker, and so on? If building such a tool were easy, it could have been incorporated into the original AI. Secondly, the cost, time, and expertise required to produce such an "AI checker" could be prohibitive. Existing AI systems, such as ChatGPT, require access to massive training data sets, massive compute power, large teams of extremely expensive experts,[26] consume massive amounts of electricity and other resources,[27] and take a long time to set up.[28] Moreover, given that these checkers would be needed for a particular context, jurisdiction, or social setting, different states and/or regulatory entities would be required to build and maintain their own versions. Sharing one general version is unlikely to be satisfactory. So only the most wealthy, well-resourced entities or states would

21 For an overview of existing guidelines and related proposals see Schiff et al., *What's Next for AI Ethics, Policy, and Governance? A Global Overview, in* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20, Feb. 7-8, 2020), pages 153-158, https://doi.org/10.1145/3375627.3375804 (last accessed Feb. 29, 2024); Hagendorff, *The ethics of AI ethics: An evaluation of guidelines*, 30 Minds and Machines 99 (2020), https://doi.org/10.1007/s11023-020-09517-8 (last accessed Feb. 29, 2024); Ryan, Stahl, *Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications*, 19(1) Communication and Ethics in Society 61 (2021), http://dx.doi.org/10.1108/JICES-12-2019-0138 (last accessed Feb. 29, 2024).

22 For an overview on some recent regulatory and legal approaches and a comparison between the AI Act and UK's legislative approach see Hacker, *AI Regulation in Europe: From the AI Act to Future Regulatory Challenges*, arxiv (Oct. 6, 2023), https://arxiv.org/abs/2310.04072 (last accessed Feb. 29, 2024).

23 Ó hÉigeartaigh, *The most dangerous fiction: The rhetoric and reality of the AI race.* SSRN Preprint (2025) http://dx.doi.org/10.2139/ssrn.5278644 (last accessed Jan. 4, 2026).

24 See Machin et al., *Specifying Safety Monitors for Autonomous Systems Using Model-Checking, in* Proceedings of the International Conference on Computer Safety, Reliability and Security (SAFECOMP, Sep. 2014), pages 262-277, https://link.springer.com/chapter/10.1007/978-3-319-10506-2_18 (last accessed Feb. 29, 2024).

25 See Elkin-Koren, *Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence*, 7(2) Big Data & Society (Jul. 29, 2020), https://journals.sagepub.com/doi/full/10.1177/2053951720932296 (last accessed Feb. 29, 2024).

26 Butcher, *A.I. Specialists Can Earn as Much as NFL Players*, Dice (Jul. 17, 2017), https://www.dice.com/career-advice/a-i-specialists-earn-much-nfl-players (last accessed Feb. 29, 2024).

27 Freitag et al., *The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations*, 2(9) Patterns 1 (2021), https://doi.org/10.1016/j.patter.2021.100340 (last accessed Feb. 29, 2024); Hacker, *AI Regulation in Europe: From the AI Act to Future Regulatory Challenges*, arxiv (Oct. 6, 2023), https://arxiv.org/abs/2310.04072 (last accessed Feb. 29, 2024); Saul, Bass, *Artificial Intelligence Is Booming - So Is Its Carbon Footprint*, Bloomberg (Mar. 9, 2023) https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure (last accessed Feb. 29, 2024).

28 cf. Marr, *A Short History Of ChatGPT: How We Got To Where We Are Today*, Forbes (May 19, 2023), https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today (last accessed Feb. 29, 2024).

stand a chance of producing such a tool, if at all.

2) An expectation of AI operators to self-regulate and produce responsible and safe AI, even without regulatory obligation, might initially seem reasonable. After all, existing liability regimes might already incentivise AI operators to avoid causing harm, as that may trigger obligations to undo the harm or to compensate (but see 4. in this list). However, sufficient self-regulation is unlikely to occur in practice, for many reasons. Firstly, there are no certified "licensing" requirements on those developing AI or other mathematical products; it is not a professional vocation, unlike law or medicine.[29] And while standards continue to be set in various ways,[30] the mechanisms to enforce adherence to such standards remain weak. Second, AI developers (working for AI operators) will likely come from various training backgrounds, including computer science, mathematics, physics, and engineering. These provide minimal compulsory training in AI ethics and responsible research and innovation,[31] so developers may have had little if any such training.[32] At the moment, responsible development and ethical training are not forced upon developers, neither by a professional oversight body, nor by their own intrinsic disciplinary culture. Finally, our experience working with startups and teaching ethics to mathematicians,[33] has led to encounters with technical experts who profess that they are "just doing the technological development, and that it is not their responsibility to ensure its safety or ethics". Just as the legal and regulatory community might expect operators to take it upon themselves to produce safe tools and products, the operators might look towards lawmakers for legislative guidance to establish at least the basic regulatory framework with an expectation that they only need to "follow the law". Thus, the responsibility to regulate AI is simultaneously passed back to lawmakers. Going beyond the existing legal framework by self-imposing further checks or constraints will also potentially affect profitability and hence concern investors or share-holders, making self-regulation unlikely unless all relevant actors in the industry agree.

3) The ability to regulate AI often presumes that its use is known, and its impact understood, or at least verifiable. However, relying on the possibility of detecting the presence or use of AI can be misplaced. Many uses of AI fail to leave any trace or "watermark". It is hence difficult if not impossible to tell when AI has been utilised and/or served as one, or even the essential, element or process leading to an outcome impacting on others. This is particularly true for AI systems that make choices, decisions, or recommendations. For example, AI used to shortlist job applicants for interviews based on their CV data may raise many ethical issues,[34] including that there may be no way of identifying the reliance on AI in the first place. The shortlisting could have been done by hand, or by some elementary algorithm giving a rudimentary score such as "(years of experience) divided by (salary requested)", or by an AI trained on a large dataset; all are plausible. Similarly, AI might be used as part of a larger techno-social system, including search engine recommendations, online advert distribution, calculating insurance premiums, and determining public transport timetables. In all these applications, and many others, the output is straightforward: the user or customer is presented with a list of search results, shown an advert, given an insurance cost, or handed a bus map and timetable. An end-user or impacted party, as much as a regulator, has almost no way of knowing if, or how, AI was used in such a process.

4) A more 'hands-off' regulatory approach might suggest relying on existing rules of tort liability (for example, for faulty products or substandard service provision) and apply them to the use of AI once harm has been caused. However, the assumption that this might induce liability-avoiding behaviour ex ante, and - more importantly - of being able to restitute or compensate for the relevant harm ex post-impact, may not always hold. Akin to the global financial crisis of 2008, when the world saw banks and insurance companies deemed "too big to fail", some AI processes may cause damages

29 Müller et al., *A Hippocratic Oath for Mathematicians? Mapping the Landscape of Ethics in Mathematics*, 28(5) Sci Eng Ethics 41, https://doi.org/10.1007/s11948-022-00389-y (last accessed Feb. 29, 2024).

30 Kilian et al., European *AI Standards - Technical Standardization and Implementation Challenges under the EU AI Act* (Feb. 26, 2025). SSRN Preprint (2025) https://dx.doi.org/10.2139/ssrn.5155591 (last accessed Jan. 5, 2026).

31 Horton et al., *Embedding Ethics in Computer Science Courses: Does it Work?, in* Proceedings of the 53rd ACM Technical Symposium on Computer Science Education (SIGCSE 2022, Mar. 3-5, 2022), pages 481-487, https://doi.org/10.1145/3478431.3499407 (last accessed Feb. 29, 2024).

32 Müller et al., *A Hippocratic Oath for Mathematicians? Mapping the Landscape of Ethics in Mathematics*, 28(5) Sci Eng Ethics 41, https://doi.org/10.1007/s11948-022-00389-y (last accessed Feb. 29, 2024).

33 Chiodo, *The Ethics in Mathematics Project*, Cambridge University Ethics in Mathematics Project (Dec. 5, 2025), https://www.ethics-in-mathematics.com (last accessed Dec. 5, 2025).

34 Kodiyan, *An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool* (Nov. 2019), https://www.researchgate.net/publication/337331539_An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazon%27s_AI_based_hiring_tool (last accessed Feb. 29, 2024).

which are too big to undo, compensate, or fine for and hence would have created harmful consequences that are too significant to 'fix' (i.e., restitute in ways commonly set out in tort law). Large-scale technology-induced harm, such as oil spills and other disasters, are often dealt with by enormous human effort and commensurate payment from the organisation at fault.[35] However, this assumes that the entity responsible, its insurer, or - as a matter of last resort - the relevant polity affected by the harm has sufficient resources to undo or otherwise address the harm caused. Given the scale and impact some uses of AI can have,[36] it might not necessarily be the case that the damage done can in fact be 'fixed' at all.[37] Since, as a matter of last resort, humanity at large, other living beings, and our planet shoulder those risks which cannot be restituted or compensated for if materialised, we believe that preventative measures are needed, not just post-harm recourse.

5) Conceptions of legal responsibility under private law generally provide an effective recourse for individuals who have been harmed or suffered losses as the result of the actions of another. As suggested above, it may be tempting to assume that tort and related liability regimes can operate satisfactorily to address the harms caused by AI. But, in addition to the ex post vs ex ante issue already considered, the entity producing or deploying the AI may in practice escape liability due to difficulties in establishing that AI has been involved, showing a sufficient causal link, or offering relevant evidence.[38] In addition, some of the most important actors involved in producing or deploying "Big AI" (as defined below) at scale (Microsoft, Google, Meta, etc.) are so well-resourced that an individual harmed stands almost no realistic chance of successfully pursuing claims against them in litigation. In addition to these hurdles, major further practical roadblocks are likely to arise in the form of finding a court with jurisdiction and determining the applicable law, contractual limits of liability imposed via terms of service (if valid), and the inability to enforce judgments in jurisdictions where the relevant actors have significant assets. Furthermore, existing laws might not always sufficiently conceptualise the end-user's interest in a way that captures the potential harm done by AI. If, for example, a large video sharing platform, such as YouTube, takes down a video of a birthday party on the basis of someone's copyright claim that is automatically enforced via Content ID,[39] most copyright laws will struggle to offer a legally recognised interest that the uploader can put forward in claiming reinstatement of the video on the platform. While users certainly enjoy (fair use and other) defences against infringement claims by rights holders, those are likely to remain ineffective when AI tools automatically enforce such claims across the platform: Not only is AI (so far) unable to comprehend contextual issues such as (fair) uses for the purpose of parody, critique or quotation - but conventional copyright doctrine is only beginning to respond to automated private ordering tools when deployed to prevent a defensible use, or to enforce the monetisation of such

35 Cf. Taylor, *Catastrophes and Insurance*, in Global Catastrophic Risks 164-183 (Bostrom, Cirkovic, Jul. 2008), https://doi.org/10.1093/oso/9780198570509.003.0012 (last accessed Feb. 29, 2024); Chang et al., *Consequences of oil spills: a review and framework for informing planning*, 19(2) Ecology and Society 26 (2014), http://dx.doi.org/10.5751/ES-06406-190226 (last accessed Feb. 27, 2024); Leach, *Litigating Catastrophe*, 110 California Law Review 2149 (2022), https://doi.org/10.15779/Z38MK65918 (last accessed Feb. 27, 2024).

36 Crawford, *Atlas of AI. Power, politics, and the planetary costs of artificial intelligence*, Yale University Press (Apr. 6, 2021).

38 Schütte et al., *Damages liability for harm caused by Artificial Intelligence–EU law*, 69 Helsinki Legal Studies Research Paper 1 (2021), https://dx.doi.org/10.2139/ssrn.3897839 (last accessed Feb. 27, 2024).

39 Grosse Ruse-Khan, *Global Content Protection Through Automation – A Transnational Law Perspective*, 49 International Review of Intellectual Property and Competition Law 1017 (2018), https://doi.org/10.17863/CAM.35223 (last accessed Feb. 28, 2024).

8,5 Pt

use in favour of someone claiming copyright.[40]

III. AI as a Product or as a Service?

10 In addition to these shortcomings in the recent proposals to regulate AI, there is an overarching ambiguity in terms of classifying AI within conventional categories of commerce: some AI systems may be embedded into a product or understood as constituting a (digital) product, while others may be seen as part of providing a service, whereas many might involve a combination of both. There is an ongoing debate about this distinction, in particular within the EU: The EU Artificial Intelligence Act (AI Act) is conceived as a product safety instrument or product regulation. The European legislator thus treats AI systems as products. This is due to the fact that product regulation is harmonised and well-functioning since the EU's New Legislative Framework (NLF) entered into force. The NLF provides for a coherent certification and conformity assessment procedure and a uniform market surveillance for products placed on the market in the EU. The provisions of the AI Act are to be integrated into these legal structures and existing processes and procedures. In particular, the interplay between abstract legal requirements from legislative acts and the concrete instructions for actions from technical norms and standards is to be used for AI systems. In concrete terms, the method of co-regulation, which has established itself as fast, flexible and practicable and in which the EU legislator lays down basic principles while standardisation bodies determine the concrete action to be taken, is to be applied to AI systems.

11 However, certain AI systems (even those embedded in a product) have very different characteristics from conventional products. Some AI systems do for example change constantly after being placed on the market, especially if the training phase is not finished at the time of market entry, or if the AI is constantly being revised and improved. For some AI systems it is impossible to foresee what they will be used for at a later stage (e.g., General Purpose AI - GPAI) which makes an ex ante conformity assessment difficult to impossible.

12 Moreover, AI systems are often offered as a service ("AIaaS") and not as a product. This may be due to the operators' efforts to maintain trade secrecy and because producing and deploying AI systems requires an enormous computing capacity. By using AIaaS, smaller companies gain access to cutting-edge and up-to-date AI developments which they would not be able to use otherwise, since they may be missing the necessary (training) data, know-how, and computing resources to create, test and maintain these AI systems on their own.[41]

13 In contrast to products, regulating services - while not uncommon in regards to financial, insurance or legal industries as well as platforms - is a significantly more complex exercise, not least since there is no specific tangible item for the regulator to focus on. Regulating AIaaS directly does not seem to be on the regulatory agenda at the time of writing. When comparing the two concepts, however, it shows that the extent to which the actors involved can actually exercise influence varies greatly depending on whether an AI system is placed on the market as a product or the AI is offered as a service, which is why the legal responsibilities should be distributed accordingly. While AIaaS providers have a much bigger (or exclusive) and ongoing influence and impact on the training of their systems as well as the use, those same providers of AI lose that influence and often their ability to control their AI after placing such an AI-equipped product on the market. This may happen if, for example, a deployer makes, or is able to make, substantial changes to the AI.

IV. Terminology Concerning the Actors Involved

14 Regardless of the distinction between product and service, there is a distinction between different actors involved in the various stages of the lifecycle of AI systems. Based on the general distinction in product regulation, we differentiate between the producer and the user side. Following the terminology of the AI Act, we call the actors on the producer side *providers* and actors on the user side *deployers* (clarifying that end users are not implied herewith).

15 Both, providers and deployers, can be subsumed under a broader understanding of *operators*, art. 3 sec. 8 AI Act.[42] However, the term operators may also be understood in a narrower sense, relating to users responsible for the safety of products, art. 3

40 Grosse Ruse-Khan, *Automated Copyright Enforcement Online: From Blocking to Monetization of User-Generated Content* (Mar. 1, 2020), https://dx.doi.org/10.2139/ssrn.3565071 (last accessed Feb. 29, 2024).

41 Cobbe, Singh, *Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges*, 42 Computer Law & Security Review 105573, 2 (2021), https://doi.org/10.1016/j.clsr.2021.105573 (last accessed Feb. 28, 2024).

42 Regulation of the European Parliament and of the Council 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence 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), hereinafter AI Act.

lit. e of the European Parliament's (now withdrawn) proposal for the introduction of an AI liability (see C.I.2).[43]

16 Throughout this paper, we use the terms *providers* on the producer side and *deployers* on the application side; the term *operator* is used as a term encompassing both. We will occasionally use the term *developers* in specific reference to those (individuals or teams) who are doing AI work within an AI operator. We use the term *users* to describe those people who engage with, or are impacted by, AI systems. These are summarised in Figure 1.

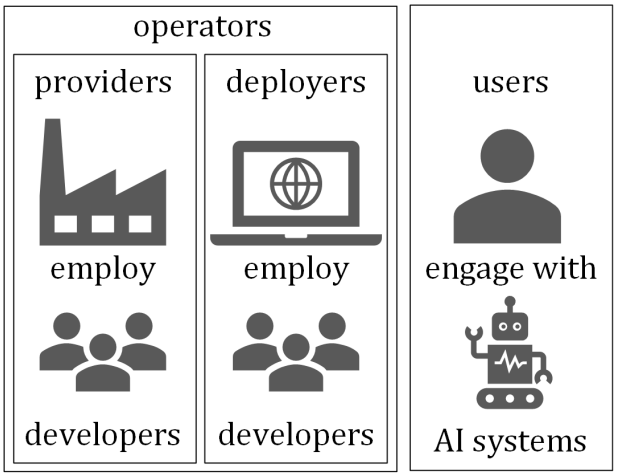


Figure 1: Terminology and actors involved in AI

V. A Feasible Way of Classifying AI

17 In order to adequately regulate AI, one needs to differentiate it along many axes. While the risk axis used by the EU in the AI Act[44] is valuable, it heavily depends on understanding, in advance, the growing list of application domains and types of AI systems, and on the capacity for classifying AI risks accordingly. Therefore, rather than building on a (potentially unlimited) number of different application domains, we propose to open two *additional* axes in addition to the existing EU risk classification, and further distinguish AI by its:

- Impact size, described as **Small AI** and **Big AI**, and
- Team size, described as **Small Team** and **Big Team**.

18 This leads to the following 2x2 matrix (i.e., grid) of "types" of AI, which we explain below.

	Big AI (big user base, big social impact)	**Small AI (small user base, small "direct" social impact, possible larger "indirect" impact)**
Big team	Examples: Twitter, Facebook, Google, Garmin, large retailers ● AI is usually part of a larger range of products or services. The product can be a corporeal item into which AI is integrated (e.g., in smartwatches). ● It takes many years to become a Big Team, working on Big AI. ● Many entities in this corner can also be called "gatekeepers".	Examples: manufacturers of specialised products, smaller medical or pharmaceutical companies ● AI is often part of an expensive product that is used by a small, often specialised group of people or sold to other businesses (B2B). ● AI is used to create and manufacture other tools or items. ● These companies are occasionally called "hidden champions" in the business literature: due to their unique role, they may be invisible to the non-expert or to people working in different areas, despite their size.
Small team	Examples: (early) OpenAI with ChatGPT, (early) Stability.ai with stable diffusion ● AI is often the only product, it gets sold to other companies or goes viral. ● The transition from(Small Team, Small AI) to (Small Team, Big AI) can be extremely fast, e.g., ChatGPT had 100million users within 2 months of launch.	Examples: Most startups in their early stages ● AI systems that stay in this corner often solve a niche problem, and regularly come with very specialised use-cases. ● For users and non-domain specialists it can be hard to understand who the AI providers are, and where the AI is used or sold. ● Startups are in this corner before they get big.

Figure 2: Matrix of AI types: Small AI vs Big AI, and Small Team vs Big Team

19 We do not specify precisely the impact size of "Big AI", in part because "impact" is a very difficult term to define. However, one useful proxy for impact, especially when considering risk and the potential for harm (as we do in this paper), is the *number of people (directly) affected.*[45] This is not a perfect measure, but nonetheless a useful measure of "impact". In addition, we note that the EU Digital Services Act (DSA) defines very large online platforms and very large online search engines (VLOPs and VLOSEs, together VLOPSEs) to be those which impact over 10% of the EU population.[46] Thus, taking "Big AI" to

43 Zech, *The European Parliament proposal for a regulation on liability for the operation of artificial intelligence systems,* 14 Zeitschrift für Geistiges Eigentum/Intellectual Property Journal 293, 298 (2022), https://doi.org/10.1628/zge-2022-0022 (last accessed Feb. 27, 2024).

44 See the "risk pyramid" in European Commission, *Shaping Europe's digital future: A European approach to artificial intelligence*, European Union (Jan. 31, 2024), https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (last accessed Feb. 28, 2024).

45 Which can come in many forms, such as "number of users", "number of people the AI makes decisions about", "number of customers sent AI-generated content", etc.

46 Art. 33 sec. 2 Regulation of the European Parliament and of the Council 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 1, hereinafter: DSA. See also section C.IV for a discussion on the DSA as an implementation framework for our proposals.

mean "AI that impacts over 10% of the population in question" would be a reasonable starting point for our discussion, though a more rigorous definition would be needed for any legislation.

20 Similarly, we do not specify precisely the size of a "Big Team". We envisage that any company which is publicly listed (or meets the threshold for public listing) would come under such a category. Or any company that meets a minimum number of employees (of, say, 1000). As such thresholds are only set for the purpose of our analysis, there is a degree of flexibility in how to interpret this term. Moreover, in our analysis we assume that "team size" correlates to the size of the overall company or organisation creating the AI, rather than the number of people directly involved in the development work. This is because all the factors that we are concerned with in Big Teams (internalised incentive to avoid doing harm, incentive for the company to not go bust, ability to defend against lawsuits, etc.) come from the overall size of the company.

1. Small Team - Small AI

21 The number of graduates focusing their studies on AI is rapidly rising and there is an ever-increasing number of companies creating AI tools.[47] Many of these companies with <50(ish) employees, in particular many startups, are producing AI tools, products and services for use in the open market. They regularly act as B2B suppliers, creating tools that other, often more prominent, corporations use for their day-to-day operations or as part of a larger product. These companies, and the work that they do, are often "invisible". Few people know they exist, and even those impacted are often unaware this is (partially) as a result of AI. It is also in this sense that Small AI leaves no easily detectable trace or "watermark". Such an AI system may still (in)directly impact millions of people, but not normally very large groups on a global scale (i.e., billions of people), and so can remain under the legal and regulatory radar for a long time. Their impact (as an individual organisation) is "small" when compared to big AI operators, and in comparison to the proportion of the overall population affected by their AI. As discussed in Section C.IV.2, digital regulation in the EU includes specific allowances and exemptions for small and micro enterprises with fewer than 50 employees and no more than EUR 10 million turnover.

22 Some of the authors have observed in their work and teaching that it is often tempting for such Small AI operators, and in particular providers, to operate with minimal, or even no domain expertise. Their team of developers may have formed around a small group of friends or associates, each expert in some technical discipline such as mathematics or computer science, or the hiring may not have been done with diversity of skills and knowledge in mind. And yet they may be working in an area in which their understanding is no better than that of a layman, with "big data" their only source of insight. In the case study of a young medical startup, Sappington and Serksnyte[48] show how an overly strong focus on expertise in engineering and mathematics can actively disregard many forms of domain knowledge, a scenario which some of the authors have directly observed too.

23 The lack of relevant expertise and resources in Small AI startups often means that they need to find alternative ways to build and maintain an ethical culture within their company.[49] However, this may incidentally lead to such companies believing they can try to contract out their responsibility, risk and liability for potential harm to the entities they provide their services or products to, and thus may take on zero moral or social responsibility themselves (and of course aim to avoid legal liability). Additionally, malicious and even non-malicious actors may occasionally operate with impunity by looking for legal and regulatory loopholes.[50] There are very few mandatory standards for them to follow, and many of the ethical frameworks[51] are merely voluntary. In their work with startups, some of the authors found that the notions of "malpractice", "negligence" or "duty of care" - let alone corporate social responsibility - appear not to have penetrated sufficiently into the team's (business) culture. Nor is the concept of "doing things properly" well-defined because there are no industry standards, accepted norms, and agreed-upon best practices to rely on.

47 cf. SIHCAI, *Artificial Intelligence Index Report 2023*, Stanford University (2023), https://aiindex.stanford.edu/ (last accessed Feb. 28, 2024).

48 Sappington, Serksnyte, *Who and What Drives Algorithm Development: Ethnographic Study of AI Start-up Organizational Formation, in* Proceedings of the Ethnographic Praxis in Industry Conference (2018), pages 245-263, https://doi.org/10.1111/1559-8918.2018.01206 (last accessed Feb. 28, 2024).

49 Sloane, *Ethics in German AI Start-Ups* (Dec. 22, 2022), https://dx.doi.org/10.2139/ssrn.4336465 (last accessed Feb. 28, 2024).

50 See Rainie et al., *Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm Within the Next Decade*, Pew Research Center (Jun. 16, 2021), https://www.pewresearch.org/internet/2021/06/16/experts-doubt-ethical-ai-design-will-be-broadly-adopted-as-the-norm-within-the-next-decade/ (last accessed Mar. 5, 2024), for some expert commentary on this and related problems.

51 For an overview see Ayling, Chapman, *Putting AI ethics to work: are the tools fit for purpose?*, 2 AI and Ethics 405 (2022), https://doi.org/10.1007/s43681-021-00084-x (last accessed Feb. 28, 2024).

Hence, attempts to punish small teams for not developing and maintaining their Small AI properly, or attempts to audit and check their output, would regularly fail. As a former member of the United States National Security Council, Roger Cressey, has argued: AI development "is the Wild West right now"[52] with nothing to keep these cowboys in check.

24 It may not seem as though Small AI should be of any broad societal concern, as their impact is rather restricted. However, in aggregate they do pose a risk to society through repeated and often undetected (hence not triggering a regulatory response) conduct, which might be described by a "death by 1000 cuts" metaphor. That is, numerous, perhaps primarily small-scale, impacts by Small AI on individuals, groups and even on how society functions that, however, in its combined effects can be severe. This might be in the form of people receiving inappropriate loan sizes or conditions, via the building of ineffective road networks, via the poor placement of hospitals, schools, and other essential services, via poor logistic networks, via the recommendation of inappropriate or inadequate products or services to people to buy or use, and so on. Aware that our categorisation involves simplifications, we suggest that AI systems that fit into the types described above share sufficient common denominators for the purpose of our risk and liability analysis to place them in one category. We call this "Small Team - Small AI".

2. Big Team - Big AI

25 Some large companies are implementing AI that regularly impacts a very large number of individuals, groups and even societies (usually beyond national borders), including large social media platforms. These companies will often be large-scale gatekeepers or decision makers, where in particular the cumulative effects of conduct by these entities are likely to have a profound effect on societies. This might be through providing a quasi-essential service such as a search engine, prolific social media platform, or online marketplace, all of which vast numbers of people have come to rely on and "could not do without". Or it could also be through the provision of platforms that other businesses build their own products and services around (such as web hosting, cloud computing, digital marketing, AI content generation, etc.). The impact of such conduct can be enormous: a Big AI used within a large organisation might change the way a significant portion of humanity acts, how people think, perceive the world, carry out work, or how they interact with each other. Even just one Big AI operator can, single-handedly, re-shape societies by impacting a significant proportion of the population. For example, a large search engine can change, or even distort, the information a significant proportion of the public can or cannot find. Or a video content recommender could recommend harmful or radicalising content to masses of people. Even if such impact is considered minor at the individual level, its sheer scale makes it significant.

26 The companies deploying Big AI are often extremely visible. Due to their large user base and ambitions, they cannot hide, and are not trying to. They aim to have as many users, customers or subscribers as possible, and their existence is usually well-known. They utilise AI and other algorithmic tools to filter and process vast amounts of data daily. This often leads to scenarios where Big AI influences how a lot of people think, and how they see the world.[53] As exemplified by the Cambridge Analytica scandal, this poses a political threat, as now one organisation can be the conduit for significant political influence[54] - without this being immediately apparent.

27 Serious problems might arise from the scale of their impact across society and the associated aggregated impact, rather than from individual actions or decisions carried out by such an organisation. For example, large platforms may use AI to recommend products or services to their users, control what they have decided to constitute hate speech, implement mechanisms to monitor other content the platform defines as 'harmful', or act to remove alleged copyright infringements. While blocking, de-listing or otherwise censoring access to any *one* article, post, video or other content may not have terribly harmful consequences for the author or society, large-scale conduct that deviates from applicable laws and societal norms on free speech is likely to have distortive effects on any democratic society, including by shaping - over time - what society considers to be acceptable as free speech. Even if no

52 Klar, *AI 'wild west' raises national security concerns*, The Hill (Mar. 8, 2023) https://thehill.com/policy/technology/3888433-ai-wild-west-raises-national-security-concerns/ (last accessed Feb. 28, 2024).

53 Susser et al., *Technology, autonomy, and manipulation*, 8(2) Internet Policy Review 1 (2019), https://doi.org/10.14763/2019.2.1410 (last accessed Feb. 28, 2024).

54 ur Rehman, *Facebook-Cambridge Analytica data harvesting: What you need to know*, 2019 Library Philosophy and Practice 2497; Berghel, *Malice domestic: The Cambridge analytica dystopia*, 51(5) Computer 84-89 (2018), https://ieeexplore.ieee.org/document/8364652 (last accessed Feb. 29, 2024); Wilson, *Cambridge analytica, Facebook, and Influence Operations: A case study and anticipatory ethical analysis, in* Proceedings of the 18th European conference on cyber warfare and security (Jul. 4-5, 2019), pages 587-595, https://www.proquest.com/openview/b2de0d9e0b088a0711696220cfa58d82 (last accessed Feb. 29, 2024).

one decision or action by Big AI may be particularly detrimental, at scale and in aggregate, its impact can shift societies in potentially unsustainable ways.[55] Once again, we see the harm here as "death by 1000 cuts". And even though the source here is centralised, the large-scale harm can arise from a compounding of minor harms.

28 In the name of enforcing ostensibly "good" rules such as the removal of harmful content, Big AI deployers might over-enforce such rules in the absence of any disincentives not to. Overall, this can lead to a systematic distortion of speech and communication, changing the entire information ecosystem. And while there are existing legal mechanisms to hold such organisations to account,[56] there are several challenges in doing so. As covered in Section A.II, there are difficulties in taking legal action against large entities deploying AI; even the Australian Federal Government needed to back down against Facebook.[57] There are difficulties in making the causal link between harm caused and the AI, let alone identifying AI was used in the first place. And there are difficulties in remedying the large-scale harms, given their type and sheer scale. All these suggest a new, tailored regulatory approach is needed. We call this "Big Team - Big AI".

3. Big Team - Small AI, and Small Team - Big AI

29 It might seem that specifying *team size* for AI is pointless, as in the examples we have given, Small Teams do Small AI, and Big Teams do Big AI. And as explained above, we assume that the size of the team generally correlates to the overall size and significance of any AI that team might produce or deploy. However, this is not always the case. Sometimes, Small AI is done by a Big Team, as might happen in a pharmaceutical research laboratory creating bespoke drugs for rare diseases. In such cases, a large organisation with substantial resources and many staff is developing an AI with a niche application which (directly) impacts only a relatively small number of people. Here it is probably safe to assume that the size and breadth of the team, combined with the very constrained application space, makes the use of AI in this setting rather low-risk for widespread societal damage, despite the costs for individuals potentially being extremely high. However, Big Teams producing B2B products and services for other organisations may have a significant, usually more indirect impact on society, not least because the consumers and other end-users of the respective products or services with inbuilt AI might not even be aware of the involvement of an AI system. This could, for example, occur in the pharmaceutical industry, either in the case of personalised medicine, or where AI is used as part of the pharmaceutical research process postulating which drugs to try (or indeed not try) to develop, and how.[58] We call this "Big Team - Small AI".

30 Conversely, a Small Team might be developing a Big AI, such as a startup that creates a widely-applied or widely-used product with significant uptake - such as, for example, an originally Small Team of OpenAI developers creating ChatGPT or the nowadays widely used Stable Diffusion from stability.ai. Here we might see a Small Team of developers create and run a product or service that directly impacts over 100 million people.[59] In this instance, we observe all the risks of Big AI described above, such as the potential for distortive effects on society that cannot be rectified ex post via means of dispute settlement. At the same time, we also see all the drawbacks of Small Teams, potentially including a lack of breadth, insight, and resources, thus making it hard for them to fully understand their impact space and hence to be able to address risks posed. Just as for Small Team - Small AI, in a Small Team - Big AI situation, one can also occasionally observe an abdication of responsibility by attempting to contract out potential harms through contracts and terms of service.[60] This combination of the risks of Big AI,

55 Arogyaswamy, *Big tech and societal sustainability: an ethical framework*, 35 AI & Society 829 (2020), https://doi.org/10.1007/s00146-020-00956-6 (last accessed Mar. 5, 2024).

56 See the discussion of different legal regimes in Section C; as well as Perc et al., *Social and juristic challenges of artificial intelligence*, 5 Palgrave Communications 61 (2019), https://doi.org/10.1057/s41599-019-0278-x (last accessed Feb. 29, 2024).

57 See Taylor, *Deliberate ploy: whistleblowers reveal why Facebook's Australia news ban included non-news sites*, The Guardian (May 28, 2022), https://www.theguardian.com/technology/2022/may/29/deliberate-ploy-whistleblowers-reveal-why-facebooks-australia-news-ban-included-non-news-sites (last accessed Feb. 29, 2024).

58 For an overview of the possibilities of AI usage in personalised (cancer) medicine we refer to Schork, *Artificial Intelligence and Personalized Medicine, in* Precision Medicine in Cancer Therapy 265-283 (von Hoff, Han, Jun. 2019), https://doi.org/10.1007/978-3-030-16391-4_11 (last accessed Feb. 29, 2024).

59 ChatGPT had more than 1 million users within 5 days, see Ul Haque et al., *"I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data*, arxiv (Dec. 12, 2022), https://doi.org/10.48550/arXiv.2212.05856 (last accessed Feb. 28, 2024), and well over 100 million users after only 2.5 months; for a rich discussion of its large-scale impact, see Leiter et al., *ChatGPT: A meta-analysis after 2.5 months*, arxiv (Feb. 20, 2023), https://doi.org/10.48550/arXiv.2302.13795 (last accessed Feb. 29, 2024).

60 One of the authors has extensive experience working

and the shortcomings of Small Teams, makes the use of AI in this setting susceptible to being extremely high-risk. We call this "Small Team - Big AI."

4. Risks and Progression in the Matrix of AI Types

31 We have established three key observations about the matrix of AI types:

- That "team size" and "AI size" are often, but not always, correlated,
- That "team size" and "AI size" can sometimes differ, and
- That the risk to society posed by the organisation increases as the AI size increases, and as the team size decreases. That is, in general terms, the riskiest AI potentially comes from "Small Team - Big AI", and the safest from "Big Team - Small AI".

32 What complicates matters further is that an AI operator can *move* to different squares in the matrix; their placement is not static. For example, a Small Team - Small AI operator might be using AI for niche research, and evolve or grow into a very large research team/organisation doing much more sophisticated and elaborate AI work, but broadly within the same niche area, thus becoming a Big Team - Small AI operator. Alternatively, a Small Team - Small AI might have a product or service that "takes off" and very rapidly gains 100 million users on a platform, or is deployed over a global supermarket chain with 100 million shoppers, thus becoming Small Team - Big AI. Or a Small Team - Big AI might scale up their operations and employ more developers and staff to cope with their increased impact and associated responsibilities, thus becoming a Big Team - Big AI. So, unlike regulating a use-case of AI (such as for insurance pricing) which is static with any given organisation, the AI type of an operator can change, sometimes rather quickly. Thus, one single regulatory approach is unlikely to be successful, and different strategies, corresponding to the dynamic developments described here, are needed. And as the context in which AI is deployed changes, there needs to be sufficient regulatory overlap to cover dynamically developing situations as they move between different squares in the matrix; whatever the development and deployment methodologies of the company are, and however they change, there should always be some, if not several, of these regulatory tools applicable. An AI operator should not be able to adjust its team size, and/or deployment size and space, in such a way that none of the regulatory tools (regarding Big/Small Team, and Big/Small AI) are in force.

with AI startups, providing advice to them for ethical and responsible AI development. They have dealt with over 25 AI startups, from 2020-2022. Part of their work involved investigating which entities within the startup, and which external entities (i.e., clients), would be responsible for the proper and safe functioning of the AI. In nearly all instances, across nearly all startups, this author was told some version of the statement "We contract out all responsibility for proper functioning, and subsequent liability for any malfunction, to the client".

B. Potential Solutions from a Practical Perspective

33 Having introduced the Big-Small axes (of team size, and AI size) to analyse the risk posed by AI systems, we now propose *two complementary regulatory tools which, when deployed together, become a solution* that we believe fits with our previous analysis. One is for Small AI: a form of strict(er) liability for the harm caused by AI. And one is for Big AI: a form of regulatory oversight by means of constant monitoring and cumulative penalties. This section outlines and addresses the technical, social and legal challenges for each solution; whereas the further issues of anchoring these solutions within a domestic legal and regulatory framework will be addressed in Section C.

I. What Can We Do to Address the Risks Posed by Small AI?

1. The Problem of Small AI

34 As explained, Small AI is often undetectable to wider society or impacted parties, its uses and domains are practically limitless and difficult to forward-predict, and it can be rapidly deployed in unregulated areas.[61] So regulatory frameworks assuming it is discoverable, or can be foreseen, or done on a per-sector basis, are unlikely to succeed. Small AI is often developed *exclusively* by technically-trained people (mathematicians, computer scientists, etc.) without domain expertise or insight into non-technical and ethical aspects of their work, assembled in Small Teams which follow a high-risk high-return business model and are prepared (or even expect) to fail. So regulatory frameworks relying on such organisations to self-regulate and develop AI responsibly, or that

61 One of the authors witnessed an AI startup team make the (paraphrased) comment "we only want to work in unregulated industries", as part of his work as an ethics consultant for Machine Intelligence Garage (part of Digital Catapult UK) in 2022.

they have a principal interest to survive, are unlikely to succeed.

35 Of course, there may already be ways to sue AI providers for damages resulting from the AI. However, as many Small AI organisations sell B2B, this "harm" manifests through the products and services offered by the B2C organisation that interacts with consumers or other end users, thus making it practically difficult to sue the AI provider, or indeed even identify that AI has been involved, its causal contribution to the harm materialised, and which entity might be responsible. The entity purchasing the AI tool could sue the Small AI provider for such damages, but in our experience most Small AI providers explicitly absolve themselves by contracting out all such risks and liability, thus (attempting to) place the responsibility for safe and effective operation of the AI, and any associated civil liability, on the AI deployer. Of course, whether such contracting out is enforceable is a different matter. While this type of contracting - even if passed on in agreements between the AI deployer and end-user - might not always serve to avoid tort liability vis-a-vis end-users (especially if mandatory consumer protection laws come into play), legal and practical difficulties remain. They include (1) difficulties for AI deployers or other actors at the end of a supply chain involving Small AI somewhere upstream to take effective redress action (if they have been successfully sued by a user for damages) against their contracting partners, or directly against the Small AI provider or developer upstream; (2) challenges in identifying (and proving) causal links, the materiality of a contribution, let alone breaches of a duty of care along complex supply chains; (3) as well as the willingness of the Small AI operator to go bust. And one must consider all the above in a cross-border context where actors along the supply chain potentially operate in different jurisdictions and subject their contractual relations to distinct applicable laws. This indicates that any sort of legal system or framework which relies on standard civil liability alone as a deterrent is unlikely to succeed.

36 In summary:

- One cannot detect the use of Small AI;
- One cannot forward-predict all uses of Small AI;
- One cannot regulate Small AI exclusively on a sectoral, industry-specific basis;
- One cannot assume Small AI entities have an inherent interest to self-regulate and act responsibly;
- One cannot rely on the survival desire of the legal form through which Small AI is deployed;
- One cannot assume that existing civil liability alone is a sufficient deterrent for Small AI.

37 In light of the above, one needs an alternative approach that convinces and incentivises those developing Small AI to address the concerns about AI risks. We outline a proposed solution below.

2. A Proposed Solution for Small AI

38 With the development of new, novel, innovative Small AI tools and applications, there is often only one group of people sufficiently well-placed to see and understand the operation, outcomes, and limitations of such tools: the development teams themselves. Thus, our proposal for how to address the problems created by Small AI will focus heavily on these development teams and the individuals within them.

39 Existing liability laws applied to AI, typically with cumulative requirements such as (1) needing to prove that harm was caused; (2) that there is a sufficient causal link to the operation of the AI; and (3) that those developing the AI acted intentionally or negligently, are simply too weak to be any serious behavioural deterrent to those developing Small AI. In cases of harm from Small AI, the existence of (individual) harm is often easier to determine and to observe when compared to Big AI, but proving a sufficient causal link between the harm and the AI's operation serves as a substantial burden. In addition, the typical requirement in conventional tort law to prove intent or negligence (in the form of the breach of a duty of care) on the part of the AI operator(s) then becomes a step too far, especially in an area such as AI where there are currently no well-established and agreed upon standards and norms.[62] So, established benchmarks for negligence hardly exist.

62 cf. Mittelstadt, *Principles alone cannot guarantee ethical AI*, 1 Nature Machine Intelligence 501 (2019), https://doi.

40 We therefore suggest the creation and adoption of a liability system that achieves the following:

- **Places a legal incentive on the Small AI operators to consider all realistically possible ways to prevent harm resulting from their AI.** This alleviates the need for regulators to be able to "review" all Small AI work - hence such an incentive, if effective, avoids the need for monitoring or auditing of the AI.
- **Extends such an incentive, and associated civil liability, to (some) individual persons involved.** This prevents such an operator from simply becoming a "phoenix" and dissolving itself to dissipate any liability.
- **Requires only proof of harm (existence, and causally linked to the AI), but not proof of negligent behaviour.** This removes the complexity for the legal system of understanding how the AI operates or was developed, and addresses the fact that standards and norms (and thus the concept of negligence) are not well-formed for AI development given how new it is. This pushes operators to do their own safety and quality assurance, to ensure their product is safe.
- **Allows the AI operator to reduce any regulatory penalties if it can prove that, despite the harm caused by its AI, it had taken all reasonable steps to prevent any foreseeable harm from materialising.** This draws from the aviation idea of proving it is safe to an acceptable level of risk (cf. Flannery 2001),[63] which is crucial to build faith and trust by the public in a new and uncertain sector. Given how unpredictable AI is, the ethos of "assume it is safe" needs to be broken.
- **Disallows the contracting out of the above defined responsibility.** This stops an AI provider from washing its hands of responsibility, and helps ensure product safety is always a concern for AI operators.
- **Applies to all forms of AI development, and is not limited by sector or industry.** This prevents Small AI operators from trying to hide in unregulated or underregulated sectors, and is necessary given that AI can be used in practically all industries now.[64]

41 By placing significant legal burden and liability onto AI operators, and in particular on AI providers who (historically) have seldom experienced such legal burden, this places the onus on the people who need to be paying the most attention to preventing harms from AI: the developers themselves. This will be an effective and meaningful change, and will assist in creating a cultural shift amongst those actually building Small AI.

42 These outcomes might be achieved through a variety of specific legal tools, some of which we present and discuss in Section C.I.

3. Additional Benefits of a Strict Liability Regime

43 Will a strict liability regime along the lines suggested above stifle innovation? No, we believe that this regulatory framework will reward the best - in the sense of socially most beneficial - innovators of Small AI. It will promote more rigorous testing of such AI, and encourage and commercially justify the diversion of resources to such activities in order to maintain or gain a competitive advantage. We believe this will lead to a safer, and thus better, product for customers, providers and investors alike. Customers are getting more for their money with a safer product, AI operators will find it ultimately more cost-effective and easier to maintain and update a product that has been thoroughly tested from the start, and investors will be at less financial risk when they invest in Small AI.

44 Objectors may argue that such an imposition delays production, reduces functionality, and increases costs, but this is not necessarily the case. Indeed, additional safety measures such as airbags in cars can be seen as an increase in functionality, giving manufacturers a competitive advantage. People are willing to pay more for safety, once they understand the consequences of not doing so. So those in industry who try to argue "No-one else is doing this, so why should we?" will soon find that their competitors *are* doing more to test and verify their product, thus saving on civil lawsuits, and gaining more customers with their safer product.

45 Such a liability system will also incentivise VCs and other investors to make safety, testing, and verification another of their key investment criteria. This turns formerly hidden costs into something public and measurable, thus encouraging companies to articulate, and investors to scrutinise, the

org/10.1038/s42256-019-0114-4 (last accessed Feb. 28, 2024).

63 cf. Flannery, *Safety culture and its measurement in aviation* (2001), https://api.semanticscholar.org/CorpusID:110462980 (last accessed Feb. 29, 2024).

64 Different sectors will naturally have different potentials and/or levels of harm. But the nature and scope of liability does not need to (and often cannot) be specified in advance as it will be highly contextual, while in principle the form and type of restitution or compensation should be relative to the harm caused.

planning, processes and budgeting that will go into additional safety checks. This also helps companies justify to investors, and investors demand of companies, specific resources and money being put towards the safety of the AI. Those companies with the weakest safety processes will pose the most risk for investors, and it is those companies that will be forced to change, or lose out on investment and thus be less likely to succeed. It will ultimately promote investments into companies that genuinely value AI safety and ethics, some of which are currently overlooked in the "arms race" of AI development.

46 With a strict liability regime as suggested in Section B.I.2 in place, this adds additional skills requirements in the corporate hiring process, and thus becomes a quality control mechanism for hiring. Each new hire will either need existing competences in AI safety, or require additional training on the job. It is usually much easier, and cheaper, to hire someone with an existing skill than to train them in it. Moreover, knowing someone has already learned that skill beforehand means it is probably a skill they are interested in and see value in, rather than something the company subsequently forces upon them, so they are likely to be more aligned with its relevance and purpose from the onset. And if this becomes a core part of the hiring process, it will encourage the education of AI developers early on, at university and/or at entry-level AI development jobs. This promotes innovation in education and training, as AI safety becomes a necessary skill for employment.

II. What Can We Do to Address the Risks Posed by Big AI?

1. The Problem of Big AI

47 The regulatory tools proposed above to combat Small AI cease to function *properly* when trying to deal with Big AI. Of course, having some tailored approach within liability law in place is certainly useful, and acts as *some* disincentive for bad behaviour and bad outcomes, so our proposed strict liability frameworks for Small AI operators should *continue to apply* for Big AI operators: just because their AI has a significant impact, does of course not mean they should be absolved of a stringent liability framework. However, as discussed earlier, several further issues arise. It may be difficult for an individual or group to directly challenge a Big AI operator in court, or to establish a causal link between a *specific* harmful event and the operation of Big AI. The harm inflicted by a Big AI might not be substantive enough in relation to a single individual or legal entity to bring a civil lawsuit, but at the same time the scale of the (systemic) damage inflicted on society might be too big to repair or compensate. Overall, the notion of post-harm responses, in particular in the form of liability for harm caused by Big AI would most likely be ineffective, or be pointless. Thus, more proactive and preventive additional measures need to be developed.

48 Unfortunately, forward-predicting how a Big AI operates and interacts with its surroundings is impossible, for several reasons. The training data is likely massive (and proprietary, or at least under the de facto control of few), and the ML training itself is often non-deterministic: repeating the same training on the same data is likely to produce different AIs.[65] While continuous (technical) audits can solve *some* problems,[66] they are fraught with difficulties owing to the nature of how AI is constructed (usually via a large matrix of "weights"). And the feedback loops Big AI may create, by changing bulk human behaviour and then "re-learning" on that, are often impossible to predict or foresee accurately.[67] Hence, any attempt to audit and understand how the AI will behave, pre-deployment, will be of little use. Thus, *real time monitoring* is needed.

49 It may be tempting to carry out such preventative and monitoring measures via "thorough checking" of all outputs of a Big AI. Of course, one cannot check all outputs and decisions by means of human review: a single Big AI might produce billions of outputs daily.[68] As described earlier, creating a separate digital entity alongside the Big AI to check all outputs simply expands the Big AI itself, and creating an independent digital checker would be practically impossible for most jurisdictions. Consequently, any attempt at comprehensive monitoring of the Big AI is probably futile. Therefore, some form of *selective monitoring* is needed instead.

50 At this point it may appear that keeping Big AI safe is impossible, as we can neither fully understand it,

65 Shumailov et al., *Manipulating SGD with Data Ordering Attacks*, 34 Advances in Neural Information Processing Systems 18021 (2021), https://doi.org/10.48550/arXiv.2104.09667 (last accessed Mar. 5, 2024).

66 Mökander, Floridi, *Ethics-Based Auditing to Develop Trustworthy AI*, 31(2) Minds & Machines 323 (2021), https://doi.org/10.1007/s11023-021-09557-8 (last accessed Feb. 28, 2024).

67 For a discussion of potential feedback loops in large-scale algorithms predicting violence see Hogan et al., *On the Ethics and Practicalities of Artificial Intelligence, Risk Assessment, and Race*, 49 Journal of the American Academy of Psychiatry and the Law 326 (2021), https://jaapl.org/content/49/3/326.long (last accessed Feb. 28, 2024).

68 In the case of OpenAI, GPT was receiving 2.5 billion prompts a day as of July 2025. See Silberling, *ChatGPT users send 2.5 billion prompts a day*, TechCrunch (Jul. 21, 2025), https://techcrunch.com/2025/07/21/chatgpt-users-send-2-5-billion-prompts-a-day/ (last accessed Jan. 5, 2026).

nor fully check it. But as discussed earlier, most Big AI harm comes from how it performs *in aggregate*[69] to pose systemic risks. This is particularly pressing in healthcare, where others have already called for a more encompassing approach to explainable AI.[70] Here we can make use of the product-service duality of AI, and view Big AI outputs as discrete items or *products, for the purpose of our regulatory control model suggested below* (that is, treat each output from the AI as a "product", just as one treats hammers manufactured in a factory as products). A Big AI might produce billions of outputs each day, and we wish to ensure that in aggregate these are acceptable. Thus, we can use established quality control mechanisms from manufacturing, where one periodically selects a small sample of objects produced, and then uses human checks on those to infer if the production line as a whole is working well.

51 In summary:

- One cannot expect substantive (compensable) harm to an individual from Big AI;
- One cannot always establish a causal link to personal harm from Big AI;
- One cannot expect that those harmed can afford to sue a Big AI operator;
- One cannot expect that full restitution is possible from suing a Big AI operator for any large-scale societal harm it causes;
- One cannot forward-predict all behaviour and impact of Big AI;
- One cannot conduct a full pre-deployment audit of Big AI;
- One cannot create another AI to test all outputs of a Big AI;
- One cannot test all outputs of a Big AI by hand.

52 In light of the above, we suggest that one needs an alternative approach. In our view, the best option is a constant safety monitoring system; one that is implementable, affordable, and detects problematic behaviour in almost real time. We outline our proposed solution below.

2. A Proposed Solution for Big AI

53 What we propose here is a constant monitoring system with human oversight for Big AI, coupled with a cumulative penalty system if significant problems are detected in its output. This is *in addition* to the liability system described in the previous section. We describe our proposal below. But first, we draw some simple analogies with existing systems in the world that most people are already familiar with.

54 Firstly, the concept of random testing of products in a factory production line is well-understood. For example, a factory filling water bottles has various ways of testing and manually inspecting some bottles, to infer whether the production line is working well. They might select a small fraction of the filled bottles each hour, to look for cracks, over/underfilling, etc. Then, if a certain percentage of these bottles are faulty, OR if the percentage of faulty bottles starts rising suddenly, OR if the percentage of faulty bottles fluctuates each hour throughout the day, it can be inferred that the production line has problems. This process can be formalised with existing statistical techniques in mathematics.

55 The overarching principle is to randomly sample objects, and ask "Is this object satisfactory?" for each one. It might be that the objects should be identical and conform to one single standard (e.g., filling the same bottle type). Or they might be checking several different products being produced together (e.g., different tiles made in the same factory). Or they could be checking completely different "objects" (e.g., a senior academic examiner, randomly sampling essays graded by more junior academics to check they are not over- or under-grading).

56 Second, the concept of a cumulative penalty system based on random sampling is also well-understood. For example, in many jurisdictions speeding cameras operate to catch drivers who exceed the speed limit. These cameras do not have *total* coverage of all drivers at all times, but take sufficient samples that a driver regularly speeding will eventually get caught, at which point, they are issued penalty points. The more often they get caught, the more points they accrue, and the greater the penalty administered (fines, driving bans, incarceration). Such a system aims to change the aggregate behaviour of drivers, by disincentivising speeding and administration of penalties,[71] and by removing drivers who repeatedly

69 For the aggregate effects of search engines see Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (Feb. 20, 2018), https://doi.org/10.2307/j.ctt1pwt9w5 (last accessed Feb. 29, 2024); for the potential aggregate impact of AI on the economy see Bughin et al., *Notes from the AI frontier: Modelling the impact of AI on the world economy*, 4 McKinsey Global Institute 1 (2018).

70 Ghassemi et al., *The false hope of current approaches to explainable artificial intelligence in healthcare*, 3(11) The Lancet Digital Health 745 (2021), https://doi.org/10.1016/S2589-7500(21)00208-9 (last accessed Feb. 28, 2024).

71 Corbett et al., *The deterrent effect of penalty points on speeding*

speed - rather than attempting to detect all instances of speeding. By using a monitoring system for Big AI, we can couple this with such a penalty system for the Big AI *deployers*.[72] The more often the Big AI gets "caught" doing something wrong, the worse the penalty.

a.) An Outline of the Constant Monitoring System with Human Oversight

57 The constant monitoring system has several components and steps, which we outline here. We start with some organisational prerequisites before moving on to the main monitoring loop involving seven steps.

Prerequisites

- The Big AI deployer that needs to be monitored must be informed that it is part of this monitoring, which will be carried out by what we call a 'Check Authority' (CA). The CA's competence and jurisdiction to monitor should have a clear legal foundation (e.g., in a tailor-made regulation) in order to ensure, and if necessary enforce, the cooperation of AI deployers in the monitoring process.

- The type of activity or outputs produced by the Big AI deployer need to be identified, as well as any relevant *existing* laws, standards or benchmarks that these outputs will be evaluated against. This may be, for example, testing if the AI is incorrectly flagging social media posts as hate speech, or scrutinising whether the AI is incorrectly taking down content as copyright-infringing. One will need to decide what is actually being tested for, and whether the concern here is false positives (such as AI wrongly identifying an upload on YouTube as copyright infringing), or false negatives (AI not detecting a copyright infringing upload on YouTube), or both.

- A team of individuals with relevant expertise will need to be assembled to undertake the testing.[73] These may be representing a regulatory body, civil servants of a government office, selected experts serving on boards reviewing the application of industry standards, or may even be judges, or other qualified people. The point is they are testing samples against *existing* laws relevant to the AI output in question (e.g., rules on hate speech, non-discrimination, competitive markets, or copyright), and already have the expertise to make evaluations or judgments on such matters.

Monitoring loop: sampling and testing decisions

1) The Big AI deployer will need to log all inputs, choices, decisions, and outputs it makes for a given time period to be determined. This could be something like "6 month rolling", to give the CA enough time to make all checks and have an appeals process, but not create a substantial data storage burden on the deployer. However, most AI deployers will already be storing their system's user inputs and AI outputs, hence the additional data burden should be negligible.

2) The Big AI deployer will then send through a (salted)[74] *cryptographic hash* (a small "unique fingerprint") of each decision or output to the CA. It preserves privacy rights by giving no information about the original file, but can be easily verified against the original file. This process also helps reduce the amount of data that needs to be transmitted to the CA. Often user inputs and the generated AI outputs will already be used by the companies to improve their product and to make new products. The consent potentially required for such data transfer could be part of the proposed legislation.

3) These hashes will need to be sent to the CA at regular intervals, in batches, by the Big AI deployer; potentially daily or weekly. To protect itself, the Big AI operator might take a "hash of each batch" and commit that to a large blockchain; this creates a very small file that can be stored on a blockchain in an efficient way. If the CA happens to lose some hashes in the future, the Big AI deployer is in a position to refer to the blockchain to verify what data had been sent.

4) The CA then randomly selects a small fraction of the hashes from the batch, and asks the Big AI deployer to send through the original files for checking. The CA cannot see in advance what data lies behind each hash, so cannot cherry-pick the "worst" decisions to test. Similarly, the Big AI deployer needs to hash all of their decisions/outputs, and cannot cherry-pick the "best" ones to send for testing. If it comes

drivers, in Proceedings of the Behavioural Research in Road Safety 2008: Eighteenth Seminar (2010), pages 175-195.

72 Our monitoring proposals are for Big AI *deployers* (not providers), as it is the deployment which determines the scale of impact. We place the onus on the deployers (who are "easy to see"), and they can then pass some responsibility on to others in the AI supply chain through contracting, careful choice of AI providers, setting of minimum performance standards, etc.

73 They act as a "human-in-the-loop". Implementing human-in-the-loop setups in practice come with their own well-studied challenges which we do not discuss here; for details see Chiodo et al., *Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility,* arxiv (Sep. 25, 2025), https://arxiv.org/abs/2505.10426 (last accessed Jan. 8, 2026).

74 A simple additional privacy step, to prevent any reverse-engineering of hashes back to outputs.

to light that a complaint is made against the Big AI deployer for an unfair or harmful decision, the deployer will need to present the output/decision, and corresponding hash, to the CA; if it was not one of the hashes originally sent over then substantial penalties are levied against the deployer.

5) The CA then presents this small fraction of AI outputs/decisions to the trained reviewers (judges, regulators, etc.) for individual scrutiny against existing laws.

6) The results of this human review process, which, in this system, are considered to be *a priori* correct (until proven otherwise), are then saved in a database. The "(time, correct/incorrect)" pair that is the result of the review process forms a time series for which the CA can then perform a change point detection using mathematical statistics.[75]

7) There are many ways to test if a time series has changed, but we outline three common changes that the analysis might look for:

- *Measure a change in the misclassification rate.* This simply measures the percentage of AI outputs/decisions determined as "wrong" by human checkers in a single batch. A threshold needs to be set initially, and if sufficiently many outputs are deemed "wrong", then the Big AI deployer is put on notice or given a penalty. One may introduce multiple thresholds with varying degrees of penalty.

- *Measure a change in the misclassification trend.* This measures how quickly the percentage of overruled outputs/decisions changes from batch to batch. For example, if most previous batches had roughly 1% errors, and then suddenly the errors start quickly climbing in successive batches to 3%, 4%, etc., then even though this might be under the tolerated error rate, the sudden uptick indicates a problem, and the Big AI deployer is put on notice or given a penalty.

- *Measure a change in the misclassification variance.* This measures how much the percentage of overruled outputs/decisions changes from batch to batch. For example, if most batches had between 0.9% and 1.1% errors, and then suddenly the errors in successive batches start fluctuating more wildly (say, 0.2%, then 1.8%, then 1.4%, then 0.3%), then even though this is all within the tolerated error rate, and the error rate overall is not increasing or decreasing much, this more extreme fluctuation may indicate a problem, and the deployer is put on notice or given a penalty.

58 The sampling and testing loop keeps repeating at regular intervals, and the costs for implementing this system should be borne by the AI deployers being monitored. Each time a "change" is detected, some sort of warning, penalty or restriction will be given to the deployer, which we detail below.

59 As part of this review mechanism, the Big AI deployer may appeal any of these human review decisions in a specifically designed dispute settlement mechanism. An updated result is recorded if such an appeal is successful, and the statistical analysis is (easily) recalculated. But given the cost of such an appeal and the fact that it would take a substantial number of human decisions to be overturned to change the outcome of the analysis, this is unlikely to occur often. In addition to this specific dispute settlement system, penalties awarded and related decisions that affect the Big AI deployer are subject to judicial review. We illustrate this whole process in Figure 3 below.

75 Many statistical tests and algorithms, including some machine learning based approaches, have been developed to detect changes in a time series. For an overview, we refer to Aminikhanghahi, Cook, *A survey of methods for time series change point detection,* 51 Knowledge and Information Systems 339 (2017), https://doi.org/10.1007/s10115-016-0987-z (last accessed Feb. 28, 2024).

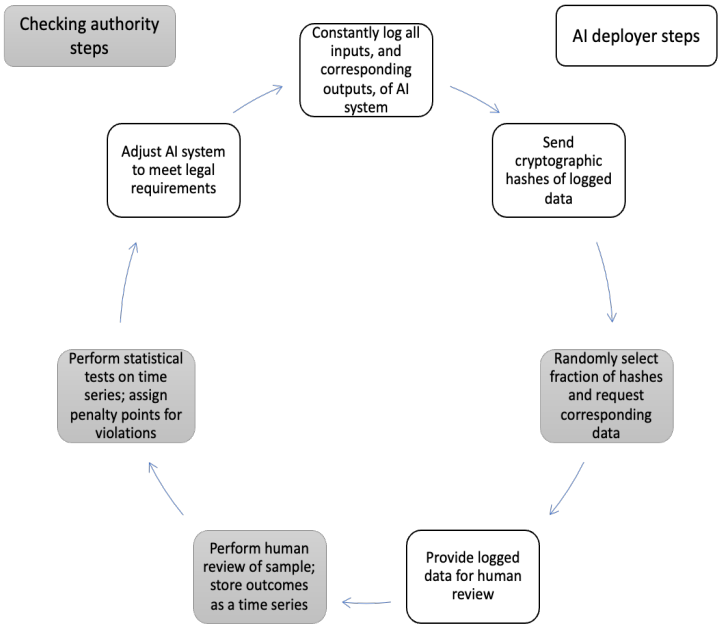


Figure 3: Intuition for constant monitoring with human oversight

Different types of Big AI to monitor

60 There are 3 types of AI that we might want to monitor as part of this Big AI approach, and the way we monitor each changes subtly.

1. Classifiers

61 These AI systems classify items into categories. For example, this could be an AI image recognition machine that scans bottles on a conveyor belt and classifies each as correctly/incorrectly filled,[76] removing the incorrectly filled bottles. So how would we monitor such an AI? From a business perspective, we want the number of bottles erroneously classified as incorrectly filled to be small. Reputationally (and legally), we want to minimise the number of bottles erroneously classified as correctly filled. Finding false positives vs false negatives can come with vastly different incentives and legal challenges. Here, we would independently sample bottles classified as incorrectly filled, and those classified as correctly filled, applying different thresholds for each group.

2. Predictors and recommenders

62 These AI systems make predictions (e.g., "It will reach 18 degrees on Tuesday") or recommendations (e.g., "These shoes will fit you") to help users make a decision.[77] They give more sophisticated outputs than classifiers, may not be falsifiable when the output is made, and are comparatively easy to manipulate.[78] It may require a long wait to observe if the output was correct. The errors produced here might be on a sliding scale, where we worry *how far* from "truth", or *how bad*, the output was. Our approach can still be applied, with a human reviewer identifying if the output was *sufficiently* good - based on either existing standards and legal principles, or those still to be developed. For example, the human might be a shoe fitter, giving their expert opinion against an AI shoe recommender.

3. Generators

63 These AI systems generate new content (such as text, image, video, etc.) based on an input prompt, like ChatGPT or AI generated news articles.[79] Reviewing such output is different from verifying "Is it correct?" as for classifiers, or "How wrong was this?" as for predictors. Understanding if the output was good or bad is often a human-subjective judgement, involving complex questions like "Is this *harmful* in some way?" requiring context-specific value judgements by the reviewer, with no immediate ground truth to check against, or process to falsify the output. These outputs require extensive human judgement, to evaluate the potential or materialised harm from the output, necessitating a human in the loop for any review process. However, as elaborated in Section B.II.4, our proposed monitoring system requires the human to evaluate outputs without significant reliance on contextual material or further external sources. Thus, it is less appropriate for determining harm which cannot be directly inferred from the output, or for evaluating broader societal harm (e.g., to a fair electoral process or democratic decision-making). Due to these inherent limits, lawmakers and regulators might consider adopting

76 Gizaw, Kebebaw, *Water Bottle Defect Detection System Using Convolutional Neural Network, in* Proceedings of the International Conference on Information and Communication Technology for Development for Africa (Nov. 28-30, 2022), pages 19-24.

77 Burke et al., *Recommender systems: An overview*, 32(3) AI Magazine 13 (2011), https://doi.org/10.1609/aimag.v32i3.2361 (last accessed Feb. 28, 2024).

78 Floridi et al., *How to Design AI for Social Good: Seven Essential Factors*, 26 Science and Engineering Ethics 1771 (2020), https://doi.org/10.1007/s11948-020-00213-5 (last accessed Feb. 28, 2024).

79 For a comprehensive overview see Cao et al., *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*, arxiv (Mar. 7, 2023), https://arxiv.org/abs/2303.04226 (last accessed Feb. 28, 2024).

a precautionary approach with regard to AI systems that pose risks of such type of harm, or rely on additional legal and regulatory tools, such as the risk assessments and audits relevant for large platforms under the EU's Digital Services Act (DSA).[80]

b.) An Outline of the Cumulative Penalty System

64 As well as monitoring, there needs to be some "consequence" that the Big AI deployer must face if the AI's performance is deemed unsatisfactory. One main benefit of a penalty system for Big AI, especially if deployed by a large corporate entity, is that many of the standard legal mechanisms and governance tools for controlling corporate behaviour, including fines, restriction of activity, and shutdown, usually *do* work. In general, Big AI deployers (usually comprised of a Big Team) will aim to comply, avoid (significant) fines, and are not prepared to be shut down, unlike Small Team operators who often are more immune to such threats, as discussed earlier.

65 The cumulative penalty system will trigger whenever a "change" is detected in the final (change-detection) step of the monitoring loop discussed above. A cumulative system's purpose is, within the principles and rules of the existing constitutional and other applicable legal orders (for example, the notions of proportionality, reasonableness and due process), to discourage any harmful or otherwise problematic behaviour of Big AI, without being so harsh as to dis-incentivise useful AI developments. Such a penalty system has two main components:

1) The size/severity of each change detected, with an associated number of "penalty points" issued to the Big AI deployer, and

2) The penalty issued to the Big AI deployer as it accumulates such "penalty points".

Penalty points

66 The exact changes and thresholds at which to issue penalty points need to be decided on, in advance. Mathematical insights can assist with making such decisions, but it will also require discourse on what constitutes bad outcomes and behaviour. This may change depending on the scale of the Big AI, the perceived risk to society (based on factors such as likelihood of occurrence and severity of harm) of the application domain and how quickly problems can be rectified (including whether the situation pre-harm can be restored, or whether financial compensation can be awarded). These are technical numbers but require extensive non-technical discussion and understanding. This is where the AI's "application domain axis" can naturally start to play a part, as the application domain often further determines many of an AI's inherent risks, and thus how severe the penalty point system should be (e.g., higher penalty points for riskier application domains).

67 Penalty points need to be proportional to the (potential) harm caused by the Big AI, pre-determined, formulaic, and the details made available to all parties beforehand. The number of points awarded, how long they stay on a deployer's record, and the consequences of accruing various numbers of points, need to be decided on. These decisions are made in tandem with the decisions about the changes to be monitored as well as the relevant thresholds. As indicated above, they have to be taken within the parameters of applicable constitutional and other relevant legal orders governing the conduct of providers of products or services that pose risks, or equivalent corporate behaviour.

Consequences of accruing points

68 The outcomes of incurring penalty points could be various different consequences or punishments, drawn from the range of legal tools available in a given constitutional and legal setting, to penalise corporations and individuals. The idea is that, as this is randomised testing, such consequences need to "ramp up", from very light consequences at the onset, to very severe ones when many penalty points are accrued. These could include:

- Warnings to the AI deployer.
- Public announcements about the AI's poor performance.
- Monetary fines (of increasing amounts).
- Restrictions/limiters on how much the AI is allowed to process (per day).
- Restrictions/limiters on how much trade or activity the deployer is permitted to do.
- Temporary suspension of the AI.
- Permanent suspension of the AI.

69 In order to "have some teeth", compliance with the consequences imposed should be monitored and if need be enforced. Failing to comply should be subject to responses that are equivalent to contempt of court.

70 At the same time, such penalty points should not be "lifelong", but rather should drop off or decay

80 See the discussion in Section C.IV below.

over (operation) time, so that a Big AI deployer is incentivised to, and rewarded for, improving performance. Overall, our suggestion is a proposal for measures that aim to prevent harm, rather than those that retroactively punish the occurrence of harm, administer recourse, or provide compensation. Thus, it should be recognised if the Big AI deployer has taken steps to reduce risk and harm, i.e., to enhance AI's safety. Any restrictions on the functioning of the AI system, or on the deployer itself, therefore should be lifted gradually.

Legal and social balance

71 Existing laws relevant in the domain the Big AI might operate in (e.g., large social media platforms and the requirement for the removal of hate speech) can be one-sided in terms of the aims pursued. They may focus on one specific objective, such as the prompt removal of hate speech, through public laws (such as the German Netzwerkdurchsetzungsgesetz, NetzDG) or even the criminal justice system. But they may leave impact on other interests and potential consequences (e.g., false positives and the accidental suppression of valid speech) as a matter for affected individuals to pursue in front of domestic courts. Thus, such Big AI deployers are inclined to unduly focus on one aspect (avoiding false negatives) to avoid potentially severe state sanctions, and be content to overdo it in the knowledge that there is little, if any, consequence for them in the other direction (creating false positives). For reasons discussed previously, we believe many Big AI deployers and platforms are not concerned enough by individual complaints brought against them, even if taken to domestic courts.

72 Our suggested approach allows a way to balance out how a Big AI deployer approaches problems. Having a state-enforced monitoring system collectivises the way society responds to the operation of the Big AI. At the same time, imposing the system's costs on the AI deployer ensures that risk prevention is appropriately internalised. Collective means of seeking redress from the AI deployer or platform via litigation (for example in the form of a class action lawsuit) is not a realistic option in many jurisdictions. And even where collective redress mechanisms are principally available under the domestic legal system, they are unlikely to succeed, or even occur, as the impacted and harmed parties would be hard-pressed to discover one another. A state-enforced monitoring system gets around this limitation and can be designed to incentivise (and if need be, pressure) the Big AI deployer to take account of all relevant interests and societal concerns that can be checked for in the monitoring process, and not just focus on complying with those legal requirements that appear most consequential, or are most likely to be invoked by the state or other actors it is most fearful of.

73 As well as there being stronger legal incentives on one side of the outputs that a Big AI deployer is producing (i.e., false negatives), there may also be enhanced social, regulatory or legal mechanisms that can address, or at least identify and highlight publicly, these same types of problematic output. If the Big AI operator consistently generates harmful output (for example, if an automated filtering system allows a lot of illegal content to be viewed by users on a social media platform), users will eventually complain and the media as well as existing regulatory authorities may pick it up, even if such occurrences are rare. This can lead to a similar adjustment of behaviour in the Big AI deployer that in turn then results in biased compliance. Again, few effective mechanisms exist to detect and/or address over-enforcement (e.g., in the form of the removal of otherwise legal content); there is unlikely to be media attention or public outcry if a legally-compliant video is removed from a platform. This is particularly true if the individual enforcement act is unlikely to be considered harmful in isolation (i.e., a platform removing a short video), even though such enforcement action may well, in aggregate, severely influence and distort the balance of interests within society, and might even shape societal behaviour (e.g., by normalising the absence of a specific type of content or expression that - while legal - tends to be caught in hate speech filters).

74 We suggest that our approach is particularly suited to identify and even prevent AI-induced imbalances in relation to different societal interests for which there is no existing mechanism to uphold the envisioned balance of interests - either because state enforcement mechanisms incentivise AI operators to prioritise a particular interest, or because the surrounding social and economic context serves to achieve the same result.

75 Of course, the sorts of errors the CA might try and check for will depend on *how* the Big AI is deployed; the fact that monitoring needs to happen is by virtue of the impact size of the AI, but *what* needs to be monitored for will be determined by where and how the AI is deployed. One might counter by asking why, then, are we proposing this additional burden on AI deployers *only* by their impact size? If a Small AI were being used in a "high-risk" situation such as in medical diagnostics or treatment, why would that only fall under our liability framework, whereas a seemingly low-risk Big AI such as a spam filter would (perhaps counterintuitively) attract an additional monitoring obligation? The reason here is that practically every Big AI will pose some *systemic* risk to society - in a way that Small AI is unlikely to do. Consider these examples: a medical Small AI will, by definition, only impact a (relatively) small number of

people, and while it may still cause significant harm to them, this will not pose a systemic risk to society. Conversely, a (malfunctioning) Big AI being used as a spam filter may let through billions of fraudulent or phishing emails, leading to recipients handing over or losing money to the spammers. If 1% of a billion recipients fall victim to this, on the scale of $1000 each, that would be a (relatively) small harm for each victim, but would raise $10bn which might go towards an organised crime syndicate, leading to their significant empowerment and additional harm to society through other sophisticated criminal activity given their new-found wealth. Here we see that the risk to individuals is relatively small, but the combined risk to society is enormous.[81]

3. Technical Advantages of This Approach

76 Our proposed regulatory framework using constant monitoring with a human-in-the-loop[82] provides several technical advantages over other existing suggestions (e.g., an AI checking other AIs). These include:

- Our suggestion is based on a mathematically mature and active research area, particularly when compared to the idea of AIs checking other AIs. Sampling and change point detection are both well-developed fields: early works of change point detection can already be found in the 1950s,[83] and many experts can help build such a system if they are given the legal framework and requirements. *[The approach builds on an "existing technique in industry".]*

- The approach is computationally efficient: in comparison to building and constantly retraining AIs to check AIs, it has comparatively low energy cost and requires minimal technology. *[The approach "does not require expensive technology."]*

- Because the review result used as the "ground truth" in the statistical analysis is figured out by human review, it is explainable and interpretable, and the mathematical tests and sampling are verifiable. *[The approach is "explainable, interpretable and verifiable."]*

- Human in the loop: the human can spot patterns and raise additional alarms. Tacit and expert knowledge is not neglected (as would be the case for a purely technical solution) but actively used and encouraged. *[The approach "retains a human in the loop."]*

- The technical system requires little maintenance and (almost) no "retraining". As long as the threshold parameters do not change, we can simply let the system run. When a company appeals decisions, the test statistic of the statistical analysis needs to be recomputed, but this task is simple and could be automated. *[The approach "requires minimal maintenance."]*

- Changes can be detected purely based on the output, and independent of the AI: it does not matter if the operator updates its AI, and the system does not need to know anything about the internal operation of the AI. *[The approach is "technologically neutral."]*

- The system is relatively easy to adapt to new laws and regulations. For example, when the definition of hate speech changes, the new law only affects the human reviewers who need to be retrained for the new laws and regulations; the mathematical analysis does not need to change. This is not true for "AIs checking other AIs": such AIs will need to undergo much more expensive retraining when the definition of hate speech changes. Our solution only needs to re-adjust the human review mechanism. *[The mathematics is "law neutral."]*

- The approach is easy to adapt to different countries, cultures, values and norms, because different jurisdictions or areas that have their own laws can simply choose their own parameters and thresholds for what is considered unacceptable. *[The approach is, as a matter of principle, "culturally and jurisdictionally neutral."]*

- The system can be adapted for different risk classifications and different problems. While the relevant thresholds, parameters, etc., need to be changed, the overall architecture stays the same. *[The approach is "risk neutral."]*

- The parameters required for the statistical analysis and the human analysis can be chosen without extreme difficulty. It is not easy, but it is not nearly as impossible as "building an AI that can check another AI". Any such AI would also require us to answer many more technical, legal and moral questions. *[The approach is "not hampered by extreme difficulties."]*

- The mathematics and the human checks are decoupled: this means that for different problems different mathematical tests can be used, so the

81 One can only imagine what harm an organised crime syndicate could inflict with $10bn. The example of prohibition in the US in the 1920s brought rapid and vast wealth to organised crime syndicates, enabling them to flourish and establish themselves for decades to come.

82 For details on what a human-in-the-loop setup is, see Chiodo et al., *Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility,* arxiv (Sep. 25, 2025), https://arxiv.org/abs/2505.10426 (last accessed Jan. 8, 2026).

83 e.g., Page, *Continuous inspection schemes*, 41(1-2) Biometrika 100 (1954), https://doi.org/10.1093/biomet/41.1-2.100 (last accessed Feb. 28, 2024); *Control charts with warning lines*, 42(1-2) Biometrika 243 (1955), https://doi.org/10.1093/biomet/42.1-2.243 (last accessed Feb. 28, 2024).

approach is easy to adapt in the future. *[The approach is "future proof and adaptable."]*

- The mathematics can be implemented: this is not a multi-billion parameter AI that only a few people can implement. This is feasible to build, even in a jurisdiction with limited resources to spend. *[The approach is "feasible to implement."]*

- There are not too many Big AI deployers (compared to Small AI operators), which reduces the effort needed. And the monitoring can scale up/down smoothly depending on the size of the AI; more frequent monitoring (daily batches) for very large, and less frequent for smaller (monthly). *[The approach "scales easily with the number and size of such AI systems."]*

4. Limitations of This Approach

77 The "incorrect", or "problematic", decisions need to occur sufficiently often for this approach to be effective. If only a very small fraction of decisions being tested is incorrect, then a sampling-based approach would require a large sample, in order for the approach to work well. The exact definition of "very small" will depend on the application domain, the enforcement capacity of the Check Authority (CA) and other factors.

78 As already hinted at above, our constant monitoring approach relies on the human reviewer being able to evaluate the outputs and decisions of the AI shortly after they happen. The longer the time lag between the AI generating its output and the consequences of that output manifesting, the greater the time lag in our approach, thus making the approach less effective as a preventative measure. A scenario where there is a significant delay between the AI producing an output and it being determinable that that output was undesirable, leaves the AI with a long window to keep generating undesirable outputs before the review mechanism is able to detect and respond. That is, the AI will keep running and keep producing bad outputs until outputs are realised to be bad. In summary: the system works best if the AI decision can be reviewed and is falsifiable *quickly,* and the false positives and/or false negatives are sufficiently dense in the set of decisions.

79 That said, there will be scenarios where Big AI deployers choose to use their AI in settings where the output is unfalsifiable. Examples could include uses in the criminal justice system, where AI systems might be used to determine the chance of a convicted criminal reoffending, and thus factor into a sentencing process, as happened in the US with the system known as COMPAS.[84] Given that the "false positive" side of this decision process is completely unfalsifiable (an incarcerated criminal lacks the opportunity to "prove" they would not have reoffended if released sooner), our constant monitoring approach would not be applicable for such decisions, as there is no "absolute truth" to refer to. In *very limited* circumstances it may be possible to adjust our system, and have the checker see if the AI followed or violated certain procedural standards. However, since advanced AI systems often reason differently from humans, such examples may be limited to specific interpretable and explainable algorithmic decision systems, such as those using decision trees. But in general, it is inadvisable to make use of AI in unfalsifiable scenarios such as the one described above, as one of the (many) criteria for responsible and safe deployment includes falsifiability.[85]

80 However, humans frequently make unfalsifiable decisions. If an AI were to be used to replace a human making unfalsifiable decisions (the authors do not advocate for this), then nonetheless our constant monitoring system could be useful. Rather than checking against some sort of "absolute truth", the checker might make their own expert determination of the decision, to compare to the AI output. Thus, our system can be used to ascertain if the AI is, *on average*, performing as well as a human in the same situation.[86] This can also be refined, for example, in our criminal sentencing example, it might check how effectively an AI is performing on certain demographics of particular concern.

81 Our proposed constant monitoring system cannot answer all possible questions about the (mal) function of the Big AI. It is very effective, and quick at detecting *if* there are problems with the AI but does not necessarily explain *why* such problems have occurred. For example, it cannot answer questions such as:

- Why was there a change in the AI output?

- Why has the system broken down?

- Who is affected by the system?

82 However, we assert that it is the responsibility of the Big AI deployer to figure out and rectify such problems, just as it is the responsibility of car drivers

84 Rudin et al., *The Age of Secrecy and Unfairness in Recidivism Prediction*, 2(1) Harvard Data Science Review (2020), https://doi.org/10.1162/99608f92.6ed64b30 (last accessed Mar. 27, 2024).

85 Floridi et al., *How to Design AI for Social Good: Seven Essential Factors*, 26 Science and Engineering Ethics 1771 (2020), https://doi.org/10.1007/s11948-020-00213-5 (last accessed Feb. 28, 2024).

86 Noting that better average-case performance does not imply better performance in each instance.

to understand what they are doing wrong in their driving and change their behaviour accordingly. If the Big AI deployer is incapable of identifying the source of such problems and/or is unable to rectify them, or even goes so far as to claim that such actions are not possible, then this brings into question whether this AI should be permitted to continue to exist and operate. After all, it is the responsibility of those deploying such high-impact AI to ensure their tool is safe, rather than it being left to the rest of society to fix the problems (or pick up the pieces when harm is done). The overriding principle here is simple: If you cannot use your technological solution *safely,* you should not use it at all.

83 Lastly, the monitoring undertaken by the CA most likely cannot appropriately test for more diffuse risks for broader societal interests such as a functioning democracy, a system free from bias or from systemic discrimination of specific groups. As already indicated above, this is because the legal principles and rules which operate to uphold such broader interests are complex and usually will require significant and careful balancing of various interests, detailed and fact-specific determinations, and usually involve quite some discretion on the side of the (usually judicial) body in applying the relevant rules and principles to the concrete facts at hand. While all of this of course forms part of what trained lawyers (especially judges) do, this process requires time to hear arguments from all sides, proper representation of the interests concerned, experience drawn from similar cases, expert evidence on the facts and careful deliberation on the side of the decision-maker. It would be illusory to expect this to be replicable at scale in the context of constant monitoring, even if undertaken by trained experts on the point of law at issue. While we refrain from drawing an exact hard line on the types of legal issues that can be reviewed by a CA, the principle set out above stands: the more an AI output poses a concrete systemic risk of harm to clearly understood protected interests such as (human, animal or plant) life or health, to property, or other interests legally recognised in concrete and defined terms, the easier this can be made subject to our proposed monitoring system. On the other hand, the vaguer or more diffuse the legal interest at stake is and the more remote or indirect the risk posed by the AI system, the less helpful checks along the lines we propose will be. Alternative mechanisms, such as the comprehensive risk assessments by and external audits for very large online platforms and search engines under the EU's Digital Services Act (DSA),[87] may be considered to address more remote or indirect risks and/or potential harm to more diffuse and vaguely defined legal interests.

5. Additional Benefits of Constant Monitoring with Penalty Points

84 Is constant monitoring with penalty points a reasonable imposition for Big AI deployers? We believe that our approach provides more flexibility than existing regulations (e.g., the GDPR, or the EU AI Act) to control AI and data. Those regulations do not permit failure, in the sense that a single violation can lead to significant penalties for an AI operator. By comparison, constant monitoring allows, and even expects, some degree of failure, within certain tolerated levels. And even after a "failure", the penalty point system is cumulative, and early on might only inflict something as minor as a warning to the company. This is much less arduous on companies than regulations that require 100% adherence to stated rules with significant penalties for even the smallest violations, thus giving the organisation and its developers room to breathe and learn.

85 Our proposed system with constant monitoring and associated penalties is also much more likely to be accepted by those deep within the AI development and deployment process. Unlike the very rigid emissions testing regime for nitrogen oxides (NOx) which was routinely cheated by car manufacturing companies around the world, our proposal breaks down the barrier of acceptance of failures and enables developers to discuss failure openly. The regulatory constraints are no longer clear "lines in the sand" which cannot be crossed, but rather performance objectives to try and stay within as often as possible. The penalty points system actually gamifies the regulations, and helps developers see these as their own scoring system for how well they have done, turning failing into an opportunity to improve. This can help instill a workplace culture of seeing the regulations as an aid and working with them, rather than as a solid barrier and trying to work against or around them.

86 Existing regulations are often outcomes-based punishment systems, with very little feedback to developers. Instead, our monitoring + penalty points system can be viewed as an institutional training regime, supporting trial and error iteration for product improvement. This promotes innovation by creating a mechanism for constant improvement. It assists with the survival of the company by detecting problems that may lead to disastrous outcomes, before they manifest. And it fosters improvement and socially beneficial innovation by giving even ethically untrained developers and managers a starting point to build from and a long-term

87 For a discussion on how the legal tools of the DSA might serve as a useful framework for implementing the monitoring system considered here, see Section C.IV.

objective to aim for. Indeed, this system provides developers with incremental, external feedback on their product's performance, which is something they might otherwise have no access to. It helps them, and their product, improve, through a form of reinforcement learning for their own institution and its people. And if explained to developers in this way ("reinforcement machine learning on machine learning developers"), they will almost certainly respond well to it as a useful method for improvement.

87 Our suggestions here also stand up well to regulatory evaluation criteria set by others as well. What we have provided is an innovation-*friendly* approach to regulation, as discussed in the 2020 report of the UK Department for Business, Energy & Industrial Strategy.[88] Our approach of monitoring and penalty points matches well with 4 of the 5 key criteria set out in that report (cf. pp. 5-6). Namely, our approach:

- Provides regulatory advice to innovators,
- Supports experimentation and testing of innovations,
- Streamlines regulatory approvals for innovators, and
- Sets regulatory challenges to drive innovation.

88 Our proposed regulatory framework is as business-friendly as anyone could possibly hope for, and simultaneously one of the best possible ways to maintain the large-scale safety of Big AI. It can keep society safe, and keep businesses engaged.

C. Potential Solutions from a Legal Perspective

89 In this Section, we elaborate how the systems described in Section B can be anchored in existing and upcoming legal frameworks, resulting in deployment that aligns with existing legal systems and tools to the maximum extent possible. For this purpose, it is examined whether and how the strict liability regime suggested above for both Small and Big AI could be introduced, and whether the constant monitoring and penalty system for Big AI is reflected in existing liability law, market regulation and regulatory law. Within the analyses of these areas of law, random testing and the penalty points system are considered separately from one another, since from a legal perspective they constitute obligations of different types and intensities.

I. Amending Liability Rules (for Small and Big AI)

1. Introduction to Liability Law

90 Liability law operates as an indirect approach to regulation.[89] In contrast to direct ex ante regulation, imposing a set of obligations which must be met before an AI system may be put on the market (see Section C.V), there are no immediate instructions for action and hence no direct command to implement a monitoring system. The regulatee can, however, be held liable for damage that occurs. Assuming that the regulatee as a rational actor has an interest not to offer reparation or pay damages and hence aims to avoid liability, by establishing liability rules, the regulator influences the regulatee's behaviour indirectly by creating an incentive to act lawfully in order to avoid liability. In this context, a system of constant monitoring could hence be adopted as a prerequisite for compliance with the duty of care - so that monitoring becomes a necessary element to avoid liability. Furthermore, looting strategies with small entities prone to bankruptcy in case of liability can be avoided by combining liability rules with compulsory liability insurance (third party insurance).

91 It should be emphasised that our contribution in this article is *not* to offer new solutions or proposals for tort liability within Law (apart from suggesting strict liability for certain AI systems), but rather to situate our suggestions for Small and Big AI within the existing legal framework and examine whether this framework can accommodate our suggestions set out in Section B.I. For this analysis, it is essential to differentiate between different liability regimes: fault-based liability, presumed fault and strict liability which are briefly introduced below.

a.) Fault-Based Liability

92 Fault-based liability means damage is compensable only if the injured party proves it was caused by the injurer's wrongful behaviour, such as failing to observe a duty of care. Random testing, as part of the

88 UK Department for Business, Energy & Industrial Strategy, *Regulator Approaches to Facilitate, Support and Enable Innovation. Final Report. BEIS Research Paper Series, Number 2020/003* (Jan. 2020), https://assets.publishing.service.gov.uk/media/5e2efd76ed915d1f26a7427f/regulator-approaches-facilitate-support-enable-innovation.pdf (last accessed Feb. 28, 2024).

89 Zech, *Liability for AI: public policy considerations*, 22 ERA Forum 147 (2021), https://doi.org/10.1007/s12027-020-00648-0 (last accessed Feb. 28, 2024). In California, the new AI Bill explicitly mentions civil penalties and monetary damages.

monitoring system in Section B.II.2, could form part of this duty. If an operator neither monitors its AI nor performs random tests, it breaches this duty—creating an indirect obligation to test.

93 Applying fault-based liability to AI is challenging. The opacity and complexity of AI systems and their supply chains make it hard to prove defectiveness, causality, or wrongful conduct. For instance, a user receiving unsuitable insurance recommendations may not know whether AI was involved or where the error originated. Legislators or courts could address this via rebuttable presumptions of fault (see Section C.I.1.b).

94 AI's learning capacity further complicates attribution. An autonomous car might cause damage through behaviour learned independently, making it difficult to link fault to the provider. Proving fault under current law is thus often impractical. Small AI adds hurdles of invisibility and inaccessibility: startups may supply AI embedded in other products, leaving users unaware of the actual operator. While similar issues exist outside AI, the stakes are higher here (see Section B.I.1).

95 Big AI introduces power imbalances. Injured parties rarely have resources for lengthy, expert-heavy litigation. Class actions can help by pooling claims, but they falter when harm is systemic, minor, or non-pecuniary, and jurisdictional issues often arise. While class actions may incentivise Big AI operators to internalise harm—especially when paired with presumed fault or strict liability—they cannot resolve all challenges. Preventive measures like constant monitoring remain essential.

b.) Presumed Fault

96 In order to solve the aforementioned problems, tort law at times turns the burden of proof so that in specific scenarios, the injurer's (in this case the operator's) fault is presumed. Even in cases where fault is presumed, damages are only compensable if it is proven that these damages were caused by the injurer's reproachable or wrongful behaviour. It is, however, not presumed that the injurer's wrongful behaviour has led to the damage. Effectively, with this mechanism the burden of proof is shifted from the injured party to the alleged injuring party which needs to show that it has complied with relevant obligations, such as a duty of care. In order to escape liability, the alleged injurer therefore can make use of an opportunity for exculpation, by offering evidence that it has complied with the relevant duty of care (i.e., by showing that as AI operator, it did perform the required random testing).

97 With regard to presumed fault, however, three central issues remain unsolved: First, the above-described problems associated with power imbalance, invisibility and inaccessibility are not tackled. Second, the burden of proof only shifts for one aspect; all other aspects (such as that the damage has been caused by AI) must still be proven by the injured party. However, an additional presumption of causality might be introduced (like the new liability Directive envisages for complex products)[90] - which in turn would lead to a significant further reduction of the requirements for liability. While AI operators might of course argue such extensive presumptions of liability would create significant disincentives for the development of new technologies, we would counter that extended liability presumptions can actually incentivise more socially-beneficial technological innovations, and prevent a 'race to the bottom' in terms of safety and security.[91] Third and relatedly, the injured party will usually not have access to the relevant information and might not have sufficient expertise to be able to assess how and what caused the damage.

c.) Strict Liability

98 Strict liability makes damage compensable if caused by the operator's AI, regardless of wrongful behaviour or due care. It is an exception to fault-based liability, applied in hazardous contexts or where risks commonly materialise (e.g., car operation or dog ownership). Its justification lies in the risks of providing AI systems. Key advantages include internalising external risks and enhancing legal certainty. Since liability applies even when due care is observed, operators must weigh benefits against risks before market entry. Legal certainty improves because proving wrongful behaviour becomes unnecessary.

99 Strict liability signals that AI operators are accountable for AI-caused harm. It does not necessarily deter AI development—goods under strict liability remain widely used; people still drive cars and own dogs. As long as damages are insurable, strict liability is no disincentive. For businesses, legal certainty is crucial for profitability. Because due diligence is irrelevant here, random testing becomes voluntary. Operators will adopt monitoring or testing mechanisms they find cost-effective to reduce errors. Insurance providers may require such mechanisms as conditions for coverage, creating incentives for safety optimisation. If these mechanisms effectively reduce harm, they serve both operators and insurers.

90 See art. 10 sec. 4 Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, hereinafter: Product Liability Directive.

91 See the discussion in Section B.I.3 and B.II.5.

d.) Recent Legislative Activities

100 In September 2022, the European Commission (EU COM) published a Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive).[92] However, on 11 February 2025, the AI Liability Directive was withdrawn. The AI Liability Directive was intended to refer extensively to and complement the AI Act (see Section C.V.3). It is not yet foreseeable whether there is going to be a specific European liability regime for AI, and what it will look like in the future.

101 The 2022 Proposal shows that the EU COM did not envisage strict liability for AI, but intended to design a complex system of disclosure obligations and rebuttable presumptions of non-compliance and causation. The Draft Directive sought to lay down rules on the disclosure of relevant evidence by those that have it at their disposal, for the purposes of establishing liability. This measure was intended to provide an additional incentive to comply with the relevant requirements laid down in the AI Act to document or record the relevant information. In addition to that, potential claimants would have a right to request a court to order the disclosure of relevant evidence before submitting a claim for damages, if the potential claimant presents facts and information sufficient to support the plausibility of a claim for damages and it has made a prior disclosure request to the defendant. In case the defendant does not disclose the relevant information it is rebuttably presumed that the respective obligations under the AI Act have been breached. In addition to that, the causal link between the fault of the defendant and the output produced by the AI system is rebuttably presumed. The causal link between the AI's output and the damage must, however, still be proven by the claimant.

102 The proposed AI Liability Directive was to follow a minimum harmonisation approach which allows claimants in cases of damage caused by AI systems to invoke more favourable rules of national law. EU member states could, therefore, maintain reversals of the burden of proof under national fault-based regimes, or strict liability regimes. However, since the proposal has been withdrawn it is futile to speculate at this point as to what liability rules are likely to be developed by the EU and other actors, and how they will apply to AI in the future.[93] If anything, the discontinuation of the EU's draft AI Liability Directive suggests that the current pro-AI innovation climate makes the introduction of express presumptions (which lead to stricter liability standards for AI) politically not feasible.

2. Implementation of a Strict Liability Regime

103 As outlined in Section B, for Small AI the introduction of a liability relying solely on causation might be a necessary and sufficient way of addressing the associated risks. Such an approach was taken by the European Parliament's proposal[94] for the introduction of a new liability regime for AI.[95] However, the European Commission focused on an extension of the existing product liability (which relies on defectiveness) to AI, instead of introducing a strict liability rule.[96] Here, only some of the issues associated with the introduction of a strict liability regime shall be presented.

a.) Definition of Specific Risk Sources

104 The first problem is to properly define the relevant AI systems, or in general the relevant risk-causing activities, that trigger the strict liability. The EU proposal relied on high-risk AI systems to trigger strict liability (other AI systems being subject to fault-based liability with a presumption of fault). Similarly to the AI Act, the question of which

92 European Commission, *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*, COM(2022) 496 final (Sep. 28, 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496 (last accessed Nov. 12, 2025).

93 In 2022, the UK's Financial Conduct Authority (FCA) released a new Consumer Duty (PS 22/9) which sets out rules for financial institutions. Amongst other things, the Duty obliges financial institutions to mitigate harms and allows authorities to check whether these obligations are met and, in case they are not, to impose fines. With their new Consumer Duty, the FCA imposes a strict liability regime for severe harms as well as a duty to take mitigating measures to avoid liability. For more information see UK Financial Conduct Authority, *PS22/9: A new Consumer Duty* (Jul. 27, 2022), https://www.fca.org.uk/publications/policy-statements/ps22-9-new-consumer-duty (last accessed Feb. 28, 2024).

94 Annex B of the European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence, 2020/2014(INL), 2021 O.J. (C 404) 107.

95 cf. Zech, *The European Parliament proposal for a regulation on liability for the operation of artificial intelligence systems*, 14 Zeitschrift für Geistiges Eigentum/Intellectual Property Journal 293 (2022), https://doi.org/10.1628/zge-2022-0022 (last accessed Feb. 27, 2024).

96 See below Section C.III; cf. Wagner, *Liability Rules for the Digital Age – Aiming for the Brussels Effect* –, Working Paper No. 10, Forschungsinstitut für Recht und digitale Transformation (2023), https://www.rewi.hu-berlin.de/de/lf/oe/rdt/pub/wagner-liability-rules-for-the-digital-age-1.pdf (last accessed Feb. 28, 2024).

AI systems fall under the high-risk category is answered by referring to an Annex which can be amended by delegated acts (i.e., by the European Commission being the executive body without further involvement of the legislative body of the European Union).

b.) Determination of the Responsible Parties

105 Another issue is to determine the parties responsible for damages caused by the AI (aside from proving the causal link between the use of an AI system and the damage). This can be either the provider (programmer, trainer) or the deployer of such a system. The Parliament's proposal focuses on the deployers.[97] However, an argument can be made that with increased automation and autonomy the deployer's influence on the behaviour of a system diminishes. Therefore, rather the provider of such a system should be addressed by a strict liability rule (liability following risk control).

c.) Compensable Damages

106 The third issue to address is the scope of the damages that may be compensated. Like with most strict liability rules, in exchange for a mere requirement of causation, the scope of liability in the Parliament's proposal is limited to personal injury or damage to property. This means that sole (or pure) economic losses are not encompassed. For example, a claimant might be compensated for damage to their health caused by AI, but not for missing out on profits otherwise earned, or for not being considered for a job. Having said this, compensation for personal injuries explicitly includes substantial non-material damages (cf. art. 6 sec. 1 lit. a Product Liability Directive: "medically recognised damage to psychological health"). This still excludes injury to a person's honour (as in the case of defamation caused by AI), but definitely includes psychological injuries suffered through AI (e.g., caused by a chatbot's responses).

3. Product Liability for AI

107 The EU is set to go another route: extending the existing product liability to software and AI. This is a central part of the recent amendment to the existing Product Liability Directive which focuses on the producer's responsibility.

a.) AI as a Product

108 Whereas under former product liability law only movables (and electricity) can be subsumed under the definition of a product, this changes under the new rules. According to art. 4 sec.1 of the new Product Liability Directive "'product' means all movables, even if integrated into another movable or into an immovable. 'Product' includes electricity, digital manufacturing files, raw materials and software". Although it may be argued that AI systems do not constitute software, recital 13 of the Product Liability Directive clearly indicates that at least the *applications* of AI systems are intended to be included ("[...] Software, such as operating systems, firmware, computer programs, applications or AI systems, [...]").

b.) Defectiveness of AI

109 Product liability relies on the defectiveness of a product (in addition to the causation of a damage) as the trigger for liability. Defectiveness might be seen as a lack of diligence on behalf of the producer in disguise.[98] According to the new Product Liability Directive, a product shall be considered defective when it does not provide the safety which the public at large is entitled to expect. This has to be proven by the claimant. However, the burden of proof is eased because the claimant only needs to prove the objective defectiveness instead of the underlying behaviour of the producer. When it comes to AI, proof of defectiveness might be problematic due to the lack of transparency inherent in many AI systems (see Section B.II.1). The new Directive addresses this problem by introducing a presumption of defectiveness under certain circumstances (as set out in art.10 sec. 2). In addition, the causal link between defectiveness and damage should also be presumed where the damage caused is typically consistent with the defect (art. 10 sec. 3). Finally, in cases where proving the defectiveness or the causal link presents excessive difficulties due to technical or scientific complexity, a presumption under art. 10 sec. 4 comes into play.

97 For the precise determination of the responsible parties with the concept of frontend operator and backend operator see Zech, *The European Parliament proposal for a regulation on liability for the operation of artificial intelligence systems,* 14 Zeitschrift für Geistiges Eigentum/Intellectual Property Journal 293, 298 (2022), https://doi.org/10.1628/zge-2022-0022 (last accessed Feb. 27, 2024).

98 Wagner, *Robot Liability, in* Liability for Artificial Intelligence and the Internet of Things 27-62 (Lohsse et al., 2019), https://doi.org/10.5771/9783845294797 (last accessed Feb. 28, 2024); Zech, *Liability for AI: public policy considerations*, 22 ERA Forum 147 (2021), https://doi.org/10.1007/s12027-020-00648-0 (last accessed Feb. 28, 2024).

c.) Compensable Damages

110 Similar to strict liability regimes, compensable damages are restricted to material losses (cf. art. 6 sec. 2 of the new Product Liability Directive), meaning death or personal injury (including medically recognised harm to psychological health), harm to property and, as a new addition, loss or corruption of data. Pure economic losses, resulting for instance from discrimination against the injured party, are not covered. This is one of the important remaining functions of national tort law where under certain circumstances pure economic losses might be covered as well.

d.) Product Liability and the AI Supply Chain

111 Product liability addresses the producer of AI systems. This regulatory approach therefore shifts the focus of legal responsibility from the deployer (user) level to the provider level. With regard to the AI supply chain, this constitutes an important shift of responsibility upstream. Furthermore, with the responsibility of producers of AI as a component (within the meaning of product liability law), responsibility is extended even further upstream along the supply chain. This extension may even encompass providers of AI models, who supply foundational AI models or systems that are subsequently integrated into downstream products or services. Such an upstream allocation of liability reflects the principle that those who design and develop AI systems are often best positioned to identify and mitigate potential risks inherent in their products, and should therefore bear corresponding legal responsibility.

112 This upstream shift in responsibility has significant implications for the distribution of liability along the AI supply chain. While deployers remain subject to obligations under the AI Act and may incur liability for their own conduct in deploying AI systems, the product liability framework ensures that providers and component manufacturers cannot escape responsibility by contractually transferring risk downstream. This approach aligns with the broader policy objective of incentivising safety and quality assurance at the source of AI development, where technical expertise and the capacity to implement effective risk mitigation measures are typically concentrated.

113 In conclusion, the existing liability framework - consisting of the amended product liability and general national negligence liability - allows an implementation of our suggestions on tort liability as both - defectiveness of AI placed on the market by economic operators and negligence of AI providers and deployers - rely on breaches of duty. Our suggestions for certain monitoring obligations can be implemented as a part of these duties of care. However, this does not encompass our suggestions for strict liability in certain cases which is a legislative proposal.

II. Constant Monitoring with Human Oversight as a Part of Diligence in Liability Law (for Big AI)

114 We continue by examining how a system of constant monitoring with human oversight (as suggested in Section B.II.2) can be a part of due diligence under existing liability law and might be applied as part of due diligence duties within a framework of liability law. In this context, we analyse what legal consequences arise from not implementing the monitoring system at all versus implementing it and still causing damage.

1. First Case: No Implementation of the Monitoring System

115 In case AI developers do not implement the monitoring system described above at all, there is in our opinion a strong case that - unless alternative means to effectively address potential harm caused by AI are implemented at the output level - they act negligently and can, subject to other requirements for liability, be held liable. Scenarios of non-implementation (and adopting no alternative means to prevent AI harm) can therefore be considered a relatively clear and unproblematic case from a legal perspective. Assuming the injured party can show that the damage it suffered has been caused by the AI provided, under both fault-based and strict liability, the developer would be liable.

2. Second Case: Compliance with Implementation Obligation

116 While the first case is relatively easy to assess from a legal point of view, the second scenario poses a more complicated problem. Once deployers have implemented the constant monitoring system and a change in the misclassification rate, trend, or variance is detected, it is in principle still necessary to prove that the AI deployer has actually breached his duty of care. In fact, such a change may also have occurred through no fault of the latter. Hence, the mere fact that the system detected a change, does not necessarily constitute the deployer's breach of duty. However, in case the AI deployer fails to take any remedial action once a relevant change has been detected, such inaction despite knowledge of a negative change in performance, combined with

the possibility to act, would very likely have to be understood as a breach of the relevant duty of care - hence triggering the provider's liability. Given that the purpose of a constant monitoring system is primarily to detect change so that the AI operator can take action to prevent such change causing (further) damage, prompt action will be required and failure to do so will usually trigger liability.

117 While this does not resolve the issue of whether and when a change in AI performance as such constitutes a breach of a duty of care, liability triggered by inaction is likely to serve as a robust mechanism to minimise harm caused by AI. This approach is known from so-called "notice and takedown" obligations,[99] where operators can rely on safe harbours (i.e., no or limited liability) until they become aware of an infringement (e.g., illegal content) on their systems and are then liable from the moment they obtain knowledge if they do not act swiftly to remedy the infringement (e.g., remove content).

118 With regard to the outstanding issue of the detection of change as such being indicative of a breach of a duty of care, one needs to consider that random tests cannot normally prove a breach of duty per se. As shown before (Section C.I.1.a), a basic idea underpinning tort liability is that wrongful behaviour must be proven by the injured party. Showing a change in AI performance as such does not necessarily serve as evidence for breaching a duty of care: much depends on what caused that change, and whether the AI deployer could have prevented that change. Nevertheless, the inaction after a negative deviation from the required performance has been detected, can trigger liability.

119 The solution to address situations where a change in AI performance as such is the basis for a tort claim could consist of reversing the burden of proof. In that scenario, the AI deployer would have to prove that the change in performance does not reflect a breach of duty.[100] Such proof should normally be possible whenever the change is triggered by conditions outside the control of the AI provider - such as a shift in how for example users engage with an AI-based recommendation system. Even in cases where the AI deployer cannot exculpate itself (i.e., where the presumed breach of a duty of care stands), it would, however, still be up to the injured party to prove that the change occurred, as well as causality and damage. Alternatively, a strict liability regime could be implemented. In that case, the detected change in performance and the operator's inactivity would be sufficient as long as the injured party can prove the damage. The latter of these suggestions would, however, require a change in the existing (EU) law. However, under art. 10 sec. 4 of the new Product Liability Directive, the "defectiveness of the product or the causal link between its defectiveness and the damage, or both" can be presumed if "the claimant faces excessive difficulties, in particular due to technical or scientific complexity, in proving the defectiveness of the product or the causal link between its defectiveness and the damage, or both" (lit. a) and the claimant demonstrates that the causal link is "likely" (lit. b).

III. Implementation of a Cumulative Penalty System as Due Diligence

120 A cumulative penalty system as described above (see Section B.II.2.b) cannot be allocated to and executed within liability law. Setting up a penalty system contingent on legal assessments executed by third parties exceeds what is to be expected under conventional duties of care or due diligence. That aside, the penalties issued under the system as discussed above have punitive character, and do not correspond to the harm caused - whereas liability under tort law is governed by the general idea that those who cause harm have to restitute the situation that existed without the harmful event, or pay damages accordingly.

IV. Constant Monitoring as Part of an Early Warning Mechanism (for Big AI)

121 This Section considers existing approaches applied to online intermediaries, such as internet service providers (ISPs) and in particular (very) large platforms facilitating various forms of user interactions. As above, the analysis focuses on whether these approaches might provide for a suitable legal framework and/or useful implementation guidance for devising and operating a constant monitoring system with human oversight for Big AI, coupled with a cumulative penalty system, while adding some remarks on questions of liability and how they had been handled in relation to ISPs. Section C.IV.1 begins with a brief description of the conventional approaches adopted in the U.S. and the EU when online intermediaries emerged as key actors in the 1990s. It also introduces the more interventionist approach under the Digital Services Act (DSA) of the EU. Section C.IV.2 then looks at the specific legal tools available under the DSA, with a focus on measures applied to the operations of very large online platforms and search engines (VLOPSEs)

99 See e.g., for hosting providers, such as Internet Service Providers (ISPs), acc. to art. 16 DSA. A more detailed explanation of the concept will be given in Section C.IV.1.

100 Bundesgerichtshof [BGH] [Federal Court of Justice], Urteil vom 8. Dezember 1992 [judgement on Dec. 8, 1992], VI ZR 24/92 ("Mineralwasser I").

such as risk assessment, risk mitigation, independent audits and other oversight and enforcement mechanisms. Since the legal framework for online intermediaries has its foundations in exemptions from legal responsibility (in particular tort liability) of such intermediaries, but (under the DSA) has been extended to cover various means of direct regulation, it is presented here as a sui generis approach - somewhat in-between indirect regulation via liability law (see Sections C.I and C.II above), and direct regulation (see Section C.V below).

1. Introduction to Platform Regulation

122 When the Internet and other means of communication and data exchange through global digital networks gained commercial importance in the 1990s, the decentralised network structure demanded tools that would help users to navigate the increasingly complex network. ISPs and other service providers that mediated interaction among users emerged as essential facilities, serving a public function in terms of access to and sharing of information and other content, offering new dimensions for free speech and the ability to communicate across the globe. These providers in principle operate as private actors, not least because the rise of the (commercial) Internet coincided with globalising the 'Washington Consensus' of neo-liberal economic policies, including in the form of deregulation and privatisation of telecommunication services and related public utilities. As private actors, for a long time and up until recently, these intermediaries had not been subject to significant public law regulation - but of course remained potentially liable under private law vis-a-vis their own users and third parties. Due to their 'bottleneck' function as an essential facility for users' interactions online, ISPs ran the risk of being held liable for the conduct of their users - such as copyright infringements, defamation or hate speech. In order not to unduly impede ISPs (and in particular the development of network technologies used in their operations), the U.S. and then the EU adopted legal solutions for intermediary liability.[101] These solutions aimed to strike a balance between (1) utilising the ISPs' ability to prevent infringements and harm spreading through a network with (2) facilitating innovation in platform and other network technology.

123 In fact, convergent developments in this area can be observed across the globe. Notably, over the past decades a variety of different countries have adopted specialised sets of rules to govern the liability of intermediaries whose services are used by third parties to commit an infringement on the Internet.[102] These often take the form of so-called 'safe harbours', i.e., conditional liability exemptions that shield intermediaries from (some) legal responsibilities that they might otherwise incur, for example for actions that constitute participation in IP infringement or dissemination of hate speech. Notable examples of such rules include sec. 512 of the US's Digital Millennium Copyright Act (DMCA) and Arts. 13-14 of the EU's E-Commerce Directive.[103] While difficult to generalise, the most important common element of these safe harbours can be described by the 'notice and take-down' formula: With regard to content that is not their own (but of their users), ISPs have to act swiftly in removing or blocking access to illegal content - once they become aware of its illegal nature, usually once being notified. However, while these regimes display obvious similarities, significant differences also exist between them. So, for example, while s. 512 DMCA immunises intermediaries in the provision of four different kinds of services (mere conduit in s. 512(a), caching in s. 512(b), hosting in s. 512(c) and search services in s. 512(d)), the EU's E-Commerce Directive only protects intermediaries in the provision of the first three of these (mere conduit in art. 12, caching in art. 13 and hosting in art. 14). The DMCA is additionally limited to copyright infringement only, while the E-Commerce Directive applies horizontally to liability in a broad array of areas of civil and criminal law.

124 Differences also exist with regard to the reach of the safeguards provided. As opposed to their US equivalents, in the EU the safe harbours immunise intermediaries against liability only in the strict sense, i.e., for monetary compensation. All three of the EU safe harbours contain express permissions in their final paragraphs regarding the imposition of any kind of injunctive order on the providers of information society services by 'courts and administrative authorities' to 'terminate or prevent an infringement'.[104] Moreover, specifically in the area

101 See generally, Riordan, *The Liability of Online Intermediaries*, Oxford University Press (Sep. 7, 2016); Angelopoulos, *European Intermediary Liability in Copyright - A Tort-Based Analysis*, University of Amsterdam (Apr. 22, 2016).

102 See Seng, *Comparative Analysis of the National Approaches to the Liability of Internet Intermediaries* (2010), https://www.wipo.int/export/sites/www/copyright/en/docs/liability_of_internet_intermediaries.pdf (last accessed Feb. 28, 2024); and a more recent overview by Schmon, Pedersen, *Platform Liability Trends Around the World* (May 19, 2022), https://www.eff.org/deeplinks/2022/05/platform-liability-trends-around-globe-safe-harbors-increased-responsibility (last accessed Feb. 28, 2024).

103 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'), 2000 O.J. (L 178) 1, hereinafter: E-Commerce Directive.

104 Art. 12 sec. 3, 13 sec. 2, 14 sec. 3 and Recital 45

of copyright, EU law goes one step beyond this passive limitation: Recital 59 of the Infosoc Directive[105] observes that, in many cases, intermediaries are best placed to bring infringing activities occurring on their digital premises to an end. Art. 8 sec. 3 of the Infosoc Directive then explicitly instructs Member States to 'ensure that rightsholders are in a position to apply for an injunction against intermediaries whose services are used by a third party to infringe a copyright or related right'. Art. 11 of the 2004 Enforcement Directive[106] has subsequently expanded this obligation to cover all intellectual property rights.[107]

125 At the same time, there are limits on the injunctive orders as well as the scope of duties of care against intermediaries under EU law: a significant limitation on the permissible scope of both is imposed by art. 15 of the E-Commerce Directive, which prohibits the imposition of general obligations on service providers, in the provision of services protected by the three safe harbours, to monitor the information which they transmit or store or to actively seek facts or circumstances indicating illegal activity. The law of fundamental- and human rights also imposes limitations in this regard. The implications of this are considered in the next section when the more specific commitments now imposed under the DSA are reviewed for their suitability as a legal framework for an AI monitoring mechanism. As a successor to indirect platform regulation via the safe harbours in the E-Commerce Directive, the DSA takes a significantly more comprehensive and 'regulatory' approach.[108] In addition to provisions on ISP liability, the bulk of DSA rules concern 'Due Diligence Obligations for a Transparent and Safe Online Environment' (chapter III, arts. 11-48 of the DSA), including provisions on platform terms and conditions, transparency reporting, notice and action mechanisms, reasoned decisions, complaint handling and out-of-court settlements, trusted flaggers, recommender system transparency, protection of minors, as well as a range of commitments for specific types of online service providers.[109] While the next section can only address some of the elements of this approach when potentially relevant to AI monitoring, the DSA overall presents itself as the EU's complex regulatory response to increasing platform power, and a more diversified field of user - platform interactions.

126 In sum, ISPs or platform regulation emerged in the early days of the Internet primarily as a safe harbour mechanism whereby intermediaries can benefit from conditional exemptions from (civil) liability. Arguably, these exemptions created important legal foundations for platforms' business models, and to thrive economically. While there is no global set of harmonised rules (neither on indirect liability, nor on the safe harbours to avoid it), national approaches have - by and large and with some important differences - converged around a 'notice and take-down' approach. Such an approach eventually aims to balance the potential for the ISP to take effective action vis-a-vis its users when they commit allegedly illegal acts with the need not to impede the development of new technologies, also taking into account the (fundamental) rights of users, ISPs and affected third parties. We argue that, despite important differences in the factual settings at issue, a similar notion of balancing is at the heart of the AI monitoring system proposed here - but with a somewhat open question who is best placed to undertake this balancing, in particular in order to avoid potentially significant or even irreparable harm for more diffuse interests at a societal level. In addition, the increasing exposure of individuals (users and those directly or indirectly affected), groups, and societies to the often global reach of the most popular platforms corresponds to the exposures discussed in Section B.II to the reach of Big

E-Commerce Directive. Art. 18 E-Commerce Directive also requires that Member States 'ensure that court actions available under national law concerning information society services' activities allow for the rapid adoption of measures, including interim injunctions, designed to terminate any alleged infringement and to prevent any further impairment of the interests involved.' Cf. s. Digital Millennium Copyright Act (DMCA) § 512(j), 17 U.S.C.(1998) which does offer some limited possibilities of injunctive relief against intermediaries protected by one of the US safe harbours.

105 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, 2001 O.J. (L 167) 10, hereinafter Infosoc Directive.

106 Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights, 2004 O.J. (L 157) 45.

107 On the specific issue of injunctions against ISPs, see generally Husovec, *Injunctions against Intermediaries in the European Union: Accountable but Not Liable?*, Cambridge University Press (2017), https://doi.org/10.1017/9781108227421 (last accessed Feb. 28, 2024).

108 For a general discussion on the (European) 'new' regulatory approach to platforms, see Wagner et al., *Taming the giants: The DMA/DSA package*, 58 Common Market L.Rev. 987 (2021), https://doi.org/10.54648/cola2021065 (last accessed Feb. 28, 2024) and Iliopoulou-Penot, *Governing the 'Governors' of online speech: a review of the Digital Services Act's first steps in regulating social media platforms*, Yearbook of European law, 2025-10, https://doi.org/10.1093/yel/yeaf006 (last accessed Jan. 5, 2026).

109 On the DSA see generally, Husovec, *Principles of the Digital Services Act*, Oxford University Press (2024).

AI - so that arguments for regulatory intervention to prevent harm also align. With this in mind, the next section considers the regulatory tools of the DSA, especially those governing platforms with an equivalent harm-potential as Big AI.

2. Implementing a Constant Monitoring System with Human Oversight Based on the DSA

127 This section looks at the elements of the DSA's regulatory response to increasing platform power with a focus on those that may serve as a useful template for our proposed AI monitoring system with human oversight and a cumulative penalty mechanism, as described in detail in Section B.II.2.

128 To begin with, one might consider the foundational approach of 'safe harbour' liability exemptions for ISPs to be antithetical to the idea of devising a legal framework under which (Big) AI operators are responsible for preventing harm, or accountable for reparations should harm occur. And since the DSA retains the notion of safe harbours, including the 'no general obligation to monitor' principle,[110] its approach might appear ill-suited to guide the legal implementation of an AI monitoring system. However, the basic liability rules in Chapter II of the DSA concern responsibility not for own conduct, but that of *third* parties (*user* content and *acts of users)*; and they are driven by the broader context described in Section C.IV.1. While this can be distinguished from our concern over harm caused by AI, the due diligence obligations in Chapter III of the DSA appear as a much closer fit - in particular those related to 'very large online platforms' and 'very large search engines' (VLOPSEs) where the DSA takes a more regulatory, interventionist approach. Given that the deployment of Big AI might frequently take place via systems operated or supported by VLOPSEs and since the systemic risks identified with Big AI significantly overlap with those posed by VLOPSEs, the DSA regulatory approach to VLOPSEs can serve as a useful template.[111] This section hence focuses on these additional commitments for VLOPSEs in arts. 33 - 43 DSA, and the corresponding enforcement provisions in Chapter IV of the DSA.

a.) General Due Diligence Commitments for ISPs

129 Before considering in more detail the commitments for VLOPSEs pertaining to risk assessment and risk mitigation, it is worth highlighting some of the DSA provisions not specific to VLOPSEs, but with a potential guiding function for regulating harm caused by (Big) AI. To begin with, the DSA emphasises in various provisions the human rights implications of ISPs, online platforms and VLOPSEs in particular. Art. 1 DSA sets out its principal aim as, inter alia, creating a 'trusted online environment that facilitates innovation and in which fundamental rights enshrined in the Charter, including the principle of consumer protection, are effectively protected'. In relation to content moderation (including by means of AI) for example, art. 14.4 obliges all ISPs to 'act in a diligent, objective and proportionate manner' (....) 'with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter.' Under the risk assessment and mitigation obligations for VLOPSEs (discussed in more detail below), one of the four systemic risks which VLOPSEs have to assess and mitigate concerns 'any actual or foreseeable negative effects for the exercise of fundamental rights'.[112] The corresponding recital 81 DSA explains that '[s]uch risks may arise, for example, in relation to the design of the algorithmic systems used by the very large online platform or by the very large online search engine'. Similar commitments to take account of the fundamental rights of users can be found in arts. 36 and 38 DSA.[113]

110 See chapter II, art. 4-6, 8 DSA.

111 See Section B.II above on a discussion of the systemic risks posed by Big AI. The systemic risks posed by ISPs and in particular by VLOPSEs are discussed in recitals 1-3, 40, 47, 75-76, 79-90 DSA with recital 76 DSA setting out the basic idea that due to the potential for VLOPSEs to 'cause societal risks, different in scope and impact from those caused by smaller platforms', they 'should therefore bear the highest standard of due diligence obligations, proportionate to their societal impact'. For a similar approach, analysing DSA transparency obligations for 'high impact' AI, see Söderlund et al., *Regulating high-reach AI: On transparency directions in the Digital Services Act*, Internet Policy Review, 2024-01, Vol.13 (1), 1 - online at https://policyreview.info/articles/analysis/regulating-high-reach-ai (last accessed Jan. 2, 2026).

112 Art. 34.1 (b) DSA. The provision continues to list 'in particular the fundamental rights to human dignity enshrined in art. 1 of the Charter, to respect for private and family life enshrined in art. 7 of the Charter, to the protection of personal data enshrined in art. 8 of the Charter, to freedom of expression and information, including the freedom and pluralism of the media, enshrined in art. 11 of the Charter, to non-discrimination enshrined in art. 21 of the Charter, to respect for the rights of the child enshrined in art. 24 of the Charter and to a high-level of consumer protection enshrined in art. 38 of the Charter.' See also art. 35 sec. 1 DSA whereby the mitigating measures adopted to address the systemic risks need to specifically focus on the impact on fundamental rights.

113 See also the further elaborations on the diversity of

130 While the approach taken in the DSA involves broader questions about the extent to which ISPs as private actors should be human rights duty bearers which we do not discuss here,[114] the principal analogy drawn between ISPs, in particular VLOPSEs, and Big AI in terms of the respective potential to cause societal (if not global) harm appears prima facie defensible: Given (1) the direct link set out in recital 81 DSA (whereby one of the risks to the enjoyment of fundamental rights is posed by the design of AI systems employed by VLOPSEs); and (2) the similarity of circumstances between ISPs, in particular VLOPSEs, and Big AI in terms of individual and societal impact, it is worth considering whether the constant monitoring system for Big AI should equally include an assessment of the impact Big AI has on fundamental rights. This would primarily focus on the rights of those directly impacted by Big AI - the end users of AI systems, as well as those in relation to whom an AI system is applied. Ideally, the human-led monitoring process we propose would also include the impact on fundamental rights of those indirectly affected - that is, to the extent such an indirect impact is reasonably foreseeable, or to the extent information about such an indirect impact becomes available. However, as the monitoring system we advocate focuses on reviewing random samples of AI output/decisions (without those being tied to an individual complaint or dispute from which to gather further information about relevant human rights impact), the review will have to focus on human rights implications that are apparent from the AI output/decision (such as freedom of speech when legal content is filtered, or the rights of children if inappropriate content is not blocked). As discussed in Section B.II.4 above, the constant monitoring system has to focus on more 'tangible' legal interests, the impact on which can actually be assessed in full (i.e., without the need for further contextual information) by the CA on the basis of the samples provided.

131 Before moving on to the specific risk assessment and mitigation obligations for VLOPSEs, one should also mention that the DSA includes rules on transparency reporting on content moderation (including the extent to which automation has been relied on, and with what result).[115] Since the reports on automated content moderation have to include 'indicators of the accuracy and the possible rate of error of the automated means used', adhering to such reporting obligations implies a constant monitoring and review of AI-based content moderation tools and hence aligns with the basic idea of the constant monitoring system we propose. In fact, complying with the reporting commitments on AI-accuracy and error rates for content moderation might well necessitate a monitoring system equivalent to the one proposed in Section B.II.2. Another basic element of our proposal that finds support in the DSA is the concept of differentiating commitments depending on the size and type of the operator.[116] Furthermore, the idea of certifying dispute settlement providers for disputes between ISPs and users (with the possibility of revocation);[117] and the notion of 'trusted flaggers' (which serve to flag illegal content)[118] are elements of the DSA that can offer useful input for

contexts for and roles of fundamental rights especially in recitals 47, 51, 54, 63, 79, 81, 86, 109, and 153 DSA. On fundamental rights under the DSA, see Frosio et al., *Taking fundamental rights seriously in the Digital Services Act's platform liability regime. Eur Law J.* 2023; 29(1-2): 31-77. https://doi.org/10.1111/eulj.12475 (last accessed Jan. 5, 2026).

114 See on this United Nations, *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*, Final report to the Human Rights Council (A/HRC/17/31, Geneva, 2011). The UN Guiding Principles have been developed by J Ruggie, then Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.

115 See art. 15 sec. 1 (c) and in particular (e) DSA whereby reporting has to cover 'any use made of automated means for the purpose of content moderation, including a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used in fulfilling those purposes, and any safeguards applied', as well as art. 24 DSA for online platforms (concerning further reporting commitment on internal complaints, disputes, and abuse-countermeasures), and art. 42 DSA for VLOPSEs. See generally Enarsson, *Transparent, Rapid and Contextualised? Comparing Content Moderation Requirements in Emerging EU Regulation*, European journal of law and technology, 2025-05, Vol. 16 (2), 1 & Söderlund et al., as note 118.

116 See arts. 4-6, 19, 20-28, 29, 33 DSA (including an exemption under art. 19 from many duties for micro and small enterprises with less than 50 employees and no more than EUR 10 million turnover, as well as detailed rules in art. 33 on how to determine when an online platform is 'large enough' to constitute a 'very large online platform', triggering the corresponding significant additional commitments).

117 See art. 21 DSA setting out the principal standards against which certifications are to be issued, the requirements and procedures for dispute settlement providers to adhere to when offering their services, and related reporting obligations. In case it is not feasible for the monitoring of AI under our proposal to be not conducted by state authorities, similar certification principles for private providers to conduct the monitoring may be considered.

118 In particular, the selection criteria for trusted flaggers (such as expertise, independence, diligence and objectivity), as well as the reporting standards of trusted flaggers could serve as a basis for the experts relied on in our system of human monitoring for AI, in particular if such monitoring had to be conducted by those other than state authorities; see arts. 16 sec. 1, 22 sec. 2 and 3 DSA.

the implementation of the proposed AI monitoring system.

b.) Risk Assessment and Mitigation

132 The most relevant aspects of the due diligence commitments under Chapter III of the DSA, however, relate to risk assessment and mitigation by VLOPSEs. Recitals 75, 76 and 79 DSA explain the motivation for imposing these commitments: Due to the 'reach' of VLOPSEs[119] they 'may cause societal risks, different in scope and impact from those caused by smaller platforms'.[120] Given that VLOPSEs as private actors design and operate their services in order to maximise profit and/or optimise their business models, the societal concerns caused by their reach are externalised. Hence, 'effective regulation and enforcement is necessary in order to effectively identify and mitigate the risks and the societal and economic harm that may arise'.[121] VLOPSEs 'should therefore bear the highest standard of due diligence obligations, proportionate to their societal impact'.[122] The DSA assumes that once a platform exceeds a given number of active users (currently set at 45 million per month)[123] the systemic risks posed 'may have a disproportionate impact in the Union'. For these reasons, the risk assessment and mitigation commitments are considered 'necessary' in order to be able to address public policy concerns resulting from the potential impact of VLOPSEs, there being - according to the EU - 'no alternative and less restrictive measures that would effectively achieve the same result.'[124] Based on our definition of Big AI, the analogous problems of (1) externalising public policy concerns, as well as (2) the far-reaching and systemic impact on societies,[125] offer a strong argument for adopting at least an equivalent approach to Big AI operators. As shown below, the specific DSA commitments for VLOPSEs align significantly with our suggestions for a constant monitoring mechanism for Big AI.

133 According to art. 34 sec. 1 DSA, VLOPSEs are under an obligation to 'diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services'. These risk assessments have to be carried out at least once every year, and 'in any event prior to deploying functionalities that are likely to have a critical impact on the risks identified pursuant to this Article'.[126] The VLOPSEs risk assessment must be 'specific to their services and proportionate to the systemic risks, taking into consideration their severity and probability'.[127] Art. 34 sec. 1 DSA then defines four particular categories of systemic risks VLOPSEs have to assess:

> (a) 'the dissemination of illegal content through their services';
>
> (b) 'any actual or foreseeable negative effects for the exercise of fundamental rights';
>
> (c) 'any actual or foreseeable negative effects on civic discourse and electoral processes, and public security'; and
>
> (d) 'any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being'.[128]

134 When assessing these risk categories, art. 34 sec. 2 DSA requires VLOPSEs to consider the

119 Recital 75 DSA describes such 'reach' in terms of facilitating public debate, economic transactions and the dissemination of information, opinions and ideas and in influencing how users obtain and communicate information online; and for search engines in terms of locating and making information retrievable online.

120 See recital 76 DSA. Recital 79 DSA describes these risks as the potential for VLOPSEs to be used in a way that 'strongly influences safety online, the shaping of public opinion and discourse, as well as online trade'. On VLOPSEs generally, see Moravcová, *Impact of the DS Regulation on Very Large Online Platforms*, Central European journal of comparative law (Online), 2023-12, Vol.4 (2), 163.

121 See recital 79 DSA.

122 See *id.* recital 76 DSA.

123 Art. 33 sec. 1 DSA - a figure that, according to recital 76 DSA, corresponds to 10% of the current population in the EU. Art. 33 DSA further sets out details on how to adjust this number, depending on the EU population, and how to determine whether a platform or search engine falls under this category. For a list of those actors that currently fall under this category, see https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses.

124 Recital 76 DSA.

125 See Section B.II above.

126 Art. 34 sec. 1 DSA. In contrast to this, the constant monitoring system proposed here would involve much more frequent, although less in-depth scrutiny of AI outputs (hence capturing primarily direct observable harm very swiftly - whereas full fledged audits under the DSA are better suited for diffuse harms).

127 Recital 79 DSA further explains that in determining the significance of potential negative effects and impacts, providers 'could assess whether the potential negative impact can affect a large number of persons, its potential irreversibility, or how difficult it is to remedy and restore the situation prevailing prior to the potential impact.'

128 See art. 34 sec. 1 (a)-(d) DSA. On point (b), see also the discussion above (incl. notes 118-120) concerning the link between human rights, VLOPSEs, Big AI impact and risk assessment.

impact of factors such as (a) 'the design of their recommender systems and any other relevant algorithmic system'; (b) 'their content moderation systems'; (c) 'the applicable terms and conditions and their enforcement'; (d) 'systems for selecting and presenting advertisements'; and (e) 'data related practices of the provider'. Assessment of the systemic risk categories (a)-(d) under art. 34 sec. 1 DSA also has to consider the potential for 'intentional manipulation of [the VLOPSEs'] service, including by inauthentic use or automated exploitation of the service', as well as the 'amplification and potentially rapid and wide dissemination of illegal content'.[129]

135 The analogy to the potential risks posed by Big AI is apparent: First, the four categories of systemic risks identified in art. 34 sec. 1 DSA equally apply to the deployment of Big AI - not least since scale, speed and severity of any potential harm caused by VLOPSEs in these categories will usually result from VLOPSEs' own reliance on AI and other means of automation when providing their services. In fact, four of the five factors in art. 34 sec. 2 (a) - (e) DSA relate to the use of AI and other automatic means in supplying VLOPSEs services, again illustrating the impact Big AI can have on individuals, groups and societies. Finally, the potential for abuse as well as the amplification of impact referred to in art. 34 sec. 2 DSA is particularly acute when AI is deployed: recital 84 DSA explicitly refers to VLOPSEs duty to consider how 'algorithmic amplification of information contributes to the systemic risks'. Given this overlap, the risk assessment obligations under art. 34 DSA can serve as a useful template for a legal framework applicable to risk assessments related to the deployment of Big AI. In fact, given that many situations where VLOPSEs have to assess (and then mitigate) risk involved in the deployment of AI, our model for a constant monitoring system in turn can, subject to some important modifications,[130] serve as a template for implementing the commitments under art. 34 DSA. At the same time, the nature of the constant output monitoring mechanism we propose does not lend itself to a thorough analysis of latent and potentially remote risks to less concretely formulated legal interests, as reflected in some human rights and in broader societal concerns for democratic, representative and non-discriminatory governance structures.

136 The principal distinction between the risk assessment and mitigation framework of the DSA and our proposal for constant monitoring of Big AI then primarily lies in the greater autonomy provided to the private operator under the DSA: our model suggests constant monitoring and oversight *by an expert third party* (our 'check authority', CA) to prevent harm from materialising - rather than leaving risk handling primarily in the hands of the private actor. Our monitoring proposal further involves more frequent, constant control that is meant to detect harm quickly and early - rather than the DSA risk assessment, in principle to be conducted annually. At the same time, our proposal is narrower in terms of *limiting the risk monitored* to legal interests whose harm can be observed from the samples available to the CA. Despite these distinctions, the DSA also foresees further oversight and control. Under art. 35 DSA, VLOPSEs have to 'put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to art. 34 DSA, with particular consideration to the impacts of such measures on fundamental rights'.[131] Art. 35 DSA leaves the specific mitigatory response to the platform - but suggests this may consist, inter alia, in adapting VLOPSEs' (1) online interfaces, design, features of functioning; (2) their content moderation systems; or in (3) 'testing and adapting their algorithmic systems, including their recommender systems'. To the extent these measures would involve automated controls or tweaks in VLOPSE's AI systems, they would be subject to the same criticism we discussed regarding relying on an 'AI checking an AI'. Checking compliance with these due diligence mitigation obligations for VLOPSEs then is implemented through reporting duties, independent audits, data access and scrutiny, as well as the enforcement powers for national authorities (acting as Digital Service Coordinators, DSCs), the European Board for Digital Services (EBDS) and the European Commission (EC), including the

129 See art. 34 sec. 2 DSA. On addressing systemic risks under the DSA see also Kaesling, Wolf, *Sustainability and Risk Management under the Digital Services Act: A Touchstone for the Interpretation of 'Systemic Risks'*, GRUR International, 2025-02, Vol.74 (2), 119; Liesenfeld, *The Legal Significance of Independent Research based on Article 40 DSA for the Management of Systemic Risks in the Digital Services Act*, European journal of risk regulation, 2025-03, Vol.16 (1), 184.

130 Since art. 34 DSA primarily foresees the risk assessment to be conducted by VLOPSEs (subject to review in accordance with arts. 37, 40, 49-50 DSA - discussed below), our proposal for a constant monitoring system by an external entity would require some important adaptation. Despite these principal differences VLOPSEs could 'test' their systems using a constant monitoring model as described in Section B.II.2.a, in particular as a precautionary pre-deployment measure to limit directly observable errors leading to potential harm.

131 Recital 86 DSA adds that these mitigation measures 'should be proportionate in light of the economic capacity of the provider of the very large online platform or of the very large online search engine and the need to avoid unnecessary restrictions on the use of their service, taking due account of potential negative effects on those fundamental rights. Those providers should give particular consideration to the impact on freedom of expression.'

possibility for significant and periodic penalties.[132] Since there is no room here to fully review these oversight and enforcement tools, the following summary merely highlights the main overlaps with our proposed constant monitoring system.

c.) Oversight and Enforcement Tools

137 Under art. 37 DSA, VLOPSEs 'shall be subject, at their own expense and at least once a year, to independent audits to assess compliance'. To enable such audits in an 'effective, efficient and timely manner' VLOPSEs have to provide independent auditors 'access to all relevant data and premises and by answering oral or written questions' - who then have to produce substantiated written audit reports.[133] In case the audit does not yield positive results, VLOPSEs 'shall take due account of the operational recommendations addressed to them with a view to take the necessary measures to implement them'.[134] Next to being subject to annual independent audits, VLOPSEs have to adopt response mechanisms in case of crisis (such as a serious threat to public security or public health in the EU)[135] and generally have to offer access to data (including data that 'explain[s] the design, the logic, the functioning and the testing of their algorithmic systems')[136] to the extent 'necessary to monitor and assess compliance with this Regulation'.[137] Data access has to be provided to the national DSC, or the EU Commission. The former function as the default oversight authority and might serve as a template for setting up a 'check authority' (CA) under our proposed monitoring system for Big AI.[138] The investigatory and enforcement powers of the DSC include powers to set fines and periodic penalty payments, escalating up to temporary access restrictions to platform services, and to adopt interim measures.[139] However, any of the enforcement measures taken must be 'effective, dissuasive and proportionate, having regard, in particular, to the nature, gravity, recurrence and duration of the infringement or suspected infringement to which those measures relate, as well as the economic, technical and operational capacity of the provider of the intermediary services concerned where relevant'.[140] In addition, these oversight and enforcement measures are subject to general safeguards under EU law.[141]

138 Subject to these safeguards, DSCs can issue penalties and fines against ISPs for breaching their due diligence obligations under Chapter III of the DSA. Under art. 52 DSA (and with some further qualifications for VLOPSEs under arts. 74-76 DSA), the maximum amount of fines for non-compliance 'shall be 6 % of the annual worldwide turnover of the provider of intermediary services concerned in the preceding financial year' and the maximum amount of a periodic penalty payment 'shall be 5 % of the average daily worldwide turnover or income of the provider of intermediary services concerned in the preceding financial year per day'.[142] The oversight

132 See generally Chapter IV, arts. 64-83 of the DSA, detailing the specific enforcement regime in relation to VLOPSEs. The EU has not shied away from relying on its enforcement powers under the DSA - for example in relation to TikTok and Meta (see: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503) as well as X (formerly Twitter - see: https://digital-strategy.ec.europa.eu/en/news/commission-fines-x-eu120-million-under-digital-services-act).

133 See art. 37 sec. 2 and 4 DSA. In addition, under art. 34 sec. 4 DSA VLOPSEs have to retain the supporting documents of their risk assessments for at least three years and, upon request, have to communicate these to the relevant enforcement agencies. Art. 37 sec.3 DSA further sets out requirements which aim to ensure the independence and impartiality of the auditors.

134 See art. 37 sec. 6 DSA.

135 See art. 36 and recital 91 DSA. As part of a crisis response mechanism, VLOPSEs have to 'apply specific, effective and proportionate measures to prevent, eliminate or limit any such contribution to the serious threat' (art. 36 sec.1 DSA). This is subject to oversight by enforcement authorities to ensure that the response measures are 'strictly necessary, justified and proportionate' (art. 36 sec. 3 DSA). In addition, there is a recommendation to draft crisis protocols under art. 48 DSA.

136 See art. 40 sec. 3 DSA. See also Liesenfeld, as note 136.

137 See art. 40 sec. 1 DSA.

138 See arts. 49, 50 DSA - whereby DSC have to 'perform their tasks (...) in an impartial, transparent and timely manner' and 'shall act with complete independence (...) free from any external influence, whether direct or indirect, and shall neither seek nor take instructions from any other public authority or any private party' (art. 50 sec. 2 DSA).

139 See art. 51 DSA.

140 See art. 51 sec. 5 DSA. In particular with regard to access restrictions (which might be an important sanctioning tool for Big AI as well) '[a]ny measure ordered shall be proportionate to the nature, gravity, recurrence and duration of the infringement, without unduly restricting access to lawful information by recipients of the service concerned' (art. 51 sec. 3 DSA).

141 See art. 51 sec. 6 DSA whereby 'any exercise of those powers is subject to adequate safeguards laid down in the applicable national law in compliance with the Charter and with the general principles of Union law. In particular, those measures shall only be taken in accordance with the right to respect for private life and the rights of defence, including the rights to be heard and of access to the file, and subject to the right to an effective judicial remedy of all affected parties'.

142 See also the specific powers to issue fines and periodic penalty payments to VLOPSEs under arts. 74, 76 DSA, discussed below. On 5 December 2025, the US-based platform X (formerly Twitter) had been fined a sum of 120

is further strengthened by an individual's right to lodge a complaint against an ISP (including of course VLOPSEs) with the DSC.[143] Somewhat related to such individual complaints, art. 54 DSA refers to national negligence liability rules whereby '[r]ecipients of the service shall have the right to seek, in accordance with Union and national law, compensation from providers of intermediary services, in respect of any damage or loss suffered due to an infringement by those providers of their obligations under this Regulation.' This suggests that a breach of the user-related obligations under the DSA also serves as basis for a tort claim under national liability law,[144] hence creating a similar parallelity of regulatory and liability frameworks as we have discussed above.

139 Finally, it is worth noting that under arts. 65-83 DSA, a detailed supervisory framework specific to VLOPSEs, operated by the EU Commission, exists. This framework aligns with the general implementation monitoring and enforcement powers discussed above, and - in addition - gives the EU Commission the right to:

- Initiate an investigation - *ex officio* or based on a request from a DSC - with a view to determining a VLOPSE's compliance with the DSA (arts. 65-66 DSA);
- Request information from the VLOPSE concerned, and/or any third parties who might be in the possession of relevant information (art. 67 DSA);
- Conduct investigations, including by means of access to and inspection of premises, books, data and other material of providers - if need be under reliance on coercive measures (arts. 68-69 DSA);
- Adopt any necessary interim measures in case of urgency due to the risk posed by the VLOPSE under investigation (art. 70 DSA);
- Make binding any commitments offered by the VLOPSE to ensure compliance with the DSA (art. 71 DSA);
- Take all necessary actions to monitor VLOPSE's compliance with the DSA, including requiring the retention of data or documents by the VLOPSE, and/or appointing independent auditors (art. 72 DSA);
- As part of a decision on non-compliance (art. 73 DSA), issue significant fines of up to 6% of the annual global turnover of the VLOPSE - but only in case of intentional or negligent breaches of the DSA (art. 74 DSA);[145]
- Engage in enhanced compliance supervision, such as demanding the VLOPSE to 'draw up an action plan setting out the necessary measures which are sufficient to terminate or remedy the infringement' (art. 75 DSA); and
- Issue 'periodic penalty payments not exceeding 5% of the average daily income or worldwide annual turnover in the preceding financial year per day', including to compel compliance with information requests, interim measures, and final decisions (art. 76 DSA).

140 All these supervisory and enforcement measures

million Euros for breaches of its transparency obligations under the DSA - see: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934.

143 See art. 53 DSA.

144 The further requirements of liability arguably derive from domestic law - see the discussion by Benjamin Raue in Hofmann/Raue (eds.), Digital Services Act: DSA, Article 54, para. 6-10.

145 In addition to fines as part of a decision on non-compliance, fines of up to 1% of the global annual turnover can be issued in cases of negligent or intentional non-compliance with the Commission's investigation (such as providing late, incomplete or misleading information); see *id.* art. 74 sec. 2 DSA.

at the hand of the EU Commission are, of course, subject to relevant limitation periods (arts. 77 and 78 DSA) as well as general due process guarantees (art. 79 DSA). The decisions of the EU Commission on measures adopted vis-a-vis VLOPSEs are to be made public (art. 80 DSA) and subject to review by the Court of Justice of the European Union (CJEU) in accordance with art. 81 DSA.

d.) Commonalities and Differences

141 In sum then, the DSA regulatory framework with significantly enhanced due diligence commitments, in particular for VLOPSEs, offers an interesting template for many aspects of our proposed monitoring system with human oversight and cumulative penalty system. While the DSA's approach to VLOPSEs does not exactly match our proposal, it reinforces the basic idea of preventing harm by a tailored mechanism of risk assessment and mitigation, further backed by monitoring, enforcement and, if necessary, sanctions. At the same time, the essential difference of internal (DSA) vs external (our proposal) monitoring makes sense in relation to the type of risks covered: we argue that external oversight based on samples of AI output is a safer, and hence preferable, way of addressing the risks which can actually be checked by a constant monitoring system - namely those where the harm materialises in relation to clearly defined and well understood legal interests. For the more remote or indirect risks to more diffuse, less 'tangible' legal interests (for example in relation to societal goals and governing principles), our constant monitoring based on samples will often not work well in detecting relevant harm merely on the basis of the facts observable from the samples. For these cases, the due diligence risk assessment and mitigation tools of the DSA might well serve as the, for now, best available option. A lot here will depend on factors such as: (1) how serious the VLOPSE takes its commitments to assess and mitigate systemic risks posed to human rights and other fundamental societal interests; (2) how rigorous external control by the DSC and the EU Commission is exercised; and - perhaps most importantly - (3) how much of the weighing and balancing inherent in making decisions about addressing systemic risks to less tangible legal and societal interest is effectively left to the private actor. One hopes that enhanced regulation *of* platforms via the DSA does not lead to ever more aspects of our (digital) lives being effectively regulated *by* platforms.

142 Returning to the comparison between the DSA model and our proposal, more fundamentally still, the underlying motivation for regulatory intervention is essentially identical: the systemic risk posed to individuals (both users and others directly or indirectly affected), groups and societies through private operators of systems that have unprecedented reach and their potential for all kinds of harm. In fact, many facets of this risk result from the common denominator of VLOPSEs and Big AI operators: the reliance on automated output that is scalable, fast and usually global in reach. VLOPSEs will frequently rely on Big AI systems in offering their services. For this reason, the regulation of VLOPSEs effectively often includes regulating Big AI (in particular to the extent VLOPSES deploy AI tools in supplying their services in the EU), along the lines of the DSA's internalised risk assessment and mitigation model with some degree of external oversight. At the same time, the DSA's approach still retains aspects of liability law (primarily in the form of the safe harbours discussed in Section C.IV.1) and hence needs to be distinguished from the direct regulatory approach to AI discussed in the next section.

V. Constant Monitoring Obligations as a Part of Direct Regulation (for Big AI)

143 We continue by examining how constant monitoring with human oversight can be a part of existing and upcoming regulatory law. As before, we start by giving a short overview of the basic concept of regulatory law. Thereafter, we will analyse if and where both random testing and the penalty system can be allocated in regulatory law.

1. Introduction to Regulatory Law

144 In contrast to liability law, regulatory law is a form of direct up front (ex ante) regulation which aims at preventing damage before it is caused. With regard to products, the main idea consists of the prevention of damage by setting rules for certification and accreditation. Regulatory law establishes obligations to be obeyed before placing an AI system on the market. In addition to that, it sets out continuous obligations which must be followed during the whole lifecycle of a product after it has been placed on the market.

145 Regulatory law is considered a direct form of regulation because its rules address the regulatee directly in a way that obligates the regulatee to act in a particular way, and/or not to commit certain acts; whereas liability rules function more indirectly by incentivising a certain behaviour in order to avoid liability.

2. Constant Monitoring with Human Oversight Under Existing Regulatory Law

146 Using EU law as principal legal reference point (not least because of the comparably more rigid approach to regulatory intervention in the EU, in contrast to an often more *laissez-faire* approach in the US), direct regulation is applied in many areas. Product safety, for example, is regulated by the "New Legislative Framework" (NLF) which sets out harmonised rules providers must comply with in order to be allowed to participate in the European internal market. In particular, the NLF contains a bouquet of Regulations and Directives laying down general rules for conformity assessment, CE marking, accreditation of conformity assessment bodies, and market surveillance.[146] Besides that, all sector-specific Regulations and Directives (i.e., Medical Devices Regulation,[147] Toy Safety Directive,[148] Construction Products Regulation,[149] etc.) are aligned with these general provisions. Hence, for each specific type of product, the same general rules concerning the conformity assessment, CE marking, and market surveillance apply. If AI systems are thus built into products whose categories are already covered by the NLF regulations, operators must comply with the NLF requirements under existing law if they wish to place their AI-driven products on the EU market.

a.) Constant Monitoring with Human Oversight as Part of Market Surveillance

147 Constant monitoring with human oversight could be implemented as part of the market surveillance laid down in the Market Surveillance Regulation (MSR).[150] As part of the NLF, the MSR aims at guaranteeing safe products by obliging operators to observe their products continuously after the latter have been placed on the market.

148 Pursuant to art. 11 sec. 3 MSR, market surveillance authorities shall perform appropriate checks on the characteristics of products on an adequate scale, by means of documentary checks and, where appropriate, physical and laboratory checks based on adequate samples. In addition to that, in deciding on which checks to perform, on which types of products and on what scale, market surveillance authorities shall follow a risk-based approach, taking into account possible hazards, the economic operator's record of non-compliance (as suggested above in Section B.II.2.b), and further factors. Hence, constant monitoring with human oversight could be implemented on the basis of the MSR.

b.) Constant Monitoring with Human Oversight as Part of Technical Norms and Standards

149 Moreover, the NLF is characterised by its specific regulatory structure: While both general and sector-specific Regulations and Directives such as the MSR and the Medical Devices Regulation set out general requirements for the respective products, concrete instructions for action are provided by technical norms and standards (provided by standardisation organisations, such as the ISO, CEN, CENELEC, and ETSI).

150 While the provisions of Regulations and Directives generally are binding (in that they set out mandatory legal commitments to be complied with), compliance with technical norms and standards is voluntary. However, compliance with such norms and standards triggers a rebuttable presumption that manufacturers comply with the mandatory legal requirements (presumption of conformity). For this reason, and because the technical norms and standards are largely developed on an industry-driven basis, they are highly relevant in practice and are implemented by most companies.

151 Monitoring obligations are already set out in existing technical standards, such as ISO 9001,[151] which lays down general rules and objectives for quality management, particularly for measuring and analysing customer satisfaction and product performance. In addition to that, sector-specific

146 European Commission, *New legislative framework*, European Union, https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en (last accessed Feb. 29, 2024).

147 Regulation of the European Parliament and of the Council 2017/745 of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, 2017 O.J. (L 117) 1, hereinafter Medical Devices Regulation.

148 Directive of the European Parliament and of the Council 2009/48/EC of 18 June 2009 on the safety of toys, 2009 O.J. (L 170) 1.

149 Regulation of the European Parliament and of the Council 305/2011 of 9 March 2011 laying down harmonised conditions for the marketing of construction products and repealing Council Directive 89/106/EEC, 2011 O.J. (L 88) 5.

150 Regulation of the European Parliament and of the Council 2019/1020 of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, 2019 O.J. (L 169) 1, hereinafter MSR.

151 International Organization for Standardization, *ISO 9001 Quality management systems – Requirements* (Sep. 2015), https://www.iso.org/standard/62085.html (last accessed Feb. 29, 2024).

standards (e.g., ISO 13485:2016[152] for medical devices) stipulate particular instructions for quality management (including monitoring obligations) which apply to products that belong to the respective sector. Hence, monitoring obligations are already part of technical norms and standards and compliance with these obligations triggers the rebuttable presumption that a product is compliant with EU law. As a result, the proposed constant monitoring system with human oversight could be implemented as part of (general or specific) technical standards for quality management as long as the AI system is embedded into a product which falls under the NLF scheme. A penalty system, however, cannot be implemented into a standard, since technical standards solely aim at providing concrete instruction for action. Thus our proposal for penalty systems (outlined above in Section B.II.2.b) goes beyond what is possible with standards.

152 At the time of writing, there are no specific standards for the quality management of AI systems. However, work is currently being done on such a standard[153] by the technical committee ISO/IEC JTC 1/SC 42. Thus, a constant monitoring system with human oversight which applies specifically to AI, could soon be implemented into a technical standard. The work of the responsible technical committee remains to be seen. The system we outlined above (see Section B.II.2) could serve as a basis for the committee's work on monitoring standards as requested by the EU COM in their standardisation request of May 2023 which has been accepted by the European standardisation organisation CEN and CENELEC.[154]

3. Constant Monitoring with Human Oversight Under Recently Enacted Direct Regulation

153 At the time of writing, Europe's first binding legal framework for AI has recently been finalised. After the Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (AI Act)[155] had been issued by the European Commission (EU COM) in April 2021, both the Council of the European Union (Council) and the European Parliament (EP) published their compromise amendments in November 2022[156] and June 2023,[157] all three legislative bodies entered trilogue negotiations in 2023. In December 2023, a political agreement was reached and a (provisional) consolidated version of the final text was issued in January 2024. Once the text was officially translated and both Council and European Parliament voted in favour of the new regulation, the regulation entered into force on 2 August 2024.[158] In general, according to art. 113, its rules will apply from 2 August 2026.

154 One central characteristic of the AI Act consists of its risk-based approach. This approach allows it to classify and regulate AI systems according to the risk they entail and the application field they are used in. The risk-based approach means that the only AI systems subject to strict requirements are those that are categorised as actually posing a high risk. Systems that pose a low or minimal risk, on the other hand, are subject to few or no obligations. A high risk is assumed where AI systems are used in areas already subject to third-party conformity assessments and certification procedures under

152 International Organization for Standardization, *ISO 13485 Quality management systems – Requirements for regulatory purposes* (Mar. 2016), https://www.iso.org/standard/59752.html (last accessed Feb. 29, 2024), which also contains monitoring obligations.

153 International Organization for Standardization, *ISO/IEC FDIS 42001 Information technology – Artificial intelligence Management System* (Dec. 2023), https://www.iso.org/standard/81230.html (last accessed Feb. 29, 2024).

154 European Commission, Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence (C(2023) 3215 final), https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en (last accessed Jan. 8, 2026).

155 European Commission, *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), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (last accessed Feb. 29, 2024).

156 Council of the European Union, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules in artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts- General approach,* 14954/22 (Nov. 25, 2022), https://artificialintelligenceact.eu/wp-content/uploads/2022/12/AIA-%E2%80%93-CZ-%E2%80%93-General-Approach-25-Nov-22.pdf (last accessed Feb. 29, 2024).

157 European Parliament, *Amendments adopted by the European Parliament on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain legislative acts*, P_TA(2023)0236 (Jun. 14, 2023) https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf (last accessed Feb. 29, 2024).

158 Regulation of the European Parliament and of the Council 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence 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).

sector-specific regulation (e.g., medical devices) or where they are used in particularly sensitive areas (e.g., critical infrastructure) and pose significant risk.

155 Hence, if an AI system is implemented into a medical device which falls under the Medical Devices regulation and must undergo a third party conformity assessment (e.g., systems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act).an AI system which is implemented into MRI scannerssystems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act). using computer vision to detect anomalies) it is automatically considered high-risk under the AI Act (art. 6 sec. 1 and annex I sec. A no 11 of the AI Act). Regarding high-risk AI systems which are used in sensitive areas (systems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act).e.g., AI systems used to assess the creditworthinesssystems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act). of natural persons or AI systems used in the job application process), the operator, after testing the system, may argue that the system does not entail a high risk, as there is no significant risk in the specific use case scenario - for example, in the case of an AI system used to support human decision-making. And those AI systems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act).

156 A differentiation between Big AI and Small AI elaborated in Section A.V is not taken into consideration within the AI Act's risk classification. Hence, Small Team - Big AI, for example, would not fall under the high-risk group by default. The matrix illustrated in Section A.V could however be added to the annexes of the AI Act. So far, the annexes serve as substantiation of the legal text. Annexes I and III, for example, concretise which harmonised regulations and which areas of application are relevant for the risk classification. Small AI and Big AI could be added to annex III which sets out the areas or providers that pose a high risk. This way, the proposed system (matrix + monitoring obligation) could be adopted within the AI Act if Annex III was to be modified accordingly. The EU COM is entitled to amend annex III of the AI Act if necessary.

157 As mentioned above, for providers of high-risk AI systems the obligations pursuant to arts. 9 - 15 AI Act apply. Among these obligations are monitoring and testing of high-risk AI systems as part of the risk management pursuant to art. 9 sec. 2, 5, 6, 7 of the AI Act. In addition to that, high-risk AI systems must be tested in order to fulfil the data quality requirements set out in art. 10 AI Act. Continuous monitoring could also be implemented as part of the obligatory record-keeping (art. 12 sec. 2c AI Act) since monitoring goes hand in hand with keeping logs. Furthermore, continuous monitoring could be implemented as one, if not the most effective way to guarantee human oversight, which is required under art. 14 sec. 3 and 4. And finally, a constant monitoring system could be part of the required post market monitoring pursuant to art. 72 AI Act.

158 For deployers of high-risk AI systems, monitoring the system's operation will be obligatory pursuant to art. 26 sec. 5 AI Act. Hence, constant monitoring with human oversight would equally serve to implement obligations for AI deployers. For providers and third parties involved in the conformity assessment of high-risk AI systems, art. 43 AI Act as well as the annexes VI and VII of the AI Act or sector-specific regulation could be the linking point for the monitoring system proposed in this paper. Monitoring and testing could hence be part of the conformity assessment. Moreover, the cooperation between providers and notified authorities (equivalent to the Check Authority or "CA" discussed in Section B.II.2.a) is covered by art. 43 sec. 3 AI Act.

159 The specific process or monitoring loop would not be described in the AI Act itself but rather laid down by technical regulation (norms and standards, see above). The central provision establishing the rebuttable presumption of conformity in case harmonised standards are complied with, is art. 40 AI Act.

160 To summarise, it can be concluded that the AI Act provides various points of reference for the monitoring system developed in this paper. It can be subsumed under the obligations for high-risk AI systems and could therefore be the basis for technical norms and standards which concretise these obligations.

4. Implementing a Penalty System Under Direct Regulatory Law

161 Finally, the question arises whether in regulatory law, as opposed to liability law (see above), the penalty system described in Section B.II.2.b could be established. Based on general rule of law considerations, the fact that such a penalty system involves decisions imposing disadvantageous effects (fines and other penalties) on the regulatees then

constitutes an interference with the fundamental rights of the regulatees which requires, among other things, a legal basis.

162 One could consider introducing the penalty system as part of the market surveillance procedure pursuant to art. 72 ff. of the AI Act. Since market surveillance falls within the member states' competency, market surveillance authorities at national or federal level would be responsible. According to art. 74 sec. 1 of the AI Act, the general rules regarding market surveillance laid down in the Market Surveillance Regulation (MSR)[159] apply to the national competent authorities for AI market surveillance, too.

163 The MSR grants market surveillance authorities the power to restrict the rights of economic operators at various points. In particular, MSR provides for the following measures:

- Art. 14 stipulates the powers of market surveillance authorities, allowing them, inter alia, to impose penalties on economic operators (see art. 14 sec. 4 (i) MSR).
- Pursuant to art. 41 sec. 1 MSR, the Member States shall lay down the rules on penalties in case market surveillance obligations have been breached. These obligations can either result from the MSR or from sector-specific legislation listed in annex II. The Member States have the responsibility to take all measures necessary to ensure that those penalties are implemented in accordance with national law.
- According to art. 11 sec. 1 (b) MSR, market surveillance authorities are allowed to take appropriate and proportionate measures where the economic operator fails to comply with Union legislation and to take corrective action afterwards.

164 Hence, the penalty system introduced above could be implemented as part of the market surveillance according to the AI Act in conjunction with the MSR.

5. The Competence Question

165 From a legal and also from a practical perspective it is essential to determine who will be responsible for continuous monitoring, for random testing, and for the penalty system. Generally, the role of a review entity (and in particular the CA introduced in Section B.II.2.a) could either be taken up by a state authority (or authorised third party), the AI deployer, or the user.

159 Regulation of the European Parliament and of the Council 2019/1020 of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, 2019 O.J. (L 169) 1.

166 A state authority (or authorised third party) would have the advantage of being neutral and having no conflict of interest. Furthermore, since we do not suggest checking *all* outputs of *all* AI systems, but instead do random sampling from a limited number of (Big) AI systems, the workforce and expertise required is limited, especially when considering the benefits associated with the monitoring system. The effort required to equip government agencies as well as third parties sufficiently to carry out monitoring would therefore be moderate. In addition to that, state authorities would be the only party who could act as market surveillance authority and have the power to impose fines.

167 AI deployers, on the other hand, generally have a higher level of expertise with regard to the AI systems they have built and run and will tend to be better able to increase the workforce. It is clear who they are, compared to AI providers who may be numerous and difficult to identify in the supply chain. And self-certification is already a process used in other industries. However, shifting the competency to the deployers could lead to a regulatory overload (especially for SMEs). In addition to that, it is questionable how effective this form of self-regulation will be. Finally, the user appears to be the least suitable party, as the user lacks the expertise, resources and competence to impose fines.

168 Hence, imposing a penalty system can and should only be done by the responsible state authorities. Testing and monitoring require more legal than technical knowledge, once the system is installed. For the system to work, it is essential to guarantee independence and impartiality of the testing entity. If the testing is done by AI deployers, the testing entity must have a position comparable to the data protection officer (DPO).

D. Conclusion

169 This paper presents new and alternative avenues for regulating AI, understood in the broad sense of state intervention that directly or indirectly affects how AI is developed, operated and deployed. In doing so, it also identifies two new regulatory axes to be addressed - Big/Small AI and Big/Small Teams - and the challenges associated with designing tools that are effective and tailored to these axes. Against this background, our proposals in this paper thus fall into two broad categories:

- Forms of strict liability, as a means to incentivise operators of Small (and Big) AI systems (where direct regulation is unlikely to be effective, and often unnecessary) to take steps preventing or at least minimising harm resulting from their AI.

- A mechanism of constant monitoring through random sampling with human testing and determination, as well as associated cumulative "penalty points", to control the impact of Big AI systems on individuals and society before they cause large-scale, and in particular systematic, harm to society.

170 Our proposals aim to have the maximal possible impact on the behaviour of AI operators and on where and how AI is used, so as to prevent harm and encourage responsible development and use of AI. In our view, these proposals are implementable, policeable, enforceable, cost-effective, and technologically-neutral; they do not focus on what the AI *is*, but rather what it *does*.

171 While it is the (impact) size of the AI that appears to be the focus here, the size of the Team and the high (but not absolute) correlation between the two help us to understand the risks and challenges of *what* AI is doing, and *who* is doing it. Ultimately, our proposals have been designed to address the risks outlined in our 2x2 matrix of AI types. Whether the regulatory and legal implementation tools we discuss will suffice *on their own* to address our concerns remains to be seen, and indeed we propose these tools to operate alongside or concurrently with others. However, what is clear in our view is that any overarching AI regulatory system which overlooks or discounts the risks outlined in our 2x2 matrix of AI types will almost certainly leave "regulatory gaps" through which AI will undoubtedly slip and cause harm.

172 Turning to the current framework for implementation and recent legislative action on AI, in particular the EU AI Act, the constant monitoring mechanism arguably may be realised as part of the AI Act's provider obligations regarding high-risk AI systems. However, the concept of risk classification under the AI Act does not perfectly coincide with our proposed concept of Big AI/Small AI. The penalty point system may be realised as part of the existing EU market surveillance regime. Concerning private (tort) liability, the old laws only provide for negligence-based responsibility. However, the new Product Liability Directive might even lead to a situation where operators of all types of AI can be held liable to an appropriate extent. We hope our contribution stimulates the debate about regulatory strategies concerning AI and invite other researchers to devise further approaches.

173 With regard to very large online platforms and search engines (which generally will rely on AI and other means of automation for content recommendation and moderation, and in supplying their services more generally), the enhanced responsibility framework under the DSA serves as a useful starting point for implementing our proposals. These platforms could implement an 'internal' version of our monitoring system as part of their risk management commitments - while the auditing and other external controls foreseen by the DSA fall significantly short of the constant monitoring with human oversight our check authority is meant to undertake. At the same time, the potential for broader and often diffuse societal harm posed by large platforms and their reliance on AI will often not be detectable and hence preventable through a constant monitoring system that necessarily produces swift results based on tests for harms primarily observable from the AI output alone. In light of what is at stake for individuals and societies, we certainly hope that all actors involved seriously engage with the risk assessment and auditing tools envisioned by the DSA.[160]

174 And finally, the risk to society of AI-related technologies is now well-appreciated by many. However, the focus of such awareness is often on a single, high-impact AI or high-impact event, such as AI starting a (nuclear) war, AI taking over (all) digital systems, an AI-powered robot army rising up, or similar dystopian scenarios. Our approach deals with the AI risk posed by a "death by 1000 cuts", whereby either a vast number of (undetectable) AI systems (Small AI) start creating overall weaknesses in society through their undesirable and harmful impact, or one AI system does likewise through a vast series of mini-impacts (Big AI), no single one of which is catastrophic but in aggregate can systematically shift how society operates, potentially for the worse. Each of these pose the same sort of threat to humanity that mosquitoes have throughout history; spreading harm to many, and difficult to detect and deal with individually. No one mosquito can be considered a significant threat, but it is estimated that over half of human deaths through history can be attributed to mosquitoes.[161] Without rapid, decisive, and impactful action, humanity may find itself facing the same sort of "undetectable" threat that it faced in the past with mosquitoes: people around us will start becoming ill, perhaps even dying and we won't really know why, or how to stop it.

160 Early analysis of DSA compliance by VLOPSEs suggests that this not an easy task in practice; see Sekwenz et al., *From Reports to Reality: Testing Consistency in Instagram's Digital Services Act Compliance Data*, 2025, arXiv:2507.01787.

161 Winegard, *The Mosquito: A Human History of Our Deadliest Predator*, E P Dutton (Aug. 6, 2019).

The Metaverse as a Privacy Mosaic: Framing Data Protection in Hyper-Granular, Data-Drenched Digital Ecosystems

by **Pablo Trigo Kramcsák and Vagelis Papakonstantinou ***

Abstract: The Metaverse unfolds as a hybrid reality, fusing physical presence with digital immersion into a sprawling ecosystem powered by continuous streams of real-time data. Behind its allure lies a sophisticated architecture, a complex web of technologies and platforms that relentlessly captures, analyses, and reshapes vast volumes of personal information. This data proliferation magnifies the stakes of privacy in ways unseen in traditional digital spaces. The Metaverse's hyper-connected nature exposes users to a novel scale of surveillance and inference, challenging established data protection norms.

This study examines these challenges in depth, moving beyond surface-level analysis to explore the intricate ecosystem of data flows that underpin virtual interactions. It proposes a tripartite taxonomy of metaverse data, distinguishing between declared, observed, and inferred data, and offering a framework to identify specific data protection risks. In examining these dynamics, the study considers the divergent governance dynamics inherent in both centralised and decentralised metaverse architectures, recognising that these structural differences profoundly influence privacy challenges and mitigation strategies. This nuanced approach supports the development of targeted, context-sensitive strategies that embed privacy as a foundational element of the Metaverse.

Safeguarding data privacy in the Metaverse demands more than technical fixes; it calls for a fundamental rethinking of how privacy is integrated into its very architecture. This study focuses on advancing Data Protection by Design from a theoretical concept to a practical element embedded within the Metaverse Frameworks. Effective privacy protection in this setting requires a coordinated and systematic implementation of technical, organisational, and governance measures across all relevant stakeholders.

Keywords: Metaverse; virtual worlds; GDPR: data protection by design; inferred data

© 2026 Pablo Trigo Kramcsák and Vagelis Papakonstantinou

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: Pablo Trigo Kramcsák and Vagelis Papakonstantinou, The Metaverse as a Privacy Mosaic: Framing Data Protection in Hyper-Granular, Data-Drenched Digital Ecosystemsl, 17 (2026) JIPITEC 141 para 1.

*

Pablo Trigo Kramcsák: PhD Researcher at the Law, Science, Technology and Society Research Group, Vrije Universiteit Brussel (VUB). Pleinlaan 2, Elsene, 1050, Brussel, België, Belgium. E-mail: pablo.rodrigo.trigo.kramcsak@vub.be. https://orcid.org/0000-0003-2385-4722

Vagelis Papakonstantinou Professor of Law, Faculty of Law and Criminology, Vrije Universiteit Brussel (VUB). Pleinlaan 2, Elsene, 1050, Brussel, België, Belgium. E-mail: Vagelis.Papakonstantinou@vub.be https://orcid.org/0000-0002-2536-2951

Both authors express their gratitude to the discussants and participants of the webinar "Personal Data Processing within Immersive Virtual Worlds: Compliance Challenges of the GDPR", hosted by the Chair for the Responsible Development of the Metaverse at the University of Alicante (Spain) on 31 October 2023.

This publication was financed by the Chair for the Responsible Development of the Metaverse of the University of Alicante (Spain).

A. Introduction

1 The Metaverse has evolved from conceptual projections into environments where physical and digital dimensions intersect. Enabled by interconnected networks and Internet-based infrastructures, metaverse spaces operate through real-time interactions, sustained by the ongoing collection, transmission, and processing of personal data.

2 As immersive, metaverse-oriented applications expand and weave more deeply into social, economic, and cultural spheres, they reveal mounting tension between their underlying information streams and the constraints of traditional data protection standards. Such friction calls for a critical re-evaluation of how privacy and individual autonomy are preserved in data-intensive settings characterised by continuous processing, widespread sharing, and complex networked systems. To address this, it is necessary to identify the core privacy challenges faced by individuals interacting with these environments, while also exploring how to design metaverse ecosystems that are compliant with data protection requirements. At the same time, the conceptual fluidity of "metaverse-type" platforms complicates efforts to delineate its legal, social, and technological boundaries, underscoring the need for analytical frameworks capable of bridging existing regulatory structures with the data-driven dynamics of metaverse environments.

3 This paper examines the data privacy implications of immersive, metaverse-oriented applications through the lens of data protection law. Rather than framing "the Metaverse" as a sui generis digital realm, it situates these environments within the broader landscape of application-layer services, focusing on the taxonomies of data harvesting and the regulatory challenges they pose under existing frameworks. To develop a comprehensive understanding, Section B delineates the foundational components of the Metaverse, critically analysing its data-intensive infrastructures, enabling technologies, and immersive attributes that collectively amplify its systemic complexity.

4 Building upon this foundation, Section C turns to the structural distinctions between centralised and decentralised metaverse platforms. It shows how these different configurations influence core issues of data governance, user control, and the allocation of responsibilities for privacy protection. These variations are not simply architectural; they bear directly on the legal frameworks applicable to different metaverse scenarios.

5 Section D proceeds by classifying the types of personal data processed within metaverse environments. The discussion ranges from basic identifiers to more complex and sensitive forms of data, including biometric information and emotional or behavioural responses. This taxonomy is essential to grasp the scope and granularity of personal data implicated in these contexts.

6 Section E then examines a set of personal data protection concerns that are either intensified or reconfigured by the Metaverse. It contrasts these with challenges typical of traditional online environments, highlighting how immersive features heighten existing risks.

7 Finally, Section F analyses the systematic integration of a Data Protection by Design (DPbD) framework within metaverse ecosystems. Such an approach, if rigorously implemented, offers a path toward embedding data protection principles directly into the architecture of these environments, rather than treating privacy as an external or secondary consideration. In doing so, the analysis also reflects on the conditions under which such integration could be meaningfully achieved.

B. The Shape of the Metaverse: Conceptual Fluidity and Technical Foundations

8 The word *Metaverse* resists straightforward definition,[1] having evolved into something of a conceptual wildcard in contemporary discourse. Its definitional ambiguity invites interpretation but also ensures it remains a moving target for analysis. Often described as a "catch-all" term,[2] the Metaverse is commonly envisioned as an immersive environment populated with 3D computer-rendered objects, spaces, and people,[3] where users interact through virtual self-representations known as avatars.[4]

1 European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, 'Metaverse - Study requested by the JURI Committee' (2023), 10 https://www.europarl.europa.eu/RegData/etudes/STUD/2023/751222/IPOL_STU(2023)751222_EN.pdf accessed 10 July 2025.

2 Luciano Floridi, 'Metaverse: a Matter of Experience' (2022) 35 *Philosophy and Technology* 1, 73.

3 Mariana Olaizola Rosenblat, 'Reality Check: How to Protect Human Rights in the 3D Immersive Web' (2023) NYU Stern Center for Business and Human Rights, September 2023, 1 https://bhr.stern.nyu.edu/publication/reality-check-how-to-protect-human-rights-in-the-3d-immersive-web/ accessed 10 July 2025.

4 Ben Chester Cheong, 'Avatars in the metaverse: potential legal issues and remedies' (2022) *Int. Cybersecur. Law Rev* 3, 467–494, 472.

9 Despite this openness, several recurring features help to sketch its contouring architecture. Virtual worlds are generally understood to possess three fundamental characteristics: persistence and real-time functionality, immersive user experiences, and integrated economic and social interaction mechanisms.[5] Supporting these features requires four core technical attributes: Realism, fostering emotional immersion and a sense of presence; Ubiquity, ensuring consistent access and identity across devices; Interoperability, allowing seamless interaction across platforms; and Scalability, enabling the system to accommodate growing user demand without compromising performance.[6] From a data governance perspective, these characteristics are significant because they shape and intensify the collection, processing, and inference of personal and non-personal data across multiple technological modalities.

10 Although often presented as a rapidly shifting frontier, the Metaverse has so far experienced measured—rather than transformative—advances in autonomy, capacity, and functionality. Developments in the Internet of Things (IoT), virtual reality (VR), augmented reality (AR), mixed reality (MR), and extended reality (XR),[7] alongside improvements in internet infrastructure and device performance, have contributed to the creation of more immersive digital spaces.[8] Virtual environments increasingly feature broader application scopes, improved cross-technology compatibility, and smoother integration with smart devices, giving rise to more convincing and interactive experiences.

11 Public and private entities are exploring ways to leverage metaverse environments, integrating them into their processes, services, and business models,[9] thereby impacting a variety of activities and industries.[10] Central to this development are the social and economic interactions that drive these virtual worlds.[11] As a result, there is increasing attention and debate about the potential socio-economic implications of the Metaverse on a global scale.[12]

12 The term Metaverse can be used both in the singular and the plural. As a singular concept, it represents a unified and interconnected digital universe, that encompasses multiple platforms and environments, similar to the internet or cyberspace. In contrast, using "Metaverses" in the plural acknowledges the current reality of multiple distinct virtual worlds, each with its own architecture, governance, and user base. This plural form, which has drawn most scholarly attention over recent decades,[13] reflects the fragmentation and diversity of virtual spaces (as "micro-universes") shaped by varying providers, technologies, and regulatory frameworks. The singular emphasises the aspirational and integrated vision of a comprehensive digital environment, while the plural highlights the practical multiplicity and heterogeneity of existing immersive ecosystems.[14]

13 This study takes a functional and conceptual view of the Metaverse as a new data-driven stage of the Web, emphasising the interactions, data flows, and governance arrangements that link its diverse

5 European Commission and the Multi Stakeholder Platform on ICT Standardisation (MSP), 'Rolling Plan for ICT Standardisation - Metaverse' (2023) <https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/metaverse#> accessed 10 July 2025.

6 European Parliamentary Research Service (EPRS), 'Metaverse: Opportunities, Risks and Policy Implications' (2022) <https://www.europarl.europa.eu/cmsdata/268589/eprs-briefing-metaverse_EN.pdf> accessed 10 July 2025.

7 Agencia Española de Protección de Datos, 'Metaverse and Privacy' (www.aepd.es, 29 September 2022) <https://www.aepd.es/en/prensa-y-comunicacion/blog/metaverse-and-privacy> accessed 10 July 2025. VR, AR, MR, and XR are key concepts in the evolving landscape of immersive technologies. VR provides a fully immersive experience within a computer-generated environment, while AR enhances the real world by overlaying it with digital content. MR takes AR a step further by enabling users to interact with and manipulate digital objects within their physical surroundings. XR serves as an umbrella term that encompasses all these forms of computer-altered realities (Olaizola Rosenblat (n 3) 4).

8 Arjun Nagendran and others, 'Avatar Led Interventions in the Metaverse Reveal That Interpersonal Effectiveness Can Be Measured, Predicted, and Improved' (2022) 12 *Scientific Reports* 1, 21892.

9 Yogesh K Dwivedi and others, 'Metaverse beyond the Hype: Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy' (2022) 66 *International Journal of Information Management* 102542, 2.

10 Lau Christensen and Alex Robinson, 'The Potential Global Economic Impact of the Metaverse' (2022) Analysis Group, 2 https://www.analysisgroup.com/globalassets/insights/publishing/2022-the-potential-global-economic-impact-of-the-metaverse.pdf accessed 10 July 2025.

11 Mitchell Goldberg and Fabian Schär, 'Metaverse Governance: An Empirical Analysis of Voting within Decentralized Autonomous Organizations' (2023) 160 *Journal of Business Research* 113764.

12 See, e.g., World Economic Forum, 'Social Implications of the Metaverse' (2023) https://www3.weforum.org/docs/WEF_Social_Implications_of_the_Metaverse%20_2023.pdf accessed 10 July 2025.

13 Lorena Arismendy Mengual, 'Liability for Wrongful Behaviour in the Metaverse' (2024) 15(3) *JIPITEC* 229, 230.

14 "[I]t may be more appropriate to refer to metaverses in the plural, assuming that there will not be a single metaverse platform." (Dwivedi et al. (n 9) 25).

applications and platforms. It encompasses a wide array of applications, platforms, services, and data-intensive technologies—both centralised and distributed—that operate atop digital infrastructures, rather than constituting isolated enclaves. These applications engage with users and one another, while operators and service providers intervene in governance, management, and data-handling processes. The analysis focuses on the shared logics of data processing across these environments, while recognising technological distinctions where they are relevant for data privacy implications.

14 To understand these shared logics in practice, it is necessary to examine the flows of information that traverse both immersive and physical environments, and the infrastructures that enable such continuous data processing.

15 The Metaverse integrates inputs from the physical world with data generated within virtual environments, creating a multifaceted informational landscape. This data varies in nature. Some remain static, while others, such as location and sensor readings, are dynamic and continuously evolving.[15] Supporting this complexity is an equally sophisticated infrastructure composed of interconnected sensors, actuators, and networks. Sensors capture real-time information from the environment, while actuators convert digital commands into physical responses.[16] Together, these components establish a bidirectional interface that facilitates interaction between physical and virtual realms, a prerequisite for Metaverse functionality.[17]

C. Who Holds the Keys? Typing the Metaverse(s) Through Governance and Structure

16 Digital environments commonly described as metaverse applications are not merely technical infrastructures; they function as social and economic systems that reflect and shape the rules governing participant interaction. Governance arrangements (including the allocation of power, the structuring of authority, and the distribution of control over resources) not only influence how users engage with these environments but also have significant implications for privacy, data protection, security, and individual agency.

17 Platforms occupy a central role in structuring the Metaverse, shaping how people move, interact, and make sense of these digital spaces. They set the rules of participation, decide who can access what, and guide how information is collected, used, and shared. Within the Metaverse, these platforms can be broadly distinguished according to their governance and technical structures as centralised or decentralised.[18] These models do more than define technical and operational structures; they fundamentally shape the governance frameworks that regulate power, control, and decision-making within these virtual environments. Centralised metaverse platforms, typically large-scale virtual environments, concentrate governance in a single entity or a small group of actors, establishing clear hierarchies and tightly controlled spaces in which authority, decision-making, and resource allocation reside with a limited few. In contrast, decentralised settings aim for a more distributed form of governance, granting users greater autonomy and influence over their virtual interactions. However, the distinction is not always absolute. Some platforms integrate elements from both models, resulting in hybrid systems.[19]

18 A centralised environment resembles traditional internet platforms, where a few entities have significant power over the infrastructure, governance, and data. Such configurations rely on central servers managed by a single entity.[20] A central authority oversees the entire ecosystem, determining and enforcing rules, policies, and governance, while users wield limited influence over decision-making processes. This central stakeholder often exerts substantial control over user data, encompassing its collection, copy, storage, and dissemination (for example, managing user identities, authentication, and profiles).

19 Decentralised platforms, by contrast, operate on distributed architectures, where governance, technical infrastructure, and key components are spread and shared across multiple nodes or participants.[21] In principle, this model allows for

15 Beng Chin Ooi and others, 'The Metaverse Data Deluge: What Can We Do About It?' in *IEEE 39th International Conference on Data Engineering (ICDE)* (Institute of Electrical and Electronics Engineers (IEEE), 2023), 3675-3687, 3678.

16 Xinyi Tu and others, 'Architecture for data-centric and semantic-enhanced industrial metaverse: Bridging physical factories and virtual landscape' (2024) *Journal of Manufacturing Systems* 74, 965-979, 968.

17 Tero Kaarlela and others, 'Towards Metaverse: Utilizing Extended Reality and Digital Twins to Control Robotic Systems' (2023) 12 *Actuators* 6, 219.

18 Liang Yang and others, 'Interoperability of the Metaverse: A Digital Ecosystem Perspective Review' (SSRN, 13 June 2024) https://ssrn.com/abstract=4929167 accessed 12 July 2025, 9.

19 Ikram Ud Din and others, 'Integration of IoT and Blockchain for Decentralized Management and Ownership in the Metaverse' (2023) 36 *International Journal of Communication Systems* e5612.

20 Yang et al. (n 18).

21 Omar Hashash and others, 'Towards a Decentralized Metaverse: Synchronized Orchestration of Digital Twins

greater user autonomy and ownership, as well as peer-to-peer interactions. This has significant implications in how digital assets are managed, how content moderation is approached, and how compatibility across different clients and devices is supported,[22] potentially enabling more flexible metaverse environments. It often uses blockchain technology and distributed protocols to support its features, which impacts the levels of decentralisation.[23] While stakeholders and users may engage actively in the governance and development of this model, they inevitably face the significant technical complexities and coordination challenges embedded within its inherently distributed structure.

20 Hybrid platforms combine elements of both models, integrating centralised oversight with decentralised participation or technical features. These systems illustrate the spectrum of governance approaches within the Metaverse, highlighting the trade-offs between control, user agency, and technical complexity.

D. Dissecting Metaverse Personal Data Processing: A Taxonomy of What Is Collected, Generated, and Analysed

21 Viewed through the lens of its core characteristics and technical foundations, the Metaverse is made up of interconnected applications and services that work together to create continuity, responsiveness, and a palpable sense of presence. Its persistence, real-time functionality, and immersive nature are not incidental features: they depend on the constant generation and processing of user data. Rather than remaining passive spectators, users serve as continuous sources of input in the creation of metaverse objects,[24] feeding ubiquitous and realistic environments engineered to track, analyse, and adapt to their every action. This dynamic is sustained through multi-sensor fusion,[25] whereby streams of data such as gaze direction, voice, motion, and biometric signals are continuously captured and processed. These inputs enable the system to respond in real time and tailor interactions to individual users, reinforcing the sense of immersion.

22 The sheer complexity and interdependence of these data flows resist simplification into mere categories. Beyond the traditional distinction between special and non-special categories of personal data, often framed as sensitive and non-sensitive, a nuanced taxonomy is necessary to differentiate data according to its origins, modes of collection, and methods of production. Such a taxonomy is crucial, as it exposes the varying degrees of control, responsibility, and legal significance inherent in each type, thereby laying the groundwork for meaningful governance and accountability.

23 Broadly speaking, personal data processed within metaverse applications can be classified into three principal categories: declared data, which users voluntarily provide;[26] observed data, which emerges from their behaviours, interactions, and preferences;[27] and inferred data, which is generated through the analysis and correlation of existing information.[28] This tripartite taxonomy thus aligns closely with the technical architectures and operational realities in the Metaverse, encompassing all data generated and used in it. Together, these categories correspond to distinct stages of data flow, beginning with user engagement and sensing and progressing to complex computation, thereby offering a comprehensive framework for understanding how information is created,

and Sub-Metaverses' in *Proceedings of the IEEE International Conference on Communications (ICC '23)* (Institute of Electrical and Electronics Engineers (IEEE), 2023), 1905-1910, 1905.

22 Goldberg and Schär (n 11).

23 Nir Kshetri, 'A Typology of Metaverses' (2022) 55 *Computer* 12, 150-155.

24 Thien Huynh-The and others, 'Blockchain for the metaverse: A Review' (2023) *Future Generation Computer Systems* 143, 401–419, 406.

25 Ruoyu Zhao and others, 'Metaverse: Security and Privacy Concerns' (2023) 3 *Journal of Metaverse* 2, 93-99.

26 It has been defined in the literature as 'data actively and knowingly provided by the data subject' (Article 29 Data Protection Working Party, 'Guidelines on the Right to Data Portability – WP 242 rev.01, 5 April 2017' (2017) 10 https://ec.europa.eu/newsroom/article29/items/611233 accessed 10 July 2025).

27 A useful way to describe it is as 'data provided by the data subject by virtue of the use of the service or the device' (Article 29 Data Protection Working Party, 'Guidelines on the Right to Data Portability – WP 242 rev.01, 5 April 2017' (2017) 10 https://ec.europa.eu/newsroom/article29/items/611233 accessed 10 July 2025).

28 Inferred data (or derived data) 'refers to data which is created by the controller on the basis of the data provided by the data subject (regardless of whether these data were observed or actively provided by the data subject, or a combination thereof)' (European Data Protection Board, 'Guidelines 8/2020 on the Targeting of Social Media Users. Version 2.0. Adopted on 13 April 2021' (2021) https://www.edpb.europa.eu/system/files/2021-04/edpb_guidelines_082020_on_the_targeting_of_social_media_users_en.pdf accessed 10 July 2025, 24.).

transformed, and mobilised within the metaverse systems.

I. Declared Data

24 This data represents information users willingly and consciously share while interacting in virtual worlds. The significance of declared data lies in its direct and explicit nature. At a general level, this category encompasses:

25 (1) User profiles and personal details. User profiles are a key component of declared data, allowing individuals to create and manage their digital identity. These profiles typically include basic information such as name, age, and gender, alongside demographic data, preferences, social connections, and virtual achievements.[29] Notably, this data underpins the creation and ongoing customisation of avatars, which act as the user's digital embodiment. Serving as the primary interface for interaction and communication, avatars anchor the user's presence and facilitate meaningful engagement within the Metaverse.[30] Declared data shapes avatars' appearance, behaviour, and backstory, reflecting users' desired identity and roles within virtual settings. It forms a curated digital self, crafted through personal choices in metaverse social contexts.

26 (2) Preferences. During account setup or at any point thereafter, users can select their interests, hobbies, and specific preferences, allowing for a more personalised virtual experience. These preferences can range from choosing themes for content to specifying the type of interactions they wish to have, reflecting their personal inclinations and goals within the virtual world.

II. Observed Data

27 This category encompasses data generated through users' actions, interactions, and behaviours. The Metaverse's technological features extend the capacity not only to collect but also to observe and record detailed information from both physical and virtual worlds, often integrating inputs from service providers or third parties, thereby broadening the scope of data capture. Metaverse devices and sensors often draw on information from the physical world to enhance immersion,[31] relying on inputs from service providers or third parties. The underlying goal is to synchronise input and output modalities, fostering a more seamless and integrated experience.[32] Observed data may include a broad spectrum of elements, such as:

28 (1) User activities. This encompasses data generated through users' movements, interactions with virtual objects, and in-world behaviours. Its implications extend beyond active participants, potentially impacting individuals not directly engaged in the Metaverse. For example, VR headsets equipped with external cameras and motion sensors may capture data from users' physical surroundings,[33] including bystanders who fall within the device's tracking range.[34] Such incidental data collection can erode anonymity, exposing individuals to forms of digital fingerprinting.[35]

29 (2) Sensorial data. Derived from users' embodied interactions and perceptual engagement, this data includes visual inputs obtained through micro-optics and image recognition systems, auditory information such as voice and ambient soundscapes, and tactile feedback mediated by haptic gloves, bodysuits, touchscreens, or biometric sensors. More broadly, it draws on corporeal cues (eye movements, facial micro-expressions, postural shifts, and gestural patterns)[36] that serve as proxies for affective or cognitive states. For instance, iris dynamics captured by VR headsets or posture variations detected by motion controllers may provide insights into a user's emotional or attentional orientation.[37]

30 (3) Communication logs. Capturing the full spectrum of interaction, these logs record both the content and metadata of user exchanges, including voice chats, written messages, gestures, and symbolic

29 See Daniel Zimmermann, Anna Wehler and Kai Kaspar, 'Self-representation through avatars in digital environments' (2023) 42 *Current Psychology* 21775.

30 Martin Ebers, 'Avatars and the Protection of Digital Identities in the Metaverse' (MetaverseUA Research Paper No 3, 2024) 4 https://metaversechair.ua.es/working-papers/ accessed 12 July 2025.

31 Lik-Hang Lee and others, 'All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda' (2021) 14 *Journal of Latex Class Files* 8, 41.

32 Ibid.

33 Bristows, 'Data protection in the metaverse' (www.bristows.com, 27 October 2022) https://www.bristows.com/news/data-protection-in-the-metaverse/ accessed 13 July 2025.

34 Vasilis Xynogalas and Mark Leiser, 'The Metaverse: searching for compliance with the General Data Protection Regulation' (2024) 14 *International Data Privacy Law* 2, 89-105, 100.

35 Daniel Berrick and Jameson Spivack, 'Understanding Extended Reality Technology & Data Flows' (Future of Privacy Forum Blog, 17 November 2022) https://fpf.org/blog/understanding-extended-reality-technology-data-flows-privacy-and-data-protection-risks-and-mitigation-strategies/ accessed 13 July 2025.

36 Olaizola Rosenblat (n 3).

37 Agencia Española de Protección de Datos (n 7).

communication alike. More than simple records, they could serve as a dynamic archive, enabling fine-grained analysis of social dynamics, linguistic patterns, and relational flows.

31 (4) Location tracking. The Metaverse relies on geospatial technologies to fuse physical and virtual spaces, supporting location-based services and digitally grounded experiences through precise spatial data.[38] In this context, location data entails more than simple coordinates. It captures users' movements, choices of virtual locations, and temporal patterns across virtual environments. This continuous mapping reveals habits of presence and affinities for specific areas, providing input for behavioural analytics.

32 (5) Virtual transactions. Each in-game purchase, currency exchange, or asset transfer produces economic metadata that extends beyond the transaction itself. For instance, in decentralised virtual environments, inventory management systems can monitor the movement and ownership of digital assets over time.[39] These transactional records form detailed consumption profiles, revealing not only what users acquire but also the timing, methods, and contextual factors influencing their decisions.

III. Inferred Data

33 This category consists of information not directly disclosed or recorded but rather generated through analytical processes applied to human traits and behaviours by third-party actors.[40] It involves interpreting patterns drawn from existing data, turning raw inputs into new insights. As a derivative form of data,[41] it reflects knowledge produced from both declared and observed information, often through advanced analytical techniques, including AI-driven decision-making systems.[42] For example, motion-based fingerprinting can reveal specific traits about a user, and when combined with other tracked data, it can significantly enhance identification accuracy.[43] Similarly, a wide range of personal information can be inferred simply by analysing motion data captured by VR devices,[44] potentially exposing sensitive aspects of an individual's life, including medical conditions, physical impairments, or past traumas.[45]

34 Inferred data quietly shapes the architecture of the Metaverse, drawing on users' expressed preferences and behavioural patterns to generate detailed profiles. These profiles drive personalised features such as content suggestions, virtual interactions, and purchase recommendations. Advertisers rely on this data to align messages with users' likely interests, increasing engagement while raising concerns about manipulation and profiling. At the same time, inferred data supports the automation of tasks and the anticipation of user needs.

E. Fractured Privacy: Mapping Personal Data Risks in the Metaverse

35 As detailed in previous sections, the Metaverse is built upon a distinctly data-centric foundation that reshapes the generation, collection, and use of information. This systemic reconfiguration underscores a fundamental departure from traditional digital environments, leading to privacy challenges that go beyond traditional concerns.

36 Powered by technologies enabling multisensory interactions,[46] such as wearable haptic devices and

38 Theofilos Papadopoulos and others, 'The Metaverse Is Geospatial: A System Model Architecture Integrating Spatial Computing, Digital Twins, and Virtual Worlds' (2025) 14 *ISPRS International Journal of Geo-Information* 126, 7.

39 Durgesh Pandey and Paul Gilmour, 'Accounting Meets Metaverse: Navigating the Intersection Between the Real and Virtual Worlds' (2024) 22(2) *Journal of Financial Reporting and Accounting* 211, 221.

40 Luca Bolognini and Marco Emanuele Carpenelli, 'The future of personal data in the Metaverse' (2022) Istituto Italiano per la Privacy e la Valorizzazione dei Dati https://zenodo.org/records/6413046 accessed 13 July 2025, 8.

41 Mark Elliot and others, 'Inferred Data' in Mark Elliot, Anna M Mandalari, Miranda Mourby and Kieron O'Hara (eds), *Dictionary of Privacy, Data Protection and Information Security* (Edward Elgar Publishing, 2024).

42 AI tools could be relevant, among other aspects, for the operation of automatic digital twins, computer agents and the autonomy of the avatar (Lee et al. (n 31)).

43 Louis Rosenberg, 'Privacy in the Metaverse might be Impossible (new research study)' (www.medium.com, 30 March 2023) https://medium.com/predict/privacy-in-the-metaverse-might-be-impossible-new-research-study-64935481c6de accessed 14 July 2025.

44 Vivek Nair and others, 'Inferring Private Personal Attributes of Virtual Reality Users from Head and Hand Motion Data' (2023) arXiv https://arxiv.org/abs/2305.19198 accessed 14 July 2025.

45 Eleonora Margherita Auletta, Francesca Tugnoli and Giada Iovane, 'Personal data protection in the Metaverse: operational challenges and regulatory uncertainties' (www.ictlc.com, 28 September 2022) https://www.ictlc.com/personal-data-protection-in-the-metaverse-operation-challenges-and-regulatory-uncertainties/?lang=en#_ftn4 accessed 14 July 2025.

46 Stylianos Mystakidis, 'Metaverse' (2022) 2 *Encyclopedia* 1, 486-497, 487.

the Internet of Things,[47] The Metaverse captures a vast spectrum of personal data, gathering far more sensitive information than conventional platforms.[48] Internet-connected wearables continuously monitor users, collecting diverse data types, including physical metrics, user habits and preferences, and communication patterns.[49] Auxiliary devices designed for metaverse applications go further, capturing increasingly intrusive biometric information.[50] The volume and granularity of information XR technologies collect far exceed that of traditional surveillance tools, making them markedly more invasive.[51] Beyond raw data, virtual spaces rely on advanced analytics,[52] allowing it to draw extensive inferences that deepen the scope and sensitivity of knowledge about users.

37 Metaverse's data practices extend their reach beyond corporeal data, raising profound implications for mental autonomy,[53] by progressively penetrating users' cognitive and emotional spheres. While earlier devices were largely limited to tracking basic physiological functions, today's tools map psychological processes, preferences, and emotional states.[54] When deployed together, these systems generate compound effects, both individual and societal, that raise serious and unpredictable risks to rights and freedoms.[55]

47 Kisoo Kim and others, 'Metaverse Wearables for Immersive Digital Healthcare: A Review' (2023) 10 *Advanced Science* 4.

48 Dwivedi et al. (n 9) 8.

49 Lee et al. (n 31) 39.

50 Kewa Jiang, 'Protecting Mental Health Data Privacy in the Metaverse: How Does That Make You Feel?' (2023) 28:1 *Intell Prop & Tech* L J 23, 28.

51 Olaizola Rosenblat (n 3) 9.

52 Leon Anidjar and others, 'The Matrix of Privacy: Data Infrastructure in the AI-Powered Metaverse' (February 18, 2023). Harvard Law & Policy Review, Forthcoming https://ssrn.com/abstract=4363208. It should be noted that convergence of AI and XR is transforming immersive technologies, making AI essential in prediction, personalisation, interaction, speech recognition, and language translation. This evolution raises the question of whether the combination of generative AI and XR will become the next major platform (Susan Ariel Aaronson and Adam Zable, 'Facing Reality: Canada Needs to Think about Extended Reality and AI', CIGI Papers No. 289 — March 2024 (2024) https://www.cigionline.org/static/documents/no.289.pdf accessed 14 July 2025, 6).

53 Council of Europe and IEEE Standards Association, 'The Metaverse and Its Impact on Human Rights, the Rule of Law and Democracy' (2024) https://rm.coe.int/the-metaverse-impact-on-and-its-impact-on-human-rights-the-rule-of-law/1680ae6bce accessed 14 July 2025, 7.

54 Marcello Ienca and Gianclaudio Malgieri, 'Mental data protection and the GDPR' (2022) 9 *Journal of Law and the Biosciences* 1, 2.

55 Agencia Española de Protección de Datos (n 7).

38 In parallel, stakeholders in the metaverse ecosystem actively monetise and trade the data they collect,[56] embracing extraction-driven business models that prioritise continuous tracking and behavioural profiling. This data possesses substantial economic value, both as a standalone commodity and as a resource that can be repurposed, aggregated, and sold across various markets and sectors,[57] including advertising, insurance, and predictive analytics.

39 Metaverse technologies, however, do not create uniform challenges, as each generates data in ways that are technically and legally distinct. VR applications capture highly detailed behavioural and spatial information, AR systems integrate environmental and sensor data, and IoT/XR platforms continuously stream telemetry from connected devices. The intensity, granularity, and type of data thus vary across systems, stretching traditional privacy safeguards to their limits. While existing regulatory frameworks already cover these activities, the sprawling and interconnected nature of metaverse systems' data flows places established protections, particularly the European Union's General Data Protection Regulation (GDPR),[58] under significant pressure.[59] The GDPR's broad material scope covers a wide range of personal data processing,[60] making models that facilitate immersive user interactions clearly subject to its regulatory framework.[61]

40 For instance, Metaverse's extensive and granular data collection dynamics strain the principles of

56 Zefeng Chen and others, 'Metaverse Security and Privacy: An Overview' in *2022 IEEE International Conference on Big Data (Big Data) Dec. 17 2022 to Dec. 20 2022 Osaka, Japan* (Institute of Electrical and Electronics Engineers (IEEE), 2022), 2950-2959.

57 See Ben Egliston and Marcus Carter, 'Critical questions for Facebook's virtual reality: data, power and the metaverse' (2021) 10 *Internet Policy Review* 4.

58 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119/1 (4 May 2016) https://eur- lex.europa.eu/legal- content/EN/TXT/PDF/ ?uri=CELEX:32016R0679&from=EN.

59 Maryant Fernández, 'Call for Evidence on an EU Initiative on Virtual Worlds - BEUC's Response' (www.beuc.eu) <https://www.beuc.eu/sites/default/files/publications/BEUC-X-2023-053_Call_for_evidence_on_an_EU_initiative_on_virtual_worlds.pdf> accessed 16 July 2025.

60 See Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10(1) *Law, Innovation and Technology* 40.

61 Baily Martin, 'Privacy in a Programmed Platform: How the General Data Protection Regulation Applies to the Metaverse' (2022) 36(1) Harvard Journal of Law & Technology 235, 256.

purpose limitation and data minimisation, which are designed to restrict the scope and necessity of personal data processing. These principles presume a degree of control and predictability that is often absent in immersive environments, where data is collected continuously, contextually, and often passively. Additionally, the Metaverse relies on layered technical infrastructures, intersecting systems, and a strong push for interoperability, making the consistent application of privacy protections across actors and platforms particularly difficult. As a result, translating the GDPR's normative commitments into operational safeguards in these environments becomes a deeply complex undertaking.

41 This scenario makes it increasingly necessary to move beyond surface-level assessments and confront the specific regulatory challenges the Metaverse poses. At stake is not merely compliance, but the viability of applying core privacy principles in an environment where data is ambient, inferential, and persistently in motion.

I. Intensive Data Collection and Detailed Profiling

42 Virtual environments' architecture fosters a level of monitoring that exceeds traditional applications,[62] allowing for the fusion of heterogeneous sources,[63] and the continuous tracking of user behaviour with heightened precision.

43 This situation not only gives rise to new privacy concerns but also amplifies existing ones,[64] elevating 'the risk of invasion of privacy'.[65] To illustrate, these spaces could monitor the user's features such as facial expressions, voice inflections, postures and, if allowed, vital signs, including heart rate, blood pressure, respiration rate, pupil dilation, and even galvanic skin response.[66] In addition, the immersive aspect of the Metaverse could reduce the users' ability and perception to make decisions that prevent the excessive collection and processing of personal data.[67]

44 All of this highlights the implications of the Metaverse's data logic, which is rooted in the sheer complexity and heterogeneity of data sources it activates. The Metaverse mirrors and duplicates lived reality,[68] generating data not only from deliberate user actions but also from ambient, often passive, sensory inputs.[69] These inputs, whether behavioural, physiological, or environmental, are persistently recorded, producing a flow of information that is continuous and profoundly difficult to contain.

45 Within sensor-rich, behaviour-driven immersive environments, the capacity to derive inferences extends far beyond personalisation or optimisation. It reshapes how knowledge about individuals is generated and applied, enabling datafication to penetrate deeply into personal identity and decision-making processes. The outcome is the construction of granular, multilayered profiles that surpass the analytical capacities of earlier digital environments.[70]

46 These enhanced profiles create fertile ground for highly targeted and, at times, intrusive marketing practices, including avatar-based promotions.[71] Advertisers can also measure the effectiveness of their campaigns more precisely by leveraging data collected through XR technologies, including eye movements, pupil dilation, and heartbeat.[72] These practices contribute to ongoing surveillance and invade individuals' intimate virtual spaces, undermining their autonomy over digital experiences. The concern lies not merely in the accuracy of these inferences and profiles but in their opaque capacity to shape perceptions, interactions,

62 European Data Protection Supervisor, 'Metaverse' (www.edps.europa.eu) https://edps.europa.eu/press-publications/publications/techsonar/metaverse accessed 16 July 2025.

63 Ooi et al. (n 15).

64 Martin (n 61) 254.

65 Hang Wang and others, 'A Survey on the Metaverse: The State-of-the-Art, Technologies, Applications, and Challenges' (2023) 10 *IEEE Internet of Things Journal* 16, 14671-14688, 14672.

66 Louis Rosenberg, 'The Growing Need for Metaverse Regulation', in Kohei Arai (ed) *Intelligent Systems and Applications. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 544* (Springer, Cham, 2023), 544.

67 Philipp Koehler, 'The Metaverse and some of its emerging challenges for data protection law' (www.taylorwessing.com, 10 October 2022) https://www.taylorwessing.com/en/insights-and-events/insights/2022/10/the-metaverse-and-some-of-its-emerging-challenges-for-data-protection-law accessed 16 July 2025.

68 Janna Anderson and Lee Rainie, 'The Metaverse in 2040' (www.pewresearch.org, 30 June 2022) https://www.pewresearch.org/internet/2022/06/30/the-metaverse-in-2040/ accessed 16 July 2025.

69 Mystakidis (n 46) 488.

70 Agencia Española de Protección de Datos (n 7).

71 See, for example, Kyle Coble, Jay Ratican and James Hutson, 'Beyond the pixelated mirror: Understanding avatar identity and its impact on in-game advertising and consumer behavior' (2023) 4 *Metaverse* 2023 2, 2377.

72 Bristows (n 33).

and decisions, often without meaningful transparency or accountability.

II. Identity, Authentication and Anonymity

47 In virtual spaces, the varied identity requirements across platforms give rise to substantial vulnerabilities for users. These vulnerabilities include threats ranging from identity theft to online harassment.[73] It is particularly important to recognise the extent to which biomechanics may serve as unique identifiers in virtual reality.[74] As users engage with virtual environments, their movements, speech patterns, and other behaviours generate data that may fully reveal their identities.[75]

48 Compounding these risks, avatars often fail to adequately shield users' true identities. They may be linked to personal information or accounts accessible to metaverse service providers and shared with other stakeholders. Even minimal data, such as the basic motion tracking used for interaction in virtual worlds, can uniquely identify individuals, undermining meaningful anonymity.[76] Researchers refer to this as "telemetry data," the minimal dataset needed for natural user interaction,[77] which nevertheless holds outsized power for profiling. Recent studies underscore just how easily VR users can be profiled and deanonymised.[78]

49 The regulatory scope is not confined to explicit data disclosures. While avatars are often designed to support pseudo-anonymity, they can nevertheless expose users' real-world identities through inference. The physical and demographic attributes assigned to an avatar, such as body shape, skin tone or clothing, may closely mirror or idealise aspects of the user's authentic self,[79] creating a connection that undermines anonymity. In some instances, these representations may engage special categories of personal data under Article 9 GDPR, particularly when avatars realistically reflect sensitive characteristics.[80] Assumptions or inferences drawn from avatar features, especially when combined with other data or assessed in the context of specific processing activities, may also qualify as special category data.[81] This includes politically sensitive attributes, sexual orientation and other protected aspects of identity that may be revealed through virtual representation and behavioural patterns.[82]

III. Data Subjects' Consent, Transparency and Control

50 In metaverse environments, many platforms and stakeholders acting as data controllers often fall short of providing clear and accessible information regarding the use of personal data.[83] Likewise, users are not always offered straightforward mechanisms to manage data-sharing preferences or to track how their information is processed over time. As a result, individuals may lack the necessary insight to make informed, specific, and voluntary decisions about their data.[84] This, in turn, diminishes the practical effectiveness of consent as a legal basis for processing, particularly when users do not fully understand the scope, purposes, or potential consequences of the data practices they are subjected to.

51 In this context, multiple data controllers frequently operate simultaneously, processing high volumes of personal data in real time. The pace and scope of data collection complicate efforts to maintain transparency, further blurring the boundaries of

73 Anna Maria Collard, 'Crime in the metaverse is very real. But how do we police a world with no borders or bodies?' (www.weforum.org, 18 August 2022) https://www.weforum.org/agenda/2022/08/crime-punishment-metaverse/ accessed 10 February 2024.

74 Vivek Nair and others, 'Unique Identification of 50,000+ Virtual Reality Users from Head & Hand Motion Data' in *32nd USENIX Security Symposium (USENIX Security 23)* (USENIX Association, 2023), 895-910.

75 World Economic Forum, Metaverse Identity: Defining the Self in a Blended Reality – Insight Report March 2024 (2024) https://www3.weforum.org/docs/WEF_Metaverse_Identity_Defining_the_Self_in_a_Blended_Reality_2024.pdf accessed 16 July 2025.

76 See Ronald Leenes, 'Privacy in the Metaverse' (2008), in Simone Fischer-Hübner and others (eds) *The Future of Identity in the Information Society. Privacy and Identity 2007. IFIP — The International Federation for Information Processing, vol 262* (Springer, Boston, MA, 2008).

77 Rosenberg (n 66).

78 Vivek Nair, Gonzalo Munilla-Garrido and Dawn Song, 'Going Incognito in the Metaverse: Achieving Theoretically Optimal Privacy-Usability Tradeoffs in VR' in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (Association for Computing Machinery, New York, 2023) Article 61, 1–16.

79 Zimmermann et al. (n 29).

80 Bristows (n 33).

81 European Data Protection Board, 'Guidelines 8/2020 on the Targeting of Social Media Users. Version 2.0. Adopted on 13 April 2021' (2021) https://www.edpb.europa.eu/system/files/2021-04/edpb_guidelines_082020_on_the_targeting_of_social_media_users_en.pdf accessed 10 July 2025, 30.

82 European Data Protection Supervisor (n 62).

83 Xynogalas and Leiser (n 34) 99.

84 See, for example, Jasmine Wang and Eddie Liu, 'Data Usage in Social Media' (www.medium.com, 24 December 2020) https://medium.com/big-data-at-berkeley/data-usage-in-social-media-21d606ec4451 accessed 17 July 2025.

consent.[85] Metaverse contexts also merge different purposes for processing, making it difficult for the data subjects to consent for each purpose separately.[86] Providing users with adequate information about what data is processed, how it is used, and for what purposes becomes not only burdensome but structurally difficult. This undermines compliance with the principles of purpose limitation, fairness, and transparency. [87]

52 Two models for operationalising consent in the Metaverse emerge. One adopts a unified consent framework applicable across the entire environment. The other allows each entity to establish and govern its own rules.[88] The effectiveness of either approach hinges on the distribution of power between users and platform stakeholders.[89] Where dominant actors exert control over infrastructure, services, or devices, users' ability to meaningfully consent is significantly weakened. This asymmetry raises concerns regarding the processing of special categories of personal data under Article 9(2) GDPR, which requires explicit consent in the form of an express, unambiguous statement.[90]

53 Consent becomes even more fraught when children are involved. Under the GDPR, valid consent requires that the child has sufficient capacity to grasp the implications of the data processing activity. Where such capacity is lacking, consent cannot be considered informed, and the data controller must instead seek authorisation from a parent or legal guardian.[91]

85 Moerel points out that '[p]rivacy and data security issues will be exacerbated in the metaverse and obtaining any form of meaningful informed consent will prove a practical impossibility' (Lokke Moerel, 'Metaverse and data protection', in Larry A. DiMatteo and Michel Cannarsa (eds) *Research Handbook on the Metaverse and Law* (Edward Elgar Publishing, Cheltenham, 2024) 118-157, 118).

86 Danique Knibbeler, Max Mohrmann and Sarah Zadeh, 'Privacy and security concerns in the metaverse' (www.nautadutilh.com, August 2022) https://www.nautadutilh.com/en/insights/privacy-and-security-concerns-in-the-metaverse/ accessed 17 July 2025.

87 Xynogalas and Leiser (n 34) 99.

88 Koehler (n 67).

89 See Ljubisa Bojic, 'Metaverse through the prism of power and addiction: what will happen when the virtual world becomes more attractive than reality?' (2022) *Eur J Futures Res* 10, 22.

90 Xynogalas and Leiser (n 34) 97.

91 Diletta De Cicco, James Downes and Charles Helleputte, 'No Children in the Metaverse? The Privacy and Safety Risks of Virtual Worlds (and How to Deal with Them)' In *Privacy Technologies and Policy: 11th Annual Privacy Forum, APF 2023, Lyon, France, June 1–2, 2023, Proceedings* (Springer-Verlag, Berlin, Heidelberg, 2023), 87-108, 97.

54 Given these limitations, exploring alternative legal grounds for data processing in the Metaverse becomes essential. One such ground, the performance of a contract under Article 6(1)(b) GDPR, may justify certain data uses. However, this basis introduces its own complications. Controllers must demonstrate that each processing operation is strictly necessary to fulfil a contractual obligation, aligning with users' reasonable expectations. They must also assess whether the terms themselves are fair and whether they disguise excessive data practices under the guise of contractual necessity, a task made even more difficult by the Metaverse's data-intensive architecture and fluid user engagement.[92]

55 Moreover, even when relying on the performance of a contract or other legal bases under Article 6 GDPR, processing special categories of data remains particularly challenging. Such processing requires satisfying one of the specific exceptions enumerated in Article 9(2) GDPR, a threshold that significantly restricts the circumstances under which this sensitive data may be lawfully handled, underscoring the necessity for rigorous legal scrutiny.

IV. Territorial Scope of Data Regulations

56 Metaverses operate on a global scale, relying on the seamless flow of information across platforms, service providers and national borders. This global reach, combined with the involvement of numerous end-users, service providers, and diverse stakeholders, complicates the task of determining which legal frameworks apply to these virtual spaces. The result is often a cumulative layering of overlapping and intersecting regulations that must be reconciled simultaneously. The Metaverse's decentralised structure, extensive interoperability, and the current lack of tailored regulation only deepen this complexity.[93]

57 Pinpointing users' locations within the Metaverse further obscures jurisdictional boundaries. Factors such as the physical location of the individual controlling an avatar, the avatar's position within the virtual environment, and the geographic

92 See European Data Protection Board, 'Opinion 08/2024 on Valid Consent in the Context of Consent or Pay Models Implemented by Large Online Platforms. 17 April 2024', paragraphs 41, 53 and 116. Available at https://www.edpb.europa.eu/system/files/2024-04/edpb_opinion_202408_consentorpay_en.pdf.

93 Schürmann Rosenthal Dreyer, 'The Metaverse: Data Protection in a New Virtual World | SRD' (www.srd-rechtsanwaelte.de, 15 June 2022) https://www.srd-rechtsanwaelte.de/en/blog/metaverse-data-protection/ accessed 17 July 2025.

location of data servers all blur traditional notions of territoriality. As Metaverses develop, the question of which jurisdiction governs disputes involving avatars grows increasingly ambiguous,[94] especially when multiple service providers and virtual identities intersect.

58 Given that data protection laws safeguard fundamental rights that cannot be waived, contractual choice of law provisions are inherently limited in their reach.[95] For example, the GDPR applies extraterritorially to non-EU platforms and controllers processing the personal data of EU residents, whether by offering goods or services or monitoring behaviour within the Union. Moreover, operators cannot segregate EU users' data from that of other users to circumvent these obligations.[96] While contractual clauses may attempt to clarify the relationship between providers and users, their enforceability varies widely across jurisdictions, exposing providers to potential litigation in multiple fora.[97]

V. Responsibility and Controllership

59 The Metaverse complicates data protection compliance by dispersing personal data processing across layered infrastructures and interdependent actors. Interoperable architectures facilitate data flows across multiple actors, blurring the lines of legal accountability as information collected by one party circulates seamlessly across platforms, services and devices.

60 Compounding this complexity is the autonomous data collection by third-party developers, which further fragments control and complicates attribution. As the Metaverse deepens its reach, the web of stakeholders and interactions tightens, making the precise assignment of GDPR-defined data protection roles extraordinarily elusive.[98] Within this tangled ecosystem, joint controllership becomes the norm rather than the exception, as personal data typically flows through an interconnected chain of actors, spanning device manufacturers, platform operators, and third-party services.[99]

61 Under the GDPR, a controller is the entity that determines the purpose ("why") and means ("how") of processing. In the Metaverse, controllers include a wide array of actors who collect or access data within virtual spaces (sometimes directly from users, other times from bystanders). While controllers shape the objectives and parameters of data processing, they may delegate implementation to processors.[100] They may include platform and service providers responsible for developing the infrastructure and devices that enable the Metaverse, and, in certain cases, individuals who interact within these environments as end-users.[101] Broadly, controllers in the Metaverse can be classified into four main categories:

62 (1) Platform Operators/Administrators. These entities shape the Metaverse's architecture, policies, and user experience. They design and maintain the virtual environment and govern interactions within it, from user behaviour to content moderation. While the purpose and scope of data collection vary, they commonly focus on personalising experiences, strengthening security, and optimising the overall performance of metaverse platforms. Beyond data management, these operators play a crucial role in safeguarding Metaverse's security and stability by protecting it from disruptions, cyberattacks, and unauthorised access.

63 (2) Service Providers. These actors deliver a range of services within the Metaverse, ranging from e-commerce platforms and social interaction tools to content hosting environments. Their role extends beyond mere facilitation; by defining the purposes and means through which personal data is processed to enable these services, they assume significant control over user data.

64 (3) Third-Party Developers. Developers of games, applications, or immersive experiences in the metaverse environment may assume the role of data controllers when they independently collect and process personal information. Their role becomes especially significant when setting terms

94 Cheong (n 4) 471.

95 Jie (Jeanne) Huang, 'Applicable Law to Transnational Personal Data: Trends and Dynamics' (2020) 21 *German Law Journal* 1283, 1299.

96 Bristows (n 33).

97 Matthias Artzt and Gary Weingarden, 'Metaverse and Privacy' (2022) 15 *International In-House Counsel Journal* 59, 7763, 7765.

98 Koehler (n 67).

99 Guillaume Couneson and Jennifer Calver, 'Metaverse Series #6: Data in the Metaverse – More Personal than Ever' (Techinsights Linklaters, 22 September 2022) <https://techinsights.linklaters.com/post/102hxlh/metaverse-series-6-data-in-the-metaverse-more-personal-than-ever> accessed 17 July 2025.

100 See European Data Protection Board, 'Guidelines 07/2020 on the Concepts of Controller and Processor in the GDPR. Version 1.0. Adopted on 02 September 2020' (2020) https://www.edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf accessed 17 July 2025.

101 World Economic Forum and Accenture, 'Interoperability in the Metaverse: Briefing Paper' (World Economic Forum 2023) https://www3.weforum.org/docs/WEF_Interoperability_in_the_Metaverse.pdf accessed 20 July 2025.

of engagement or embedding data analytics within their tools, thereby shaping data collection and use.

65 (4) Users. End-users may also function as controllers, particularly when they generate content, interact with other participants, or manage their personal data across metaverse platforms. This role implies processing data about other users or bystanders, involving decisions about virtual interactions, data collection, and sharing.

66 Responsibility in the Metaverse depends on both functional positions and its structural form. In centralised platforms, operators exert significant control over architecture, data flows, and compliance, while service providers generally align with their policies. Third-party developers may act as data controllers or processors depending on the autonomy and scope of their data collection. In decentralised or hybrid systems, responsibilities are distributed across operators, providers, developers, and users, creating shared accountability and more complex governance arrangements.

67 Roles in virtual worlds are fluid and context-dependent, making it difficult to determine who qualifies as a controller, joint controller or processor under the GDPR.[102] This uncertainty makes it harder to apply key data protection principles and can get in the way of users fully exercising their rights.

68 Despite these challenges, the responsibilities of actors within the Metaverse largely fall within existing regulatory frameworks. Platform operators, service providers, third-party developers, and even users can typically be classified under the GDPR's definitions of controllers, joint controllers, or processors, with comparable obligations arising under the Digital Services Act (DSA)[103] and the Digital Markets Act (DMA)[104] where applicable. Rather than requiring entirely new regulatory categories, these frameworks offer the tools to assign accountability, enforce obligations, and protect users' rights, provided regulators and stakeholders interpret and apply the rules with attention to the Metaverse's distributed structures.

VI. Seamless Integration, Interoperability and Portability

69 Metaverse ecosystems encompass a mosaic of distinct systems and technologies, each shaped by different providers.[105] Frictionless movement within and between distinct metaverse platforms hinges on two interconnected challenges: seamless interoperability and effective data portability. Crossing these boundaries exposes users to a patchwork of technical standards, policies, and legal frameworks, complicating efforts to protect user control and ensure consistent safeguards.

70 Interoperability can be defined as the "ability to interact, exchange and make use of data and resulting information to enable movement, transactions and participation across systems, platforms, environments and technologies".[106] Often termed "cross-metaverse interoperability," it ensures a continuous user experience,[107] minimising interruptions and discontinuities as individuals transition between virtual worlds.

71 Complementing interoperability, the right to data portability under Article 20 GDPR empowers data subjects to obtain and transfer their personal data between service providers.[108] Within the Metaverse, this right enhances user autonomy by enabling access, reuse, and transfer of data across platforms, thereby fostering a more user-centric ecosystem. To achieve these goals, common technical standards, such as "standard data formats",[109] are necessary, which mitigate vendor lock-in, i.e., situations where users are confined to a single metaverse platform and unable to move their personal data easily,[110] limiting

102 Knibbeler et al. (n 86).

103 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), OJ L 277/1 (27 October 2022) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2065.

104 Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ L 265/1 (12 October 2022) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R1925.

105 Bristows (n 33).

106 World Economy Forum and Accenture (n 101).

107 Taotao Li and others, 'MetaOpera: A Cross-Metaverse Interoperability Protocol' (2023) 30 *IEEE Wirel. Commun.* 5, 136-143, 136.

108 Sanna Toropainen, 'The right to data portability in the fair data economy – extending the right of individuals to benefit from managing their data' (2023) Sitra https://media.sitra.fi/app/uploads/2023/09/sitra_the-right-to-data-portability-in-the-fair-data-economy-final.pdf accessed 20 July 2025, 13.

109 Franklyn Ohai, 'The Threat Landscape of Extended Reality: Beyond Privacy and Data Protection' (www.law.kuleuven.be, 21 November 2023) https://www.law.kuleuven.be/citip/blog/the-threat-landscape-of-extended-reality-beyond-privacy-and-data-protection/ accessed 20 July 2025.

110 Knibbeler et al. (n 86).

consumer choices.[111] Currently, switching between metaverse ecosystems is difficult for individuals, as each setting has its own protocols, and there is limited interoperability.[112]

72 Interoperability challenges in centralised metaverse platforms could find a more natural remedy in competition law. Many of the largest platforms act as digital gatekeepers under the DMA, granting regulators the authority to enforce openness and, ultimately, enable meaningful data portability. In this way, market-oriented interventions can complement GDPR safeguards, addressing technical bottlenecks and structural barriers.

VII. Data Security Risks

73 Distributed, sensor-driven digital ecosystems amplify cybersecurity risks due to the vast quantities of information harvested by interconnected devices and embedded sensors,[113] particularly with respect to the availability, resilience, and confidentiality of personal data processed.[114] The very attributes that define these spaces, such as social interaction, immersive engagement, real-world simulation and scalability, intensify these data security concerns.[115] Threats arise across multiple layers, encompassing risks related to data inputs, processing, outputs, user interactions and device operation.[116]

74 The scope and complexity of these risks depend on factors such as the platform's architecture (centralised or decentralised), the nature of processing operations and the types of devices employed. Integrating data from the IoT further enlarges the potential attack surface.[117] These vulnerabilities can manifest as personal information leakage, eavesdropping, data theft, unauthorised access, phishing, data injection, broken authentication, insecure design and related exploits.[118]

75 The fluid and real-time exchanges on which metaverse services depend make safeguarding personal information especially challenging.[119] Protections at the device, software component, or unit level, while necessary, are insufficient when applied in isolation, particularly in decentralised systems where fragmented control can create gaps.

VIII. Exercise of Data Subjects' Rights in the Immersive Digital World

76 As noted earlier, Metaverse dissolves geographical boundaries, bringing disparate data protection regimes into constant collision. Within these varying regimes, each with its own regulatory approaches, data protection rights and obligations can diverge significantly, depending on the specific data protection law(s) applicable to a given case.[120] The result is a fragmented legal terrain where procedural requirements, substantive protections and remedies differ sharply. Opaque and overbroad privacy policies deepen the uncertainty, leaving users without a clear sense of how their rights apply in practice.[121]

77 Identity in virtual spaces is often fluid, with users operating through pseudonyms or avatars. In certain cases, the tenuous link between virtual personas and real-world identities can make it difficult to verify requests under the GDPR or other regimes, and to respond appropriately.

78 The governance models of decentralised metaverse settings and distributed ledger technologies introduce further obstacles. Without a central authority, responsibility could splinter, and accountability dissipate. Beyond joint controllership, pinpointing who bears legal liability for privacy breaches becomes a daunting task. Accessing, rectifying or erasing personal data can be particularly difficult when it is stored in encrypted or hashed form. Finally, users might face significant obstacles in identifying or contacting metaverse data controllers or processors, who may obscure their identities by operating through intermediaries or proxies.[122]

111 European Parliamentary Research Service (n 6).

112 Toropainen (n 108) 28.

113 Chris Warin and Delphine Reinhardt, 'Vision: Usable Privacy for XR in the Era of the Metaverse' in *Proceedings of the 2nd European Symposium on Usable Security (EuroUSEC)* (Association for Computing Machinery, 2022), 111-116, 113.

114 Agencia Española de Protección de Datos (n 7).

115 Yan Huang and others, 'Security and Privacy in Metaverse: A Comprehensive Survey' (2023) 6 *Big Data Mining and Analytics* 2, 234-247, 235.

116 Warin and Reinhardt (n 113).

117 IBM Institute for Business Value, 'Seven Bets' (IBM Corporation 2023) https://www.ibm.com/downloads/cas/GN7B276L accessed 20 July 2025, 26.

118 Huang et al. (n 115).

119 Dwivedi et al. (n 9) 8.

120 Artzt and Weingarden (n 97).

121 Natalia Menéndez González and Efe Bozkir, 'Eye-tracking Devices for Virtual and Augmented Reality Metaverse Environments and Their Compatibility with the European Union General Data Protection Regulation' (Robert Schuman Centre for Advanced Studies, Centre for a Digital Society, Working Paper RSC 2023/69, 2023) https://cadmus.eui.eu/bitstream/handle/1814/76162/RSC_WP_2023_69.pdf?sequence=1&isAllowed=y accessed 20 July 2025.

122 Artzt and Weingarden (n 97). "[I]n the operation of the Metaverse, adherence to the rights of the individuals

F. Can Data Protection by Design Truly Work in the Metaverse?

79 Previous sections have revealed how the Metaverse exposes deep tensions within established data protection frameworks, particularly the GDPR's purpose limitation and data minimisation principles. These principles are designed to impose meaningful limits on data collection and processing, yet the Metaverse's immersive, continuous, and highly interconnected data flows challenge their practical enforceability. Far from isolated transactions, data exchanges in these environments are dynamic and layered, complicating efforts to apply regulatory safeguards. The result is a growing friction between the expansive data ambitions of metaverse technologies and the foundational rights-based limits that seek to govern them.

80 Data Protection by Design theoretically offers a pathway to align rapid technological innovation with privacy imperatives; however, the Metaverse's architecture exposes significant fault lines in this ideal. Rather than reflecting a genuine integration of DPbD principles, current systems risk reducing them to performative gestures that obscure deeper regulatory and technical shortcomings. The continuous flow of data, the shifting nature of user identities, and the complex interactions across multiple platforms undermine efforts to embed privacy from the outset. Fragmented governance further hinders the coordination needed for a unified data protection approach.

I. Integrating Privacy into the Core: The Strategic Role of Data Protection by Design

81 DPbD is a forward-looking and comprehensive strategy that embeds privacy into every layer of the system's architecture. It requires elevating privacy beyond an afterthought to a core tenet that shapes the development of information systems, technologies, and data management processes, thereby enhancing regulatory flexibility.[123] This approach, paired with data protection by default, embraces a risk-based framework,[124] that tailors safeguards to the nature, scope, and potential impact of data processing activities.

82 Fundamentally, DPbD recognises that processing of personal data transcends mere technical challenges, encompassing systemic dimensions of governance, architecture, and user empowerment. Data protection and security considerations should be incorporated throughout all stages of processing, starting from the earliest phases of development,[125] enable data processing stakeholders to effectively manage the risks inherent in these operations, being far more effective than retrofitting these measures after a product or service is built.[126]

83 The integration of data protection requirements from the outset and as a default, as mandated by Article 25 of the GDPR, represented an important advancement in the reform of the European personal data framework.[127] The significance of data protection by design in this framework lies not only in its cross-cutting role as a guiding principle,[128] but also in its status as an enforceable legal obligation, exemplifying what is termed "legal protection by design."[129] Its purpose is to embed legal safeguards

had not been feasible in some instances [for example] due to difficulties in identifying the parties responsible for data security" (Vibhushinie Bentotahewa and others, 'Privacy and Security Landscape of Metaverse', in Nitin Naik and others (eds) *Advances in Computational Intelligence Systems: Contributions Presented at the 22nd UK Workshop on Computational Intelligence (UKCI 2023), September 6–8, 2023, Birmingham, UK* (Springer, Cham, 2024), 407).

123 Avner Levin, 'Privacy by Design by Regulation: The Case Study of Ontario' (2018) 4 *CAN. J. COMP. & CONTEMP. L.* 115.

124 Mireille Hildebrandt, *Law for Computer Scientists and Other Folk* (Oxford University Press, Oxford, 2020), 155.

125 Marco Almada, Juliano Maranhao and Giovanni Sartor, 'Art. 25. Data protection by design and by default', in Indra Spiecker gen. Döhmann, Vagelis Papakonstantinou, Gerrit Hornung and Paul de Hert (eds), *General data protection regulation : article-by-article commentary* (Nomos, Beck, Hart Publishing, 2023), 581-582.

126 European Union Agency for Cybersecurity, 'Privacy and Data Protection by Design' (2015) https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design accessed 23 July 2025, 11.

127 Christopher Kuner, Lee A Bygrave, and Christopher Docksey, 'Background and Evolution of the EU General Data Protection Regulation (GDPR)', in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press, Oxford, 2020), 26. Article 25 of the GDPR is a new provision under EU law; the 1995 Data Protection Directive, which preceded the GDPR, did not specifically address this concept, although it briefly mentions the idea in Recital 46 (Christina Michelakaki and Sebastião Barros Vale, 'Unlocking Data Protection by Design & By Default: Lessons from the Enforcement of Article 25 GDPR' (Future of Privacy Forum 2023) https://fpf.org/wp-content/uploads/2023/05/FPF-Article-25-GDPR-A4-FINAL-Digital.pdf accessed 25 July 2025).

128 "A broad vision of Article 25(1) transforms it into a kind of catch-all provision, scooping up anything omitted from other, more specific sections of the GDPR" (Ari Ezra Waldman, 'Data Protection by Design? A Critique of Article 25 of the GDPR' (2020) 53 *Cornell International Law Journal* 1, 147-167, 164).

129 Mireille Hildebrandt and Laura Tielemans, 'Data Protection

directly into the information and communication technology infrastructure that supports data-driven environments.[130] In this line, DPbD underscores the importance of recognising the significant and inherent regulatory potential embedded within the architecture of information systems.[131]

84 This concept typically involves minimising data use while enhancing security, restricting access, and limiting data distribution.[132] Framed in this manner, the duties of data protection by design require the systematic application of fundamental principles such as transparency, lawfulness, fairness, purpose limitation, data minimisation, accuracy, storage limitation, integrity, confidentiality, and accountability, as set forth in Article 5 and Recital 39 GDPR.[133]

85 Data Protection Impact Assessments (DPIAs), mandated by Article 35 GDPR, reflect the proactive spirit of DPbD. They systematically identify and manage risks, implementing safeguards before harm occurs. DPIAs offer a structured process to evaluate the necessity, proportionality, and potential risks of data processing,[134] supporting compliance,[135] and providing a clear framework for managing high-risk activities in accordance with GDPR requirements.

II. Scaling Data Protection by Design in the Metaverse: Key Challenges

86 In the Metaverse, implementing DPbD is not a procedural formality; it is a contested issue where technical, legal, and governance dimensions converge. Metaverse data fabrics strain governance structures, expose weak points in infrastructure, and erode user control. The velocity and granularity of these virtual worlds outpace traditional safeguards, demanding a shift to a holistic framework. This transformation calls for more than patchwork fixes, requiring an ecosystem-wide strategy. Against this complex backdrop, three central challenges emerge as decisive factors shaping the path forward:

87 (1) Gatekeepers and Platform Authority. The operators of metaverse platforms occupy a position of unparalleled control. They do not simply host virtual environments; they architect the rules, set the technical boundaries, and determine the very conditions under which interaction and data flows occur. Their role extends to shaping every stage of data handling: from limiting collection to what is strictly necessary, to embedding advanced safeguards such as state-of-the-art encryption and privacy-preserving architectures, including blockchain-based solutions.[136] In theory, such control could enable a rigorous application of DPbD, integrating privacy safeguards into every layer of the system from inception.

88 In practice, however, these gatekeepers often operate under economic incentives that conflict with such goals. The difficulty lies not merely in technical implementation but in a structural dilemma: the very actors expected to integrate data protection safeguards are also those with the strongest interest in maximising data extraction. Their business models thrive on pervasive collection, behavioural profiling, and the monetisation of user engagement, practices that directly undermine data minimisation principles. Proprietary architectures and closed ecosystems entrench these dynamics,[137] consolidating control over design choices. When flaws arise, this centralised control acts as a force multiplier, allowing shortcomings to cascade across vast portions and data controllers of the Metaverse,

by Design and Technology Neutral Law' (2013) 29 *Computer Law & Security Review* 5, 509-521, 516.

130 Mireille Hildebrandt, 'Code Driven Law. Scaling the Past and Freezing the Future', in Simon Deakin and Christopher Markou (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing, 2020), 67-84, 82.

131 Lee A. Bygrave, 'Article 25 Data protection by design and by default', in Christopher Kuner, and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press, Oxford, 2020), 573.

132 Kostina Prifti and others, 'Regulation by Design: Features, Practices, Limitations, and Governance Implications' (2024) 34 Minds & Machines 2, 13.

133 European Data Protection Board, 'Guidelines 4/2019 on Article 25 Data Protection by Design and by Default' (20 October 2020) Version 2.0 https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf accessed 22 July 2025.

134 De Cicco et al. (n 91) 94.

135 Article 29 Data Protection Working Party, Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679, WP 248 rev 01, adopted 4 April 2017 https://ec.europa.eu/newsroom/article29/items/611236 accessed 8 August 2025, 4.

136 Artzt and Weingarden (n 97) 7767.

137 Carlos Cantú, Cecilia Franco and Jon Frost, The economic implications of services in the metaverse (BIS Papers No 144, Bank for International Settlements February 2024) 21 https://www.bis.org/publ/bppdf/bispap144.pdf accessed 8 August 2025.

becoming deeply ingrained in its infrastructure and making remediation far more complex.

89 Additionally, dominant platforms may employ lock-in strategies to establish proprietary ecosystems that restrict users' ability to transfer data, assets, or identities across services. Several competing metaverse platforms appear increasingly monopolistic,[138] which may further reinforce these barriers by imposing closed standards and limiting open collaboration. This concentration of power risks consolidating control over data flows and complicating efforts to achieve comprehensive data protection.

90 (2) The Decentralisation Paradox. The shift to decentralised architectures, often heralded as a pathway to greater user control and resilience, introduces profound complexities for DPbD. Unlike centralised systems, decentralisation disperses control across a multitude of actors, devices, and services. This fragmentation complicates the comprehensive implementation of consistent safeguards, undermining the necessary holistic approach to privacy. As a result, it fosters a patchwork of fragmented solutions, each addressing isolated aspects, which perpetuates disintegration and weakens the overall effectiveness of data protections.

91 The challenge lies in reconciling the collaborative, interoperable nature of federated networks with the stringent demands of data privacy preservation. Every node in the network, whether a metaverse service provider, device manufacturer, or software developer acting as a data controller, must implement protection measures that adhere to shared standards. However, the absence of a centralised authority complicates coordination and dilutes accountability. The dispersion of responsibility creates vulnerabilities in privacy protections, eroding trust and exposing users to heightened risks. Moreover, it is important to recognise that many technology developers who contribute to the ecosystem but do not qualify as data controllers remain outside the direct scope of DPbD obligations,[139] leaving additional blind spots in data protection.

92 (3) The Regulatory Maze in the Virtual World. The Metaverse's borderless nature clashes head-on with a patchwork of privacy regimes. While this digital frontier transcends traditional jurisdictional boundaries, metaverse stakeholders and data controllers remain subject to diverse and often conflicting legal frameworks tied to users' locations. As a result, metaverse developers must tailor their services to comply with the specific laws and regulations applicable in each country or region.[140] This dynamic exposes a fundamental tension: the Metaverse aspires to offer seamless, global interaction, whereas it remains constrained by the territorial reach of national laws.

93 Multiple privacy laws may simultaneously apply to the same data or individual,[141] imposing varying obligations regarding transparency, lawfulness, user rights, breach notifications, and cross-border data transfers. This multiplicity makes universal terms and conditions unlikely,[142] particularly when virtual worlds are designed primarily to comply with the laws of their gatekeepers or main operators' home countries. The underlying infrastructure and devices often fail to accommodate the diverse regulatory demands of all jurisdictions where these environments are accessed.

94 This mosaic turns DPbD into a daunting puzzle. Consider the United States, where privacy by design remains more a whisper than a mandate, though rising calls for risk assessments signal a subtle shift. Whether these assessments will fully capture the comprehensive spirit of DPbD is yet to be seen.[143] Meanwhile, regulatory bodies like the Federal Trade Commission hint at a growing appetite for integrating privacy principles upstream.[144]

III. Data Protection by Design at the System Level

95 Scaling DPbD in data-saturated ecosystems defies simple solutions. The sheer density and interconnectedness of data streams across devices, platforms, and services expose the limits of isolated technical patches. This necessitates an integrated, ecosystem-level overhaul, one that fundamentally shapes software design patterns, infrastructure

138 Nick Munn and Dan Weijers, 'The Real Ethical Problem with Metaverses' (2023) 5 *Frontiers in Human Dynamics* 1226848, 3.

139 Hildebrandt and Tielemans (n 129) 519.

140 Chuan Chen and others, 'Privacy computing meets metaverse: Necessity, taxonomy and challenges' (2024) 158 *Ad Hoc Networks*, 11.

141 Artzt and Weingarden (n 97).

142 Norton Rose Fulbright, 'The Metaverse: The evolution of a universal digital platform' (www.nortonrosefulbright.com, November 2022) https://www.nortonrosefulbright.com/en-nl/knowledge/publications/5cd471a1/the-metaverse-the-evolution-of-a-universal-digital-platform accessed 8 August 2025.

143 Artzt and Weingarden (n 97) 7767.

144 Richmond Y. Wong and Deirdre K. Mulligan, 'Bringing Design to the Privacy Table: Broadening "Design" in "Privacy by Design" Through the Lens of HCI', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, 2019) Paper No 262, 1.

architecture, and the deployment of privacy-enhancing technologies.[145]

96 On a broader scale, modular and composable system architectures are crucial to enable privacy-preserving components to be updated or replaced as new threats emerge, facilitating ongoing adaptability and resilience. Integrating privacy into the underlying infrastructure through secure communication protocols and identity management systems ensures that protection is not an afterthought but a core design principle.

97 At a granular level, practical measures include empowering users with meaningful controls to limit data exposure, even when this might affect certain facets of user experience.[146] Features such as automatic blurring of bystanders, selective data obfuscation,[147] and real-time consent management help reduce inadvertent exposure within immersive environments. Furthermore, fine-grained access controls, data minimisation protocols, and encryption techniques for data at rest and in transit serve as key tools to safeguard sensitive information.

98 Equally critical are governance mechanisms that extend beyond technical solutions. Continuous and iterative DPIAs integrated early and throughout the development lifecycle are essential to identify, evaluate, and mitigate risks as these complex systems evolve. Transparent accountability frameworks must clearly allocate responsibilities across a wide range of stakeholders including platform operators, service providers, device manufacturers, and third-party developers. Without such clarity, the distributed and interconnected nature of the Metaverse risks creating regulatory blind spots where protections falter. Governance in this space must grapple with the complexity of overlapping jurisdictions, diffuse authority, and competing interests, ensuring that fragmentation does not translate into loopholes that erode privacy and user trust.

99 However, a critical question persists: can immersive architectures undergo fundamental redesigns at this stage? Keeping in mind the technological momentum and current maturity of these systems, some room for adjustment still exists. Although still in its developmental phase, the Metaverse is expected to evolve rapidly,[148] characterised by speculative features, uneven implementation, and the absence of a fully established business model.[149] At the same time, rapid technological advances continue to outpace regulatory efforts, making data protection a moving target rather than a fixed guarantee. In this shifting landscape, it is unclear whether integrating DPbD in virtual spaces remains more an aspirational ideal than an operational reality, transcending rhetoric to deliver meaningful, lasting safeguards.

100 Centralised metaverse platforms could provide a conducive setting for implementing DPbD duties. Their cohesive structures may allow a single actor to integrate privacy considerations consistently, coordinate data management, and deploy protective technologies across the platform. Such oversight can make monitoring and auditing more straightforward, potentially enabling faster detection and resolution of data protection issues. This apparent advantage, however, comes with a trade-off: concentrating control generates points of failure and structural vulnerabilities that require careful management.

101 Moreover, while centralised architectures can support more integrated privacy measures at the level of structures and processes, they also risk entrenching or exacerbating data protection challenges that remain unaddressed in the earliest design stages. Absent deliberate foresight, such systems risk reproducing or magnifying data protection gaps, underscoring that structural coherence alone cannot secure genuine compliance with the DPbD principle.

102 A clear example is Horizon Worlds, a Meta's virtual reality social platform.[150] It allows users to create, explore, and interact in three-dimensional virtual spaces through avatars.[151] Within the platform, users can build their own worlds, design games, and share experiences, engaging with others in real time. As a centralised system, control over servers, governance, rules, and most data processing remains with Meta, limiting the degree of influence users have over how the platform operates.[152]

103 The platform has attracted attention in debates surrounding privacy, data protection, and interoperability, particularly due to its extensive collection of behavioural and spatial data. These practices have been identified as posing potential

145 European Union Agency for Cybersecurity (n 126).

146 Olaizola Rosenblat (n 3) 18.

147 Ibid.

148 Arismendy Mengual (n 13).

149 Wang et al. (n 65) 14673.

150 Meta Horizon (Web Page) https://horizon.meta.com/?locale=en_US accessed 20 January 2026. Meta Horizon <https://horizon.meta.com/?locale=en_US> accessed 20 January 2026.

151 Kethryn Marie Protz, *Walk on the Metaverse Side: Data Protection in the Virtual Worlds* (Master of Science in Law, Digital Innovation and Sustainability, LUISS Guido Carli University 2023) page 11 https://tesi.luiss.it/37451/1/630803_PROTZ_KETHRYN%20MARIE.pdf accessed 20 January 2026.

152 Shymaa Arafat, 'Open vs Closed Metaverses: Decentraland & Horizon Worlds as Featured Examples' *Medium* (9 April 2025) < https://medium.com/%40shymaa.arafat/open-vs-closed-metaverses-decentraland-horizon-worlds-as-a-featured-examples-2501923286b6> accessed 20 January 2026.

risks to users' rights and freedoms, given the nature, context, and purpose of the data processing.[153] Horizon Worlds actively records user behaviour, including audio logs, as part of its data retention policy,[154] and it has the capacity to monitor users invisibly in real time to detect potential rule violations.[155]

104 Concerns have also been raised regarding the platform's priorities, emphasising user growth, while safety and privacy considerations (such as protecting children) were managed primarily as potential liabilities rather than core design principles.[156] The data collected by Meta could be leveraged to enhance its metaverse offerings for advertisers and business partners, raising additional questions about the commercialisation of user information.[157] A 2025 complaint to the U.S. Federal Trade Commission alleges that Meta allowed children under 13 to access Horizon Worlds, collected their personal data without parental consent, and thereby exposed them to potential harm under the Children's Online Privacy Protection Act (COPPA).[158]

G. Conclusion

105 The Metaverse is not a single platform or isolated enclave. It is an interconnected ecosystem of applications, services, and virtual environments. It includes both centralised and decentralised architectures, allowing users to interact with one another and with data-intensive technologies in real time. It does not exist as a separate digital realm or purely fictional space. Instead, it represents an evolution of the Web, building on established digital practices while expanding interactions and data exchanges in ways that intensify existing governance challenges.

106 The operational reality of the Metaverse depends on extensive and persistent information flows, collecting and generating massive volumes of data. Its infrastructural dynamics and processing logics make it more difficult to apply fundamental data protection principles, such as purpose limitation and data minimisation. Fragmented regulations and inconsistent standards further exacerbate these risks.

107 Data Protection by Design provides a clear framework for addressing these challenges. Integrating privacy considerations into every stage of system development transforms personal data protection from an optional feature into a structural safeguard. This approach (often described as "techno-regulation" or "legal by design") treats technology not merely as a tool but as a force that shapes legal and ethical practices.[159] It promotes proactive risk management, minimal data collection, and protective measures throughout the lifecycle of a system,[160] demonstrating the regulatory potential embedded within digital infrastructures. [161]

108 Centralised metaverse platforms can make this approach easier to implement because a single authority or small governing group can enforce privacy rules consistently, coordinate data handling, deploy privacy-enhancing technologies, and monitor compliance across the platform. However, concentrated control carries inherent risks. Single points of failure can amplify threats to fundamental rights, and the sheer scale and complexity of the Metaverse often exceed the capacity of any single authority to maintain comprehensive protection. Structural redesigns are rarely feasible amid

153 Protz (n 151) 43.

154 Aryabrata Basu, Mohammad Jahed Murad Sunny and Jayasri Sai Nikitha Guthula, 'Privacy Concerns from Variances in Spatial Navigability in VR' (arXiv:2302.02525, preprint, 6 February 2023) https://arxiv.org/abs/2302.02525 accessed 21 January 2026.

155 Ben Lang, 'Facebook Horizon Can Invisibly Observe Users to Spot Rule Violations' Road to VR (28 August 2020) https://www.roadtovr.com/facebook horizon privacy monitoring moderation/ accessed 22 January 2026.

156 https://www.businessinsider.com/meta-horizon-worlds-child-safety-whistleblower-ftc-fairplay-2025-4See Jyoti Mann and Pranav Dixit, 'Meta Whistleblower Claims the Company Put Profit over the Safety of Children in Horizon Worlds' *Business Insider* (10 April 2025) <https://www.businessinsider.com/meta-horizon-worlds-child-safety-whistleblower-ftc-fairplay-2025-4> accessed 21 January 2026.

157 Brooks Canavesi, 'The Foundation of the Metaverse: Centralization Versus Decentralization' ATD Blog (2 February 2022) https://www.td.org/content/atd-blog/the-foundation-of-the-metaverse-centralization-versus-decentralization accessed 24 January 2026.

158 Fairplay, Complaint against Meta filed at FTC alleging widespread Horizon Worlds child privacy violations (Press release, 10 April 2025) https://fairplayforkids.org/complaint-against-meta-filed-at-ftc-alleging-widespread-horizon-worlds-child-privacy-violations/ accessed 24 January 2026.

159 Carlotta Rigotti and Gianclaudio Malgieri, *Sexual Violence and Harassment in the Metaverse: A New Manifestation of Gender-Based Harms* (The Alliance for Universal Digital Rights (AUDRi), Equality Now and The International Observatory on Vulnerable People in Data Protection (Vulnera) 2024) <https://audri.org/wp-content/uploads/2024/04/EN-AUDRi-Sexual-violence-and-harassment-in-the-metaverse-03.pdf> accessed 6 August 2025.

160 Agencia Española de Protección de Datos, 'Guia de Privacidad desde el Diseño' (2019) <https://www.aepd.es/guias/guia-privacidad-desde-diseno.pdf> accessed 5 August 2025, 6.

161 Bygrave (n 131).

rapid technological development and commercial pressures. This tension underscores a persistent trade-off: centralised systems enable more effective oversight but concentrate power, whereas decentralised systems disperse risk but complicate enforcement.

109 Although the temptation exists to craft a specialised legal regime for the Metaverse, such a measure appears unnecessary. Existing regulatory instruments (including the DMA and DSA) already provide the tools to address its key challenges. Platforms can integrate privacy protections directly into their architectures, operationalising DPbD to make privacy a structural feature rather than an afterthought. At the same time, these frameworks empower regulators to monitor compliance, enforce obligations, and maintain oversight across platforms. Merging technology/application-level protections with regulatory oversight, this approach addresses ongoing privacy and governance challenges from the broader Web, while respecting the limits of both centralised and decentralised control.

110 The path forward in the Metaverse lies not in new rules, but in how we apply and shape the ones we already have. Metaverse platforms and applications must make privacy a structural component, while regulators must use established instruments to enforce compliance and guide development. Success will depend on a dynamic interplay: technological design that anticipates risk, and technology-neutral legal frameworks designed to adapt to innovation.

The Mythical Code-Authorship

by **Ewa Laskowska-Litak ***

Abstract: Modern copyright law emerged during the transition from the 18th to the 19th centuries, centred on the figure of the author as the originator of unique, creative works. Authors were seen as deserving legal protection, both because of their creative labour and their vulnerable position in publishing contracts. Over time, this framework has expanded to include various creative outputs, including musical works, films, and computer programs. However, the incorporation of software into copyright law occurred without sufficient understanding of its development process, leading to conceptual tensions.

This paper addresses two interrelated aims. First, it examines, from a primarily European Union perspective supplemented by illustrative United Kingdom authorities, whether changing perceptions of software—particularly its collaborative, iterative and functional nature—require a recalibration of copyright analysis. Second, it challenges two persistent misconceptions: (1) that software authorship is straightforward and legally unproblematic, and (2) that most source code qualifies as protectable under copyright. These assumptions, foundational to existing copyright regimes, fail to reflect the realities of programming and contribute to a growing disconnect between legal theory and software practice. The paper is divided into two major parts: the first critiques prevailing copyright assumptions using sociological and psychological insights into authorship and creativity, revealing the mismatch between the solitary-author model and the inherently collective, structured nature of programming. The second part proposes a revised framework that accounts for the collaborative and social character of code creation. Here, the main observation is that authorship is understood more technically, through programmatic attribution, rather than through traditional legal concepts.

In conclusion, the paper advocates for a more nuanced and interdisciplinary approach to copyright law—one that acknowledges the complexities of software development and better aligns legal doctrine with technical and social realities. This shift is crucial to addressing the growing misfit between copyright's foundational assumptions and the nature of modern software production.

Keywords: Authorship; code authorship; copyright law; collaborative authorship; software development

© 2026 Ewa Laskowska-Litak

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: Ewa Laskowska-Litak, The Mythical Code-Authorship, 17 (2026) JIPITEC 161 para 1.

A. Introduction

1 The genesis of copyright thought can be traced back to the transition from the 18th to the 19th centuries, a period that witnessed the establishment of foundational principles that persist to the present day: that it is the author, as the central figure of copyright law, who deserves protection because of their creative effort; that they are the weaker party to the contract and the legislator should protect them to a certain extent from exploitation by publishers; that the object of protection should be an original and unique work, carrying the genius of the creator, distinctive from others.[1] In the course of time, only

* Ewa Laskowska-Litak, PhD habilit., LL.M. (Heidelberg); Faculty of Law and Administration, Future Law Lab, Jagiellonian University, Kraków, Poland
e-mail: ewa.laskowska@uj.edu.pl

Acknowledgments: This work was supported by the National Science Centre in Poland (Narodowe Centrum Nauki) under grant no. 2023/49/B/HS5/00772.

1 Martha Woodmansee, 'The Genius and the Copyright: Economic and Legal Conditions of the Emergence of the "Author"' (1984) 17 *Eighteenth-Century Studies* 425.

further protection instruments growing out of these foundations have been added. This evolution has encompassed the establishment of normative protection over author's personal interests, the protection of musical, and subsequently, also of cinematographic works, computer programs, and, to a certain extent, the protection of artistic performers. Nevertheless, the decision to extend the scope of copyright protection to computer programs, without conducting a thorough examination of the software development process, may have inadvertently changed the perspective of copyright law as such. The purpose of this paper is twofold: firstly, to challenge a hypothesis that a shift in the manner in which software is perceived may facilitate a concomitant change in perspective with regard to copyright law. And secondly, to address two misconceptions relating to software protection in copyright law:

(1) that software authorship constitutes a straightforward and uncomplicated case; and

(2) that the majority of source code is subject to copyright protection.

2 Both are foundational assumptions of all copyright instruments. Both also, in my view, not only fail to capture the specificity of computer programs, but also lead to a significant and progressive disconnect between programming pragmatics and legal evaluation.

3 In other words, the main aim of this research is to dispel the prevailing myths surrounding the copyright narrative concerning software, with a particular focus on these two gravitational points: authorship and copyright subject matter. In the first, mythological part, we will confront these assumptions in turn by addressing the issues of authorship and the protection of computer program code. Utilising examples drawn from sociological and psychological research, this paper aims to highlight the inaccuracies in the understanding of authorship and the dissonance that arises when confronted with the reality of programming and with doctrinal analysis of authorship, joint authorship and contractual allocation of rights. The second (demythologising) part will be devoted to an in-depth analysis of two elements that copyright law does not take into account, but which it should systemically include in this assessment: the collaborative element of code authorship and the sociality of programming. In the latter case, the focus is primarily on the technical understanding of authorship (as programmatic attribution, not about authorship in the legal sense). The primary postulate emphasises the necessity for enhanced caution in the formulation of legal authorship assessments. Additionally, it underscores the imperative for a more interdisciplinary and community-based approach in copyright law, particularly in the context of assessing collaborative authorship. The concluding part then develops a more concrete reform agenda directed at legislation, adjudication and contracting practice.

4 This article proceeds primarily from the legal framework of the European Union, especially the Software Directive and the CJEU case law on originality and the scope of protection for computer programs,[2] while using United Kingdom doctrine on joint authorship as a comparative illustration where it sharpens the analysis. The point is not that one jurisdiction can simply be universalised, but that the EU/UK materials make especially visible the tension between copyright's inherited assumptions and contemporary software production.

B. Myth No. 1: Software Authorship Does Not Require Special Treatment

5 The status of author is generally attributed to the individual who created the work. The veracity of this assertion remains unassailable, irrespective of the legal system under consideration, be it common law[3] or continental law.[4] In principle, therefore, one can assert that the creator of a work can be considered the author, irrespective of geographical location, provided that the country in question has adopted the Berne Convention.[5] Although the dogmatics of copyright law are, in general, very conservative on this aspect, the legal scholarship completes the system. The prevailing tendency is to conceptualise authorship as a romantic phenomenon[6] (or not

2 Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) [2009] OJ L111/16.

3 Jane C Ginsburg, 'The Concept of Authorship in Comparative Copyright Law' (2003) 52 DePaul L Rev 1063.

4 Lionel Bently, 'Authorship of Popular Music in UK Copyright Law' (2009) 12 *Information, Communication & Society* 179; Lior Zemer, *The Idea of Authorship in Copyright* (Ashgate 2007); Mireille van Eechoud (ed), *The Work of Authorship* (Amsterdam University Press 2014).

5 Sam Ricketson, *The Berne Convention for the Protection of Literary and Artistic Works: 1886–1986* (Centre for Commercial Law Studies, Queen Mary College 1987); Berne Convention for the Protection of Literary and Artistic Works (adopted 9 September 1886, as revised at Paris 24 July 1971 and amended 28 September 1979) 1161 UNTS 3.

6 Megan Richardson and Julian Thomas, 'Rethinking "Romantic" Authorship' in *Fashioning Intellectual Property: Exhibition, Advertising and the Press, 1789–1918* (Cambridge University Press 2012) 103; Isabella Alexander, 'The Genius and the Labourer: Authorship in Eighteenth and

at all[7]). Another well-documented assumption on authorship is that the grant of copyright to the creator is of paramount importance, as it serves to safeguard the creator's work and creative genius,[8] or that the author's character requires special protection because, as an independent yet inexperienced person in the trade, they are vulnerable to exploitation.[9]

6 Yet, even at this basic level, software forces a more careful doctrinal analysis than the orthodox narrative often suggests. In the EU framework, computer programs are protected as literary works, but only if they are original in the sense of being the author's own intellectual creation;[10] the legal question is therefore never merely who contributed, but which contributions count as authorial expression and under what doctrinal test. At the same time, the law of joint authorship, employee creation and contractual allocation of rights shows that copyright has long had to mediate between creativity, collaboration and market organisation.

7 Current copyright frameworks do address collaborative creation, though unevenly. A useful illustration is the United Kingdom notion of a "work of joint authorship", which requires collaboration and non-distinct contributions. In Kogan v Martin [2019] EWCA Civ 1645, the Court of Appeal restated the inquiry in a way that is particularly relevant for software disputes: the court must ask whether the putative co-author made an authorial contribution that formed part of the creation of the work, whether that contribution is quantitatively and qualitatively sufficient, and whether it is part of a collaborative process rather than merely advisory criticism or editing. The attraction of this approach is its flexibility; its weakness, for software, is that it still presupposes a relatively intelligible chain between contribution and expressive outcome. In iterative software projects that chain is often dispersed across tickets, pull requests, reviews, architecture choices and dependency management.

8 The situation becomes especially difficult where there is more than one person working on the work and where contractual or employment structures detach legal ownership from factual creation. Work-for-hire and employee-creation rules differ across jurisdictions, but their common function is to facilitate exploitation by reallocating rights away from the natural person who physically created the work.[11] Co-creation raises a different problem: not who will own the rights after creation, but which participants count as authors at all. That question has long been difficult in copyright law—in music, photography, film and social media alike—and software does not create the problem from nothing. What software does is intensify it by making cumulative, interdependent and partially invisible contributions the rule rather than the exception.

9 Whilst the problems presented by social media, music and other forms of collaborative authorship to the existing copyright paradigm have been well-discussed and researched for years,[12] it is contemporary software pragmatics, particularly those based on agile frameworks, that represent a particularly salient threat.[13] For instance, one can agree with the assertion that a co-author is a person who exerts the most significant influence on the final form of the work.[14] However, the extent to which

Nineteenth-Century Copyright Law' in Lionel Bently, Jennifer Davis and Jane C Ginsburg (eds), *Copyright and Piracy: An Interdisciplinary Critique* (Cambridge University Press 2010) 300; Margaret Chon, 'The Romantic Collective Author' (2012) 14 Vand J Ent & Tech L 829; Mark A Lemley, 'Romantic Authorship and the Rhetoric of Property' (1997) 75 Tex L Rev 873.

7 cf Michael J Madison, 'Beyond Creativity: Copyright as Knowledge Law' (2010) 12 *Vanderbilt Journal of Entertainment and Technology Law* 817; Annemarie Bridy, 'Coding Creativity: Copyright and the Artificially Intelligent Author' (2012) 2012 Stan Tech L Rev 5; Olufunmilayo B Arewa, 'From J.C. Bach to Hip Hop: Musical Borrowing, Copyright and Cultural Context' (2006) 84 NC L Rev 547; Elena Cooper, 'Copyright and Mass Social Authorship: A Case Study of the Making of the Oxford English Dictionary' (2015) 24 *Social & Legal Studies* 509.

8 Ole-Andreas Rognstad, *Property Aspects of Intellectual Property* (Cambridge University Press 2018); Carys J Craig, 'Locke, Labour and Limiting the Author's Right: A Warning against a Lockean Approach to Copyright Law' (2002) 28 Queen's LJ 1; Lior Zemer, 'Three Arguments on Locke, Authorship, Communication and Solitude' (2015) 7 *The WIPO Journal* 40.

9 Diane Leenheer Zimmerman, 'Copyrights as Incentives: Did We Just Imagine That?' (2011) 12 *Theoretical Inquiries in Law* 29; James Griffin, 'The Need for a New Paradigm in IP Law: A Focus on Authorship' (2005) 14 *Information & Communications Technology Law* 267; Ronan Deazley and Bartolomeo Meletti, 'Copying, Creativity and Copyright' (CREATe Working Paper, University of Glasgow, February 2016).

10 Directive 2009/24/EC (n 2) art 1(3); Case C-393/09 *Bezpečnostní softwarová asociace – Svaz softwarové ochrany v Ministerstvo kultury* ECLI:EU:C:2010:816.

11 Catherine L Fisk, 'Authors at Work: The Origins of the Work-for-Hire Doctrine' (2003) 15 *Yale Journal of Law & the Humanities* 1.

12 Peter Jaszi, 'Toward a Theory of Copyright: The Metamorphoses of "Authorship"' (1991) 40 Duke LJ 455; Jason Borenstein and Adil E Shamoo, 'Rethinking Authorship in the Era of Collaborative Research' (2015) 22 *Accountability in Research* 267.

13 Philippe Kruchten, 'Contextualizing Agile Software Development' (2013) 25 *Journal of Software: Evolution and Process* 351; Alistair Cockburn, *Agile Software Development: The Cooperative Game* (2nd edn, Addison-Wesley 2006).

14 Peter Jaszi, 'On the Author Effect: Contemporary Copyright

this (causal?) influence can justify the co-authorship status is still not yet determined. Since we have not so far extended co-author status to the producer of a film, can a scrum master be a co-author of software? After all, their role is to facilitate the success of the team, to ensure effective communication within the team, or to remove impediments that hinder the team's progress. Their role, although not directly related to the writing of code, facilitates the same act:

10 the Scrum Master is a servant-leader for the Scrum Team. The Scrum Master helps those outside the Scrum Team understand which of their interactions with the Scrum Team are helpful and which aren't. The Scrum Master helps everyone change these interactions to maximize the value created by the Scrum Team[15].

11 The aforementioned concerns are often managed in practice through contracts rather than through fine-grained judicial inquiry into authorship. In software projects, common contractual devices include employment clauses assigning economic rights, commissioned-development agreements, contributor licence agreements, open-source licence terms and repository governance rules. These instruments reduce transactional uncertainty and enable exploitation, but they do not eliminate the underlying doctrinal question of authorship, because author status is generally a matter of fact and law rather than pure contract.[16] Nor do they fully solve evidentiary difficulties. IT contracts are typically multi-layered and multi-party, regulating scope, methodology, security, risk allocation, change management, acceptance criteria and delivery structures.[17] In that environment, the practical success of contracting can mask the conceptual instability of authorship doctrine: the project proceeds, rights are allocated for business purposes, yet the legal map of creative contribution remains obscure unless a dispute forces the issue.

12 Despite the fact that the role of figures such as scrum masters and product owners does not directly pertain to the creation of software code, it is challenging to refute their, at times even direct, influence on the eventual functionality and appearance of the software. In their study, Shastri et al.[18] examined the impact of the presence of scrum masters in IT teams, highlighting a positive association between specific agile practices and the effectiveness of the scrum master (definition of done, iteration (sprint) planning, project velocity measurement, and short iterations (sprints))..[19] In particular, the ability to mediate (negotiate) between client expectations and the team's capabilities seems to have a crucial role in the context of copyright, as it always translates into the final shape of the work—a requirement for the copyright protection to subsist. The coordination of collaborative efforts between teams from diverse organisational domains, the safeguarding of teams against external interferences, and the rebuff of scope creep can also be of equal importance.[20]

13 Nevertheless, even if the causal link between the role of the scrum master and software production is considered to be too speculative, the exclusive focus on software engineers alone generates further uncertainty. The GitHub platform is recognised as one of the most prevalent tools for software development.[21] The platform itself offers a range of options for storing computer program code in dedicated repositories (generally either commercial-private or public). Research conducted on the methods of software development within this platform has yielded several noteworthy insights in this regard. Firstly, a small proportion of public repositories (i.e. those created by a single individual for personal, scientific or development purposes) are single-author: only one quarter of all developers

and Collective Creativity' (1992) 10 Cardozo Arts & Ent LJ 293.

15 Ken Schwaber and Jeff Sutherland, *The Scrum Guide: The Definitive Guide to Scrum — The Rules of the Game* (Scrum.org and Scrum Inc 2016).

16 van Eechoud, *The Work of Authorship* (n 4); Bently (n 4); Arewa (n 7).

17 Michael J Gallivan and Wonseok Oh, 'Analyzing IT Outsourcing Relationships as Alliances among Multiple Clients and Vendors' (*Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, Maui*, IEEE Computer Society 1999); L Guan, 'Managing Risk in IT Contracts' (2012) 32 *Government News* 40; John Warchus, 'Avoiding Litigation: Drafting Successful IT Contracts' (1999) 15 *Computer Law & Security Report* 117.

18 Yogeshwar Shastri, Rashina Hoda and Robert Amor, 'Spearheading Agile: The Role of the Scrum Master in Agile Projects' (2021) 26 *Empirical Software Engineering* art 3.

19 ibid.

20 Shastri, Hoda and Amor (n 18); Maja Due Kadenic, Konstantinos Koumaditis and Louis Junker-Jensen, 'Mastering Scrum with a Focus on Team Maturity and Key Components of Scrum' (2023) 153 *Information and Software Technology* 107079; Hannes Salin, Felix Albrecht and John Skov, 'The Power of Scrum Mastery: An Analysis of Agile Team Performance and Scrum Master Influence' (*Proceedings of the 18th International Conference on Software Technologies, Rome*, SCITEPRESS 2023) 377.

21 Laura Dabbish and others, 'Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository' (*Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, Seattle*, ACM 2012) 1277; Antonio Lima, Luca Rossi and Mirco Musolesi, 'Coding Together at Scale: GitHub as a Collaborative Social Network' (*Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, AAAI Press 2014) 295.

within GitHub create their code independently.[22] And even when they do work together on a public project, research indicates that their participation is usually very cautious and rather small: limited to an average of 10 files.[23] In contrast, private (commercial) repositories are overwhelmingly multi-author and decidedly more extensive.[24] Furthermore, an examination of the relationship between the choice of programming language and the personnel complexity of the project suggests that the decision in this regard predefines collaborative authorship:[25] some programming languages require collaboration (such as JavaScript or Ruby), while others allow for limited communication between teams. Therefore, while for other creative fields, the decision to include another author in a project may be driven by creative invention, for some programming languages this becomes a key element of running a technology project.[26] Another study on public and private GitHub repositories suggests that it is unlikely to be possible to identify one person responsible for specific parts of the source code: at least several people are more or less engaged in every smaller or bigger part of the code.[27] The above studies very clearly challenge the copyright paradigms, based on the underlying assumption that authors tend to work independently on the work and co-authorship is generally the exception to this rule.

14 The concerns mentioned earlier may be alleviated through the application of contractual clauses and the mutual consent of the parties with regard to the collaborative nature of the creative process and its extent. Whilst there is a degree of consensus that author status is a matter of fact rather than contract,[28] it is suggested that some regulatory guidance could be helpful in the event of a legal dispute. However, the software development context does not facilitate this resolution either. Contracts in the IT sector are among the most multi-faceted, multi-level and hence multi-party.[29] They typically regulate not only the project's commercial object, but also development methodologies, security obligations, compliance requirements, data protection, allocation of implementation risks, testing and acceptance procedures, change-management mechanisms, maintenance expectations, and delivery schedules. Even so, no contract is capable of exhaustively capturing the full reality of software production. Software development is not a linear execution of pre-defined obligations, but an iterative and dynamic process shaped by continuous adjustment, unforeseen dependencies, emergent technical constraints, and evolving business expectations. As the project progresses, new functionalities are proposed, bugs reveal architectural weaknesses, integrations with external services reshape technical priorities, and security or regulatory requirements may require substantial redesign. In such an environment, not every relevant contribution can be anticipated ex ante, named in advance, or reduced to a stable contractual category. The legal document may define roles formally, but it cannot fully predict how influence over the final form of the software will in fact be exercised throughout the lifecycle of the project. Software development has evolved into a multifaceted process that demands a substantial investment of resources, particularly human capital. Modern software is rarely created within a single homogeneous team. Rather, it emerges from the interaction of developers, testers, UX designers, DevOps engineers, security specialists, data analysts, compliance officers, project managers, product owners, and, frequently, external vendors or client-side stakeholders. Their contributions operate on different levels, in different temporal phases, and according to different institutional logics. Some shape the literal code directly; others determine the architecture, priorities, feasibility, sequencing, or permissible scope of implementation. In practice, the final work is therefore often the result of distributed and cumulative acts of coordination, negotiation, review, acceptance, and revision. For that reason, the significance of particular individuals may exceed what is visible from the contractual allocation of

22 Lima, Rossi and Musolesi (n 21); Guilherme Avelino and others, 'Measuring and Analyzing Code Authorship in 1 + 118 Open Source Projects' (2019) 176 *Science of Computer Programming* 14.

23 Avelino and others (n 22).

24 Avelino and others (n 22); Dabbish and others (n 21); Fabian Stephany, Fabian Braesemann and Mark Graham, 'Coding Together – Coding Alone: The Role of Trust in Collaborative Programming' (2021) 24 *Information, Communication & Society* 1944.

25 Gabriel P Oliveira and others, 'How Do Developers Collaborate? Investigating GitHub Heterogeneous Networks' (2023) 31 *Software Quality Journal* 211; Natércia A Batista and others, 'Collaboration Strength Metrics and Analyses on GitHub' (*Proceedings of the International Conference on Web Intelligence, Leipzig*, ACM 2017) 170.

26 Batista and others (n 25); Dabbish and others (n 21).

27 Casey Casalnuovo and others, 'Developer Onboarding in GitHub: The Role of Prior Social Links and Language Experience' (*Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, Bergamo*, ACM 2015) 817; Pedro Henrique Pereira Braga and others, 'Not Just for Programmers: How GitHub Can Accelerate Collaborative and Reproducible Research in Ecology and Evolution' (2023) 14 *Methods in Ecology and Evolution* 1364; Ines Mergel, 'Open Collaboration in the Public Sector: The Case of Social Coding on GitHub' (2015) 32 *Government Information Quarterly* 464; Nikolas Zöller, Jonathan H Morgan and Tobias Schröder, 'A Topology of Groups: What GitHub Can Tell Us about Online Collaboration' (2020) 161 *Technological Forecasting and Social Change* art 120291.

28 van Eechoud, *The Work of Authorship* (n 4); Bently (n 4); Arewa (n 7).

29 Gallivan and Oh (n 17); Guan (n 17); Warchus (n 17).

duties alone. Persons, not merely clauses, influence the shape of the resulting software product. Accordingly, while contractual arrangements remain indispensable for reasons of legal certainty and risk allocation, they should not be assumed to resolve, in any definitive manner, the problem of authorship in software development. The complexity, variability and human embeddedness of software production mean that authorship-related questions cannot always be settled by reference to formal provisions alone. In many instances, the decisive factors will lie in actual collaborative practice: who formulated the decisive solution, who redirected the project at a critical moment, who resolved a bottleneck that altered the final implementation, or who exercised meaningful control over the program's structure and functionality. The software environment therefore demonstrates particularly clearly that legal categories of authorship may be underinclusive when confronted with a creative process that is fluid, collective, and only partially foreseeable. From this perspective, contractual regulation of authorship should be treated as necessary but inherently incomplete, since the practical significance of human relationships, situated decision-making and cross-functional cooperation cannot be fully translated into ex ante legal language.[30]

15 In conclusion, it seems that the question of authorship of a computer program is by no means a simple and straightforward matter. It is part of a very complex and multifaceted technological process involving various actors, to varying degrees and intensities, making the determination of who has rights, and to what, an important question. Despite the fact that the prevailing legal paradigm asserts that copyright is inherently attributed to the creator, this approach appears to be inadequate when confronted with the complexity, hybrid nature and multi-faceted character that are increasingly prevalent in the domain of modern computer programs and mobile applications.

C. Myth No. 2: The Majority of Source Code Is (and Should Be) Protected

16 Any claim that copyright overreaches in relation to software must begin with the concept of originality. Copyright does not protect all code, but only code that satisfies the applicable originality threshold. In the EU framework, that threshold is whether the relevant expression is the author's own intellectual creation; by contrast, elements dictated solely by technical function, constraints of the programming language, interoperability requirements or standard coding conventions fall outside protection. Framed this way, the argument is not that copyright law automatically protects all source code and therefore needs radical reduction. Rather, the problem is that software discourse often moves too quickly from "there is code" to "there is protected expression", without sufficiently testing originality, functionality and the narrow scope of protectable expression in computer programs.

17 This doctrinal baseline matters because the legal materials already contain resources for a more restrained analysis. The CJEU's originality jurisprudence, usually associated with *Infopaq*, insists on free and creative choices;[31] the software acquis adds that no criteria other than originality apply, but also that protection does not extend to ideas and principles underlying any element of a program.[32] Likewise, the case law on computer programs stresses that functionality, programming languages and data formats are not monopolised as such.[33] Properly understood, then, software copyright is neither absent nor all-encompassing: it protects expressive choices in code, but not the technical necessities, methods and standards through which software remains usable and interoperable.

18 Within the domain of copyright scholarship, the concept of 'kleine Münze' holds particular significance. It was proposed that this term should be used to denote types of works which, due to their 'insignificance', 'minimalist content' or 'utilitarian character', should not be considered as subjects of copyright protection.[34] Examples include, for example, the design of a simple kitchen tool (knife); a bus timetable; a monthly calendar; and, in modern times, also: emoji.[35] It is challenging to unequivocally identify the individual who first recognised the issue concerning the copyright protection of a computer program. It has been argued that it should not be subject to the provisions of copyright law, but rather those of patent law.[36] However, as Boyle has

30 Guan (n 17); Trif (2024).

31 Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* ECLI:EU:C:2009:465.

32 Directive 2009/24/EC (n 2) art 1(2).

33 Case C-406/10 *SAS Institute Inc v World Programming Ltd* ECLI:EU:C:2012:259.

34 Annette Kur, 'Die Alternativen zum Schutz durch das Urheberrecht in Deutschland' in Reto M Hilty and Christophe Geiger (eds), *Impulse für eine europäische Harmonisierung des Urheberrechts* (Springer 2007) 193; Marcel Bisges, *Die Kleine Münze im Urheberrecht: Analyse des ökonomischen Aspekts des Werkbegriffs* (Nomos 2014).

35 Eric Goldman and Gabriella Ziccarelli, 'Emojis and Intellectual Property Law' (Santa Clara University Legal Studies Research Paper, 2018).

36 James Boyle, *Shamans, Software, and Spleens: Law and the Construction of the Information Society* (Harvard University Press 1996).

noted, copyright protection is significantly more favourable for a number of reasons. These include its global nature, the limited formalities required to obtain protection, and the complexity and duration of the rights granted.[37] The benefit of copyright protection has ultimately led to a state in which we assume that each element of a computer program is treated in the same way as a literary work, i.e. it is protected primarily in its textual layer.[38] However, a closer inspection of this assertion reveals certain inaccuracies.

19 A single written line of source code does not necessarily translate into action or specific functionality for the entire system. Indeed, in the majority of cases, it can be considered a directive to invoke or transmit specific information to a designated location (e.g. when displaying a particular screen in a mobile application). In many instances it merely invokes a command, implements a standard library call, satisfies a syntactic requirement or reproduces a functional solution that is largely dictated by the chosen language or framework. For that reason, the originality inquiry in software must be conducted at an appropriate level of abstraction. Some protectable expression may lie in a larger sequence, in a distinctive arrangement, or in a specific implementation that reflects free and creative choices. But much code is better understood as constrained, interoperable and heavily shaped by function. That insight does not abolish copyright in software; it narrows it to those aspects that are genuinely expressive rather than technically inevitable.

20 Moreover, the formulation is predominantly contingent on the characteristics inherent to the selected programming language. In a manner analogous to the capacity to articulate the aesthetic appeal of weather through diverse linguistic expressions, a programming language possesses limited capacity to convey a single 'creative thought'.[39] Only the writing of a few (and sometimes dozens) lines of code translates into a greater meaning for the entire program, and on top of that, it still has to correctly interact (compile) with other elements usually written by other software engineers. By default, then, developing software is far more limited, in comparison to other creative activities, in terms of freedom of expression. And even setting aside the question of protecting the literal part of the code, for some time the copyright jurisprudence has recognised the so-called non-literal elements of computer code, understood mostly as structure, sequence and organisation, provided they were adequately distinguishable from other programs of a specific type.[40] Although this position has been revised slightly over time,[41] the fact that these non-literal blocks of source code may be the result of a need related to the compilation of the code, the implementation of a relevant API, or an accepted programming convention, is still not currently taken into account. It is possible to provide numerous examples of the misunderstanding of software in the copyright system, but these do not alter the dogma accepted in the literature as well as in the case law that the source code of a program is generally protected, even if it is only a line of such a code, especially if it contains 11 words.[42]

21 This romantic notion of the creative genius, uniqueness and originality of computer program code does not align with the reality that a significant proportion of code is merely copied as part of software development. Golubev et al.[43] conducted a survey of 23,378 projects on GitHub, each containing at least one line of Java code. The study exclusively focused on projects that had amassed a minimum of 50 stars, a metric that serves as an indication of user interest on GitHub. The study found evidence of at least 11.7% of exact clones, defined as copied code snippets from one project to another, implemented without any alterations.[44] Meanwhile, Baltes et al.[45] conducted

37 Boyle, *Shamans, Software, and Spleens* (n 36); James Boyle, 'The Search for an Author: Shakespeare and the Framers' (1988) 37 Am U L Rev 625; James Boyle, 'A Politics of Intellectual Property: Environmentalism for the Net?' (1997) 47 Duke LJ 87.

38 Jon Crowcroft, 'Copyright, Piracy and Software' in Lionel Bently, Jennifer Davis and Jane C Ginsburg (eds), *Copyright and Piracy: An Interdisciplinary Critique* (Cambridge University Press 2010) 209; K Tumbraegel and R de Villiers, 'Copyright Protection for the Non-Literal Elements of a Computer Program' (2004) 10 *Computer and Telecommunications Law Review* 34.

39 Mark Gabel and Zhendong Su, 'A Study of the Uniqueness of Source Code' (*Proceedings of the Eighteenth ACM SIGSOFT International Symposium on Foundations of Software Engineering, Santa Fe*, ACM 2010) 147.

40 *Whelan Associates Inc v Jaslow Dental Laboratory Inc* 797 F 2d 1222 (3d Cir 1986).

41 *Computer Associates International Inc v Altai Inc* 982 F 2d 693, 706–711 (2d Cir 1992); Joshua Bloch and Pamela Samuelson, 'Some Misconceptions about Software in the Copyright Literature' (*Proceedings of the 2022 Symposium on Computer Science and Law*, ACM 2022) 131.

42 Mireille van Eechoud, 'Along the Road to Uniformity – Diverse Readings of the Court of Justice Judgments on Copyright Work' (2012) 3 JIPITEC 60; Eleonora Rosati, 'Closed Subject-Matter Systems Are No Longer Compatible with EU Copyright' (2014) GRUR Int 1112.

43 Yaroslav Golubev and others, 'A Study of Potential Code Borrowing and License Violations in Java Projects on GitHub' (*Proceedings of the 17th International Conference on Mining Software Repositories, Seoul*, ACM 2020) 54.

44 ibid.

45 Sebastian Baltes, Richard Kiefer and Stephan Diehl, 'Attribution Required: Stack Overflow Code Snippets in

a study that analysed copied snippets from Stack Overflow, a forum where developers pose questions regarding specific issues in code development. These studies estimated that approximately 25% of Java projects contained exact clones. In a slightly earlier analysis, this figure was marginally elevated (27%), with the observation that the greater the popularity of a repository among developers, the higher the proportion of code duplicated from another project (7.33%). On average, within a single project, the proportion of clones from external sources is 3.22%.[46] Indicated in a study on cross-project clones that the proportion of clones varies between 10% and 30%.[47] Following the construction of tools capable of tracing clones across a number of widely utilised programming environments, and subsequent analysis of over 428 million files written in Java, C++, Python and JavaScript, a corpus containing 85 million unique files was identified by other researchers. This finding suggests that approximately 70% of the code on GitHub is derived from pre-existing files. It was also noted that JavaScript has the highest rate of file duplication.[48] The statistics demonstrate that, on average, 25–30% of the code in a project is likely to be a literal copy from external sources. In some repositories, this proportion may exceed half.[49] And while a certain amount of repetitiveness may be present as an element of style, artistic convention or motif, when it comes to programming that is protected without closer review by the courts, this becomes an example of a failure to adapt the law to reality, and consequently opens up the field to develop its own tools to resolve any dissonance. It also stands in contrast to the prevailing assumption that an author's individual creative invention is primarily protected because of the creative effort they have invested in creating an original work. According to Gharehyazie et al., the majority of clones stem from the same project. In the event that suitable code is absent from a project, developers will subsequently turn to code from projects in the same application domain, and finally from projects in different domains. An algorithmic solution within the same application is therefore sought first, leading to the question of whether subsequent code fragments within the same software or application become repeatable. The larger and more elaborate the project, logically the smaller the proportion of clones that occur, and the larger the copied fragment, the more additional modifications have been made to it relative to the original.[50] Therefore, it seems that, at least in the initial development phase, much of the code is neither original nor creative, but largely a duplication of existing structures from this or other projects. The authors of the study observe that, interestingly, in the majority of cases, the rationale behind clone copying is relatively prosaic. It is common practice among developers to replicate the complete source code of popular libraries in order to utilise their APIs; in the absence of this, developers would be unable to implement them, or the implementation would be less efficient. Furthermore, in more than half of the cases (56.12%), such cloned files are never modified, and in 75% of cases they are modified by only a small fraction (2.94%). The role of authorship in these clones, or 'super-sources', does not appear to be a significant factor in the cloning process.[51] In such cases, should such fragments, where they merely implement commonly available solutions, be protected from a copyright perspective? It seems not, especially as further research indicates that there are such projects that function directly as "hubs" for other projects, i.e. they provide more clones than the other projects.[52]

22 Open-source licensing is especially important here. Copyright is not displaced by open-source development; rather, it is operationalised through licences that permit copying, modification and redistribution on defined terms. This means that the prevalence of copying in software ecosystems does not by itself demonstrate legal failure. What it does demonstrate is that software practice relies on norms of attribution, compatibility and compliance that are often more granular than the abstract figure of the solitary author. The mismatch, therefore, is not simply between law and copying, but between a doctrinal emphasis on author-centred originality and an operational environment organised around reuse, modularity and licence stewardship.

23 Consequently, a more precise account of the mismatch between copyright and software practice is required. The problem is not that copyright law invariably prevents software from being written—plainly it does not, as the scale of commercial and open-source production shows. The problem

GitHub Projects' (*2017 IEEE/ACM 39th International Conference on Software Engineering Companion*, IEEE 2017) 161; Sebastian Baltes and Stephan Diehl, 'Usage and Attribution of Stack Overflow Code Snippets in GitHub Projects' (2019) 24 *Empirical Software Engineering* 1259.

46 ibid. In contrast, Mohammad Gharehyazie, Baishakhi Ray and Vladimir Filkov, 'Some from Here, Some from There: Cross-Project Code Reuse in GitHub' (*2017 IEEE/ACM 14th International Conference on Mining Software Repositories, Buenos Aires*, IEEE 2017) 291.

47 ibid.

48 Cristina V Lopes and others, 'DéjàVu: A Map of Code Duplicates on GitHub' (2017) 1 *Proceedings of the ACM on Programming Languages* art 84.

49 Gharehyazie, Ray and Filkov (n 47).

50 Gharehyazie, Ray and Filkov (n 47); Mohammad Gharehyazie and others, 'Cross-Project Code Clones in GitHub' (2019) 24 *Empirical Software Engineering* 1538.

51 Gharehyazie and others (n 51).

52 Gharehyazie and others (n 51); Lopes and others (n 49).

is subtler. First, uncertainty about authorship and scope of protection shifts risk into contracts, compliance procedures and private platform governance. Secondly, it complicates attribution and licence management when projects involve layered contributions from employees, contractors, occasional contributors and code borrowed under different licence conditions. Thirdly, it increases the cost of dispute resolution because courts must reconstruct a collaborative and technically mediated process using doctrinal categories developed for more linear forms of creation. In practice, software continues to be produced, but at the price of extensive private ordering and with recurring uncertainty at the margins—precisely where disputes over provenance, credit, relicensing and enforcement arise.

24 The primary nature of software source code is syntactic: programming languages tend to be fully defined by relatively simple formal grammars that specify structured and idiomatic source.[53] It is grammar that largely dictates that statements be confined to method bodies, which in turn must reside within class declarations, each of which contains strictly defined sequences of tokens that include keywords and required punctuation. The possibilities of free structuring of the content, as demanded by copyright law, are significantly limited and sometimes even impossible, leading to the fact that we are dealing with a typical 'kleine Münze'.[54] Additionally, style conventions restrict the space of available 'creative' choices for a developer, and are constantly being worked out by the community on platforms such as GitHub. Gabel and Su[55] further point out the problem of commonality in software engineering, which may lead to further similarities in source code.[56] This is particularly evident in low-level granularity, where the repetitive and idiomatic character of the code predominates. An exemplar of this is C-like languages, which are frequently replete with simple indexed for loops over known bounds. Consequently, the researchers predict in the extreme case that every token type in the language has been used at least once, and it is likely true that every legal two-token sequence has been written as well. In their view, while the number of legal statements in the language is theoretically infinite, the number of practically useful statements is considerably smaller and potentially finite. This concept is pivotal in the context of copyright law, as it underscores the legal protection afforded to unique codes, phrases, statements, or tokens.[57] Ultimately, Gabel and Su argue that the uniqueness of a code is in fact something extremely rare, precisely because of the specificity of the language and its use.

25 Consequently, a certain dissonance emerges when comparing the assumptions inherent in copyright law with those observed in programming practices. Copyright legislation treats a program as a literary work, thereby forcing us to assume that every literal line of code created is in principle subject to protection. However, it does not consider the possibility that such a line may in fact invoke or implement a function, library or other source to the program that is ultimately intended to define the value of such a program. The same line of code is repeated on GitHub itself, and over time, it becomes a 'standard' among the development community because its implementation does not entail an excessive number of bugs. Also, the same copyright law treats a single developer as the author, disregarding the circumstances, personnel structures and teams with which they work, or who and to what extent binds them with certain instructions regarding the content, functionality or quality of the code they create. A certain 'workaround' solution would probably be to regulate this appropriately in the implementation contract, but firstly, such a possibility is not always allowed by law at all, and secondly, it would significantly expand the already complicated contractual pragmatics and negotiation procedure.

D. Software Creation Is Not for Lone Wolves

26 What then is the programming reality? The studies suggest that private repositories and commercial ones function in a somewhat different manner than those run publicly (and for free). The majority of activity is centred around private repositories, which are established by software companies for the purpose of developing commercial software.[58] It could be tentatively assumed that a significant part of the creative activity in programming is initiated when it is linked to an incentive in the form of remuneration; whereas software created in open and accessible repositories is the part of the creative effort that is actually created for purposes other than economic (for oneself, for 'romantic' purposes, in any case different from utilitarian kleine Münze). However, at this time, there is insufficient evidence to support this position. Research findings also indicate that economic motivation is not the sole catalyst for software creativity. As Stephany et al.

53 Gabel and Su (n 39).

54 Stefan Martin and Jonathan Boyd, '"Software": Clear and Precise?' (2021) 16 *Journal of Intellectual Property Law & Practice* 459; Paolo Guarda, 'Looking for a Feasible Form of Software Protection: Copyright or Patent, Is That the Question?' (2013) 35 EIPR 445.

55 Gabel and Su (n 39).

56 ibid.

57 Boyle, *Shamans, Software, and Spleens* (n 36).

58 Batista and others (n 25).

noted,[59] a significant factor in software development is the comfort of working in a specific environment, usually local, in which social ties and personal acquaintances play an important role. It was the social trust that motivated Stack Overflow users to respond to queries and disseminate knowledge pertaining to code development. It also plays an important role in explaining the geography of contributions, even though it does not explain the popularity of specific questions or forums. From a legal perspective, this is significant because it shows that participation in software-related knowledge exchange is often structured by motivations that do not fit neatly within the economic assumptions traditionally associated with copyright. Contributors are frequently driven by non-economic considerations, such as reciprocity, peer recognition, community standing, and the maintenance of reliable professional relationships, rather than by the expectation of direct remuneration. These forms of engagement are embedded in networks of social trust, personal ties, and informal dependencies, which both enable and incentivise the sharing of knowledge, code fragments, and problem-solving strategies.[60] In many cases, contributors respond not because they are contractually obliged or financially rewarded, but because they operate within networks of mutual recognition and professional dependency, in which helping others reinforces one's own credibility and position within the community. In such settings, creative contribution is not adequately explained by ownership structures, contractual allocation, or market exchange alone, since the decision to contribute often arises from relational commitments and community norms rather than from formal legal entitlements. This, in turn, complicates any attempt to conceptualise software authorship exclusively through the prism of individual economic incentive, and suggests that legal frameworks based on that assumption may fail to capture the social conditions under which software is actually produced, improved, and circulated.

27 The study on the relevance of society and interpersonal relationships within programming communities seems to confirm previous intuitions and corresponds with similar studies, although so far social trust has not been considered as an element of support in the creative process. In this regard, Batista et al. observed the interactions amongst users of GitHub and noted that most of them do not interact directly, but instead they work independently until a certain event brings them together: they have a common problem to solve and communication is the only way to work together to circumvent or identify the source of the problem. However, before they enter into direct communication, they are restrained and attempt to find the source of the problem independently. Thus, despite technological and organisational pressures to collaborate, certain psychological factors suggest that developers are cautious in their interactions with others and limit them to purely technological matters. Empirical evidence also suggests that, in direct interactions, individuals tend to demonstrate a preference for collaboration with those in close physical proximity (once more: it demonstrates the importance of social relations and local connections). This finding lends further support to the initial conclusion that social trust and local community are integral components of the programming environment.[61] From the standpoint of authorship, this observation is particularly important, because it suggests that software creation depends not merely on the aggregation of technical inputs, but also on relational conditions that make such inputs possible in the first place. Mutual trust between developers facilitates the disclosure of unfinished ideas, the sharing of tentative solutions, and the acceptance of interventions by others into one's own work, all of which are central to collaborative coding practices but difficult to capture in formal legal categories. In this sense, programming appears as a creative domain with a distinctive social structure: unlike more traditional models of authorship, often built around the image of a comparatively isolated creator or a more clearly bounded act of expression, software development is frequently iterative, distributed, and dependent upon repeated acts of interpersonal coordination. The originality embodied in software may therefore emerge not only from individual expression in isolation, but also from a trusted environment in which contributors can refine, test, challenge, and rework each other's solutions. This specificity matters for copyright doctrine, because it complicates the application of authorship concepts developed with other forms of creative production in mind, and suggests that, in the software field, the social architecture of creation is often inseparable from the creative process itself.

28 The present findings correspond with the research conducted on the phenomenon of the popularity of certain repositories.[62] Popularity among repositories essentially means that such a project becomes a reference point for many subsequent ones: This phenomenon can be attributed to the frequent replication of code components within the development of emerging projects.[63] It can thus be

59 Stephany, Braesemann and Graham (n 24).

60 ibid.

61 Batista and others (n 25).

62 Oskar Jarczyk and others, 'GitHub Projects: Quality Analysis of Open-Source Software' in Luca Maria Aiello and Daniel McFarland (eds), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona* (Springer 2014) 80.

63 Gharehyazie, Ray and Filkov (n 47).

deduced that the duration of a project's existence is positively correlated with its popularity, particularly if it attains a degree of renown during its initial phase. Nevertheless, the volatility of technology and rapid changes in circulation often result in even the most popular and copied projects becoming obsolete rapidly. Of greater significance is the affiliation of the project with a particular organisation. The more prominent the brand of the company under whose umbrella the repository is operated, the greater the interest in the content of the project. As Jarczyk et al. observe, two further significant attributes—the average number of followers of developers in the project and the average number of developers/projects followed by them—can be regarded as an approximation of social relations among project members. The popularity of the developer, as determined by the researchers based on the number of followers, does not exert a significant influence on the popularity of the project itself. However, a correlation was observed between the involvement of more developers following other users and the popularity of the project. In essence, the prevalence of a project is amplified in scenarios where multiple developers are engaged, as opposed to scenarios where a solitary developer (yet recognised in the community) exhibits substantial involvement. The study found that involving more people in a project, even if they were not yet strongly recognisable and connected to the community, was more beneficial than the involvement of a single recognisable developer, who nevertheless has a rich portfolio of interdisciplinary projects.[64] As demonstrated above, repositories with a demonstrable history are generally regarded as the most reliable and authoritative sources. In most cases, such repositories are linked to well-known organisations or to identifiable developers acting under their official name, which makes the authenticity of the author a factor of considerable importance. In the software environment, the value of a contribution is therefore rarely assessed in complete isolation from the person or entity behind it. What matters is not only whether the code works, but also who produced it, under what name, with what history of prior contributions, and with what degree of recognised competence. A visible authorial track record functions here as a mechanism of trust: it helps users and collaborators assess reliability, predict quality, and decide whether a given repository, package, or solution is worthy of adoption or further development. This is significant because it points to a model of authorship that departs, at least in part, from the canonical assumptions of copyright theory. Traditionally, authorship has been closely linked to intellectual creation, individual expression, personality, and the justificatory framework associated with freedom of expression and moral rights. In that picture, the author is recognised primarily because he or she originates an expressive form that reflects individual creative choices. In software development, however, public recognition of authorship is often shaped by a wider and more socially mediated set of factors. Authorial visibility may depend not only on the act of creation itself, but also on authenticity, continuity of participation, responsiveness to the community, technical authority, and the capacity to maintain a credible and recognisable public presence over time.

64 Jarczyk and others (n 63).

29 It is also noteworthy that active participation in the development of communities and the expansion of projects does not necessarily translate into the popularity of the projects in which users are involved. Yet such participation still performs an important authorial function, because it enables individuals to establish themselves within the development community and to build a personal brand that becomes relevant to the reception of their work. The formation of authorial recognition is thus interwoven with many elements that are more closely related to the person of the author than to the abstract need to create or to express ideas. Reputation, recognisability, prior conduct, community standing, and perceived reliability all contribute to whether a contribution will be noticed, trusted, and attributed significance. From this perspective, software authorship appears as a partly networked and reputation-based phenomenon, in which the social construction of the author may be just as important as the creative act itself. This does not displace the traditional copyright image of authorship altogether, but it does suggest that, in the software field, that image is incomplete unless it also accounts for the relational, public, and credibility-based conditions under which authors come to be recognised as authors.

30 The above research demonstrates that within the domain of programming, genuinely solitary authorship is relatively rare. Contrary to intuitions drawn from other forms of creative activity, code development appears to be predominantly a collective practice, shaped by social trust, collaborative methodologies such as agile and scrum, and, to some extent, economic incentives, including remuneration for the work performed. This observation is particularly important because it indicates that software creation cannot be adequately understood through models that presuppose an individual creator acting in relative isolation. Even where financial compensation is present, the production of code remains deeply dependent on interpersonal coordination, iterative revision, and shared technical environments. In that sense, software development differs not only from more individualised forms of authorship, but also from forms of voluntary or undercompensated creativity in which collaboration may occur without becoming so structurally central to the act of creation itself.

At the same time, the use of the same technological tools, especially GitHub, appears to strengthen collaboration without necessarily increasing the discoverability or originality of solutions.[65] What gains importance instead are factors such as the authenticity of the author and the reliability of the author's track record. Affiliations with recognised organisations managing a project, as well as standing within the relevant community, substantially affect the project's popularity and reputation. This means that the reception and authority of software outputs are shaped not only by the intrinsic quality of the code, but also by the social identity and credibility of those associated with it. The practical construction of authorship in software therefore involves more than the production of original expression; it also depends on visibility, trustworthiness, continuity of contribution, and community recognition. These elements are not peripheral. On the contrary, they are often central to determining whose contributions will be accepted, developed further, or treated as authoritative within the collaborative environment. When these findings are juxtaposed with the prevailing legal framework, a significant gap becomes apparent. Copyright law, particularly in its treatment of co-authorship, continues to rely on assumptions that fit poorly with the realities of software production, including the expectation of a concluding agreement between creators. Yet, as the analysis of GitHub pull requests suggests, software creation often unfolds through open, incremental, and socially mediated interactions in which formal closure between contributors may be absent or legally irrelevant to the actual emergence of the work. The context of collaboration, the architecture of social relations, the presence of trust, and the openness of the development process—factors repeatedly identified as central by empirical research—remain largely invisible within the doctrinal model. This is precisely why the software context calls for a revision of the classical model of authorship. A legal framework developed around more bounded and individualised acts of creation may be insufficient where creative production is distributed, iterative, and dependent on relational infrastructures that the law does not presently recognise. The problem, therefore, is not merely that copyright law offers incomplete answers, but that its underlying image of the author may itself be too narrow to account for how authorship is actually formed in contemporary software practice.

E. The Social Reactions

31 A substantial proportion of software development occurs not only on the GitHub platform, but also on Stack Overflow. Unlike GitHub, however, StackOverflow operates as a forum in which users pose questions and receive responses in the form of comprehensive solutions. These replies often include lines of code that users can directly incorporate into their own projects.[66] Research into the phenomenon of this platform (98% of developers report using it with high frequency[67]) has indicated that users are motivated to respond not necessarily by an altruistic need to support others, but rather by a need to build their own brand and recognition.[68] In contrast to other instances of community creation and collaboration, Stack Overflow users primarily seek knowledge but are reluctant to share or assist others, unless they recognise in this an opportunity for social empowerment.[69] In essence, software engineers adopt a selective approach when it comes to making voluntary contributions to publicly accessible projects. Gabel and Su's study revealed that more experienced and active developers were more inclined to reproduce code from other projects than to generate it autonomously. This finding suggests that software development proficiency may not depend on illusory creative genius, but rather on a skilled capacity to locate information on Stack Overflow and GitHub, and to navigate between different projects. It points to two insights that have already been touched upon.

32 First, the motivations underlying creation may be even more significant in software engineering than in many other creative fields. This does not mean, however, that economic incentives remain the dominant driver; social recognition, professional credibility, and inclusion within trusted technical communities may operate as equally powerful catalysts. From a legal perspective, this matters because copyright doctrine has traditionally been built around a comparatively narrow image of the author as an individual creator moved either by economic reward or by the expressive logic associated with personality and freedom of expression. In software practice, by contrast, the incentive to contribute is often bound up with collaborative reputation, peer validation, and the accumulation of technical authority within a community. That broader motivational structure should not be treated as merely incidental, since it shapes both the conditions under which software is created and the ways in which particular contributors come to be recognised as authors in practice.

65 Braga and others (n 27).

66 Iraklis Moutidis and Hywel T P Williams, 'Community Evolution on Stack Overflow' (2021) 16 *PLOS ONE* e0253010; Sohaib Mustafa, Wen Zhang and Muhammad Mateen Naveed, 'What Motivates Online Community Contributors to Contribute Consistently? A Case Study on Stackoverflow Netizens' (2023) 42 *Current Psychology* 10468.

67 Baltes and Diehl (n 45).

68 Mustafa, Zhang and Naveed (n 67).

69 ibid.

33 Secondly, authenticity and a proven track record are important not only because they help to build and consolidate the popularity of projects, but also because they may themselves become objectives pursued within the software community. A visible history of reliable contributions functions as a marker of trustworthiness, making it easier for a developer's work to be accepted, reused, and extended by others. In that sense, authorial recognition in software often emerges through a process of public credibility-building rather than through a single, isolated act of creation. This sits somewhat uneasily with legal models that continue to focus primarily on the moment of fixation or on formal categories of authorship, while paying less attention to the social mechanisms through which authorship is actually attributed and validated. The point may be illustrated, indirectly, by UK case law: in *Nova Productions v Mazooma Games*,[70] the courts drew a firm distinction between the protected expression found in code and the broader "look and feel" of the resulting software, thereby revealing how copyright doctrine tends to isolate discrete acts of expression even where the practical value of software is shaped by a wider collaborative and reputational context. More recently, *THJ Systems v Sheridan*[71] reaffirmed that originality depends on the author's own intellectual creation and on identifiable creative choices, yet the dispute also shows how strongly software authorship questions are intertwined with issues of technical control, contribution history, and attribution in ongoing collaborative settings. Taken together, these considerations suggest that, in software, the making of an "author" is often inseparable from authenticity, reputation, and community-based trust, even though such factors remain only imperfectly reflected in the classical legal account of authorship. In this context, Baltes et al.[72] conducted a study on the social reactions of users who copied other people's code into their projects. The need to reinforce one's own authenticity contradicts the common practice of not attributing authorship to the person who actually created the code. Indeed, 49% of developers reported that the last time they copied or adapted a code snippet from Stack Overflow, they did not attribute its origin.[73] This apparent contradiction can be explained by the fact that authorship in the legal sense—that is, being the author of the code as a copyright matter—is not necessarily aligned with attribution as it functions in programming practice. Judging from developers' reactions, the latter is often more important. Even where programmers do not identify the author of the copied code by name, more than 40% of respondents reported adding a source-code comment containing a link to the relevant question or answer. In this environment, classical attribution through naming or signing the work is frequently displaced by reference to a repository, thread, or platform entry, which may be associated with a person, but does not necessarily belong to that person in any legally meaningful sense. What circulates, therefore, is not always the author's name as such, but a traceable point of origin within the technical ecosystem. This form of attribution is usually invisible to outsiders without access to the code itself, yet it remains visible, intelligible, and normatively significant within the developer community, where it is treated as an ordinary and legitimate practice.

34 Although such attribution is not legally sufficient, it performs an important social function in programming pragmatics. In a field where authenticity must be established continuously rather than presumed, developers reinforce their credibility not primarily through "authored publications" in the traditional sense, but through visible activity, sustained participation, and the number of references, forks, citations, and reuses attached to their work. Authenticity is thus built cumulatively and relationally: it depends on being seen, referenced, and recognised across the network, rather than on the formal assertion of authorship over discrete works. This is an important departure from more classical understandings of authorial recognition, in which the author is identified through the work as a bounded expressive object. In software culture, by contrast, recognition often attaches to a contribution history, a repository presence, or a pattern of problem-solving that others learn to trust. From a legal perspective, this suggests that practical forms of attribution in software are organised around reputation and verifiability rather than around the canonical copyright model of authorial designation.

35 The survey results concerning the source of copied code illustrate this point further. The majority of respondents (58%) identified Stack Overflow as the source. By contrast, some of those who claimed that the code had not been copied, even where an exact equivalent could be found on that platform, stated either that they had written it themselves or appeared simply unaware that they had reproduced existing material. This may partly reflect social desirability bias,[74] but it may also indicate a deeper uncertainty within the community as to where reuse ends and authorship begins. That uncertainty is legally significant. If developers do not clearly

70 *Nova Productions Ltd v Mazooma Games Ltd* [2007] EWCA Civ 219, [2007] RPC 25, affirming *Nova Productions Ltd v Mazooma Games Ltd* [2006] EWHC 24 (Ch), [2006] RPC 14.

71 *THJ Systems Ltd v Sheridan* [2023] EWCA Civ 1354, [2024] ECDR 4.

72 Baltes and Diehl (n 45).

73 ibid.

74 Anton J Nederhof, 'Methods of Coping with Social Desirability Bias: A Review' (1985) 15 *European Journal of Social Psychology* 263.

distinguish between original creation, adaptation, and copying, then the doctrinal assumptions on which copyright attribution is based become difficult to apply in practice. The problem is not merely one of compliance, but of conceptual mismatch between legal categories and everyday programming behaviour.

36 This mismatch is reinforced by the strikingly low level of legal awareness revealed by the available research. As Baltes et al. observe, up to 75% of respondents were unaware of the licensing terms governing the code snippets shared on Stack Overflow, namely CC BY-SA 3.0.[75] More broadly, the evidence suggests that many developers remain uncertain not only about licensing conditions, but also about the basic boundaries of copyright protection itself—what is protected, what is not, and under what conditions reuse is lawful. This points to the distinctly secondary role of copyright law in the normative universe of the programming community. In practice, developers appear to orient themselves less by formal legal rules than by community customs, technical convenience, and shared expectations about acceptable reuse. Copyright may continue to structure legal risk ex post, once a dispute arises, but it seems to exercise only limited influence ex ante over the everyday behaviour of developers. For copyright theory, this is a significant finding: in the software context, legal authorship and licensing rules do not appear to function as the primary behavioural framework for creative practice. Instead, programming communities are governed to a substantial extent by informal norms of cooperation, authenticity, and reputational trust, which often matter more in practice than the law designed to regulate authorship itself.

37 These findings help answer the objection that current copyright law cannot be much of a hindrance because software production is flourishing. Flourishing output does not mean optimal legal fit. A legal regime may still be poorly calibrated if its frictions are absorbed by platforms, employers, maintainers and developers through extra-legal workarounds. The evidence surveyed here suggests exactly that: attribution is often technical rather than legal, authorship is often settled ex ante by contract rather than ex post by doctrine, and compliance with licence conditions is uneven because many developers are only partially aware of the legal terms attached to the code they reuse. The law's categories are therefore not irrelevant; they are redistributed into governance costs, evidentiary burdens and downstream disputes.

75 Baltes, Kiefer and Diehl (n 45); Baltes and Diehl (n 45).

F. Conclusions

38 The foregoing analysis leads to four principal conclusions, each of which responds to a specific doctrinal weakness in current software copyright discourse. First, the concept of software authorship is not as simple as we imagine it to be for the purposes of copyright discourse. Rather, it is a form of social, multifaceted teamwork, whereby participants influence the entire software production process in various ways. The team's relations primarily affect the production process itself, thereby influencing the final shape of the work. Furthermore, the underlying assumptions about the programming language and platform on which the code is created and stored, as well as the pragmatics of program development, effectively force participants to collaborate. Comfort and social trust in the working environment (broadly defined, and primarily in the online space within repositories) are also important factors in this, although the latter is quite specifically built. Here, there are no transparent, classical agreements that are legally binding. The interactions are purely technical and are not the first choice of the developers, even though all the other elements of this environment compel them to cooperate.

39 Secondly, the assumption that all computer program code is automatically protected by copyright is incorrect. The nature of programming languages significantly limits the choices available to those creating code, and there are also restrictions relating to functionality, the implementation of specific elements and even style conventions. Furthermore, a large proportion of the code in almost every project contains elements that have been copied from elsewhere. The personal and economic motivations of those creating the code also play a considerable role in software engineering. Although no research has been conducted into the altruistic motives of professional software developers, the significant activity in commercial projects and reluctance to interact directly, except when unavoidable, suggests that incentives are necessary to encourage collaboration. The popularity, development and quality of a project may also be influenced by the authenticity of the author and the proven track record. In this environment, the author's position and the ability to verify their achievements and project portfolio seem to be more important professionally than in other creative fields, especially compared to artistic or musical activities. Thirdly, the mismatch between copyright theory and software practice is best understood as a problem of calibration rather than outright incompatibility. Software production continues at scale, but legal uncertainty is absorbed through private ordering: contracts, repository rules, contributor policies, compliance workflows and informal attribution practices. This does not negate the mismatch; it shows where its costs

are borne. As argued above, contractual practices undoubtedly play an important coordinating role, yet they cannot fully capture the fluid, iterative and multi-party character of software development, nor can they anticipate every socially relevant contribution made during the life of a project. For that reason, many of the tensions left unresolved by copyright doctrine are managed in practice not by law alone, but by community conventions grounded in trust, reputation, authenticity and shared expectations concerning acceptable reuse and attribution. These conventions often prove more immediate and behaviourally effective than either statutory rules or contractual provisions, precisely because they are embedded in the everyday pragmatics of programming. The resulting picture is not one in which law disappears, but one in which legal and contractual regulation is supplemented—and in important respects functionally displaced—by social norms that more closely track how software is actually created, shared and recognised.

40 Fourthly, and normatively, reform should proceed on several levels. At the interpretive level, courts should apply a stricter originality analysis in software cases and should resist collapsing technical contribution into legal authorship without evidence of authorial expression. They should also distinguish more carefully between joint creation, technical facilitation, editing, project management and code review. At the legislative or policy level, lawmakers and soft-law bodies should encourage clearer guidance on collaborative software authorship, evidentiary presumptions for repository-based contribution records, and the interface between code attribution practices and moral-rights discourse. At the transactional level, software contracts and open-source governance documents should be drafted not merely to assign rights, but also to document contribution pathways, licence provenance and responsibility for compliance. Taken together, these steps would not abolish software copyright. They would make it more modest, more transparent and better aligned with the realities of contemporary software development. The revised framework proposed here is therefore not a replacement for copyright, but a methodological correction to its application in the software field: specify the territorial baseline, test originality rigorously, separate legal authorship from technical attribution, and treat collaboration and licensing as central rather than peripheral features of code production.

41 Taken as a whole, the foregoing analysis points towards a broader conclusion for copyright law: in the software environment, authorship is not primarily organised through the legal grammar of exclusive creation, formal attribution, and clearly bounded ownership, but through conventions that are social, technical, and reputational in nature. What governs behaviour in practice is often not the copyright rule or even the contractual clause, but the community norm—how copying is perceived, when reuse is accepted, what counts as proper acknowledgement, and whose contributions are treated as credible within the relevant network. The frequency of copying, and the relative normality with which it is socially received, show that the movement of code across projects is not experienced as an exceptional infringement-prone event, but as an ordinary feature of programming pragmatics, provided that it remains within recognised conventions of use, reference, and trust. At the same time, the motivations for creating and sharing code remain normatively ambiguous: economic reward matters, but so do reciprocity, peer recognition, technical curiosity, community standing, and the cultivation of a visible, authentic track record. This gives rise to a model of authorial recognition that differs markedly from the historical copyright image of the author as a solitary, expressive, romantic genius. In programming culture, recognition is often accumulated through iterative participation, problem-solving visibility, repository presence, and repeated citation by others, rather than through the production of a discrete and self-contained "work" tied to a single authorial personality. The implication for copyright law is not simply that doctrine requires minor adjustment, but that its underlying assumptions about authorship may be misaligned with one of the most important contemporary domains of creation. Where conventions are more behaviourally effective than legal rules, and where social legitimacy matters more than formal entitlement in structuring everyday practice, the law cannot plausibly continue to treat such conventions as peripheral. Rather, they should be understood as central to any serious rethinking of authorship in software.

Acknowledgments: This work was supported by the National Science Centre in Poland (Narodowe Centrum Nauki) under grant no. 2023/49/B/HS5/00772.

The Code as an Element of Trust in Smart Contract Arbitration Clauses

by **Eduardo Andrés Calderón-Marenco, Tania Rodríguez Palacios, Jaime Elias Torres Buelvas and Piedad Barreto-Granada ***

Abstract: This paper addresses the growing intersection between blockchain technology, specifically through the use of Smart Contracts (SC), and international arbitration as a mechanism for resolving contractual disputes. In the context of a rapidly evolving digital era, this paper proposes the standardization of the code used to trigger arbitration clauses in Smart Contracts, with the aim of fostering legal certainty and trust in these automated processes. This research is based on an in-depth documentary analysis, examining academic sources and online dispute resolution platforms, to identify and synthesize best practices in the codification of such clauses. Through this exploration, the complexity inherent in translating legal agreements into computer code is revealed, addressing the need for tools that enable users to design and implement arbitration clauses with technical precision and legal validity. This article suggests that, while Smart Contracts offer a promising avenue towards efficiency and transparency in commercial and legal transactions, their full integration into the international arbitration arena requires a standardized framework to ensure legal consistency and accessibility for non-programmers.

Keywords: Contract, Code, Intelligent, Blockchain, Block, International arbitration

© 2026 Eduardo Andrés Calderón-Marenco, Tania Rodríguez Palacios, Jaime Elias Torres Buelvas and Piedad Barreto-Granada

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: Eduardo Andrés Calderón-Marenco, Tania Rodríguez Palacios, Jaime Elias Torres Buelvas and Piedad Barreto-Granada, The Code as an Element of Trust in Smart Contract Arbitration Clauses, 17 (2026) JIPITEC 176 para 1.

A. Introduction

1 SCs are efficient instruments for contracting based on new technologies, and although they facilitate their automatic execution, they are not infallible. In other words, disputes may arise as a result of non-compliance, and in such circumstances which must be provided for in the contract arbitration is the ideal means of resolving such disputes due to its characteristics.

2 SCs are a type of electronic contract, which implies a drafting in computational language or codification of content, and in this process the validity of the contract may be put at risk. Although this risk is born of the autonomy of the will of the parties, it is possible that the code will not translate the true will of the parties as it is a language in Boolean logic, generating distrust regarding the activation of legal mechanisms in the event of non-compliance, which are expressed in the arbitration clause or agreement.

3 Thus, this research is born from the interest of studying the codification of the clauses that activate the arbitration mechanism, its components, and the codified languages that currently exist, to finally propose a standardization of the codification of the arbitration clause in order to provide elements of confidence to cyber-consumers and legal validity to such clauses in the event of such an eventuality.

4 However, this paper starts from the premise that the automation of arbitration through smart contracts is not a neutral or axiologically innocuous solution, but rather a phenomenon that needs to be analyzed

in light of the structural principles of contract law and international commercial arbitration.

B. Methodology

5 This research is of a theoretical, descriptive legal nature, which is why the analytical- synthetic method has been adopted. Therefore, it does not seek to formulate definitive regulatory solutions, but rather to identify doctrinal tensions and structural limitations in the interaction between smart contracts and arbitration.

I. Planning the Search and Selection of Information

6 A review of academic literature was conducted in web databases of scientific dissemination, corresponding to the publication of the last seven years (2018-2025) from the following search keywords (Codification of the arbitration clause, Law applicable to Smart Contracts, Lex Cryptographia, Blockchain, arbitration, programming language). Likewise, master's and doctoral papers deposited in open access institutional repositories from the same time period were consulted.

II. Inclusion and Exclusion Criteria

7 Subsequently, a preliminary information filtering process was carried out, as follows:Filter 1-The inclusion criteria taken into account were:

a) articles in scientific journals,

b) authors with a master's degree or doctorate,

c) articles in Spanish or English and in institutional repositories, and that they were master's or doctoral works; 136 documents were obtained from this process.

8 Filter 2-In this stage, the following exclusion criteria were taken into account:

a) that the articles were repeated in several databases,

b) that the reading of the abstract did not correspond to the research topic; thus, 106 corresponding; documents were obtained: 106

Table 1.Results of the information search

Source: Created by authors based on information search results.

Author	Concept	Technical aspects	Legal aspects
Szabo	"A smart contract is a computerized transaction protocol that executes the terms of a contract"	Protocol Computerized transaction	Execute terms
Szabo	"A smart contract is a set of promises, specified in digital format, that includes protocols in which the parties fulfill those promises".	Protocol	Set of promises
Sklaroff	"Smart contracts are decentralized agreements created in computer code and stored on a blockchain".	Computer code Stored on a blockchain	Decentralized agreement
Werbach & Cornell	"Smart contracts are digital transactions that execute automatically and use decentralized cryptographic mechanisms for enforcement"	Digital transactions Use cryptographic mechanisms	They run automatically decentralized
Hu; Zhang; Liu & Lin	"Smart contracts are typically defined as event-driven computer programs that are executed and enforced by all participants in a peer-to-peer network in the context of blockchain"	Computer programs P2P network in the context of blockchain	Events executed by participants
Shamala; Balasaraswathi & Gayathri	"Smart contracts are undoubtedly a significant advance in the field of blockchain technology. They are self-executing agreements in which the terms of a transaction between buyer and seller are explicitly encoded in the program. Provided that the conditions specified in the code are met, the smart contract carries out the agreed activities immediately, without the need for intermediaries or third parties".	Blockchain technology Condition coding	Automatic enforcement agreements

C. Smart Contracts (SC)

9 Undoubtedly, the Internet has facilitated the creation of new legal businesses[1], it has rapidly blurred territorial borders in the conclusion and execution of contracts, and today we are witnessing a complex network of information transmission and transactions at a global level[2], in this scenario SCs are born to legal life.

10 Contracts are particularly important for the development of societies on a scale that goes beyond the merely commercial, with a legal tradition dating back thousands of years that seeks to provide security regarding the essential elements and aspects related to their formation and execution. The SC scenario makes it necessary to consider an analysis from a dual perspective, both technical and legal, in search of new balances. From a technical perspective, SCs offer many advantages, such as minimizing risks, the involvement of trusted third parties, commercial costs, and transaction costs, among others[3]; while from a legal perspective, new challenges and possible inefficiencies have arisen, even from basic aspects such as moving away from human language in their construction[4].

11 In this scenario, SCs come to legal life. The following table compiles contributions from some seminal authors, which demonstrate the maturation of the elements and characteristics of the concept:

Table 2. Academic contributions to the concept of SCs

Platform	Description
Kleros	Kleros is a decentralized platform that uses blockchain technology to provide a transparent and fair arbitration system. It uses a jury selection mechanism based on game theory, where jurors are randomly selected to evaluate evidence and make decisions in disputes related to smart contracts and other Digital agreements (Aouidef et al., 2021)
Aragon Court	Aragon Court is a platform designed to resolve disputes with in the ecosystem of Aragon's Decentralized Autonomous Organizations (DAOs). It provides a dispute resolution mechanism that allows DAO users to submit their disputes to a virtual panel of judges, who evaluate the evidence resented and make a collective decision based on pre-established community rules (Aouidef et al., 2021).
Jur	Jur is an innovative platform that offers arbitration services for commercial and legal contracts, leveraging blockcha in technology to ensure fairness and transparency in the dispute resolution process. It focuses on providing an efficient and accessible solution for commercial dispute resolution, facilitating an environment of trust for all parties involved (Schmitz and Rule, 2019).
OpenLaw	OpenLaw is a powerful tool that facilitates the creation, execution and Management of legally supported smart contracts. It allows users to easily integrate arbitration clauses into their digital contracts, providing a framework for automated, blockchain-based dispute resolution. Open Law stands out for its flexibility and for offering a user-friendly interface for Drafting digital legal contracts (Schmitz and Rule, 2019).
Mattereum	Mattereum is a platform that seeks to bridge the gap between physical and digital assets through smart contracts, providing a robust legal infrastructure for trading these assets. It includes an arbitration service to handle disputes, ensuring that agreements are executed fairly and efficiently. Mattereum is pioneering the use of Blockchain technology to create an automated and Decentralized legal system (The Mattereum Team,2018).

Source: Created by authors based on the sources cited.

1 Tatiana D Zabala, 'La ética en la inteligencia artificial desde la perspectiva del derecho' (2021) 16(2) *Via Inveniendi et Iudicandi* 1.

2 Eduardo Perafán, Sebastián Polo, Marco Sánchez and Catalina Miranda, 'Estado y soberanía en el ciberespacio' (2021) 16(1) *Via Inveniendi et Iudicandi* 1.

3 Nick Szabo, 'Smart Contracts' (1994) http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart.contracts.html accessed 13 September 2026.

4 Jeremy M Sklaroff, 'Smart Contracts and the Cost of Inflexibility' (2017) 166 *University of Pennsylvania Law Review 263.*

12 Amid this dual technical and legal relationship of SCs, in order to avoid risks and inefficiencies, Szabo, as a pioneer in conceptualizing this new contracting environment, proposed four basic objectives for the design of SCs in 1996: observability, verifiability, privity, and enforceability. Their meaning, in this same order, is related to the following aspects: a) ease of access and ability of the parties to closely monitor compliance with the contract, b) monitoring or auditing functions regarding compliance, c) contractual privacy that must be maintained regarding the content and control of compliance with a contract exclusively by the parties; and finally, d) when "self-enforcement" protocols exist, it is important for the parties to know the phases in which these circumstances occur and to be able to verify them, as external elements such as the proper functioning of the technological infrastructure can have a particular impact on self-enforceable phases.

13 More recently, authors such as Hu, Zhang, Liu & Lin (2021) propose security and correctness as the pillars of SCs; security as the "ability to resist unauthorized state changes, including fund transfers, state manipulation, and accidental self-destruction" and correctness as the "ability to correctly perform the expected function," while privacy and efficiency are properties provided by cryptography and hardware equipment.

14 From the perspective o flaw, some authors consider that SCs are computer programs that translate agreement previously agree upon by the parties, i.e.they are software, given that they are expressed in a computer code containing an agreement for automatic execution, a new computerized form of executing all or some of the obligations contained in a previously agreed contract[5].

15 In other words, the software is the computer support of a contract that according to Centeno[6] is agreed to in traditional language and then encoded and subsequently included in the blockchain or *Blockchain* technology, "in this way, the contract would have two languages. One composed of the codes that are going to be managed in the *Blockchain*, and another that is going to be of legal content"[7].

16 Thus, meaning in the context of *Blockchain*, computer code that is stored in a *Blockchain* and can be accessed by one or more parties[8];and with respect to the SC part of legal content, it contains the information on how the arbitration mechanisms are activated in the face of any conflict that may arise between the parties[9].

17 Some countries have begun to recognize in their legislations the conceptualization of SC, i.e., through the Decree-Law of December 14, 2018, Article 8 Italy proposes as characteristics, that SCs are "computer technologies and protocols that use a shared, distributed, replicable, simultaneously accessible and architecturally decentralized record on a cryptographic basis"[10], accessible by participants, but not alterable or modifiable. The Decree-Law further specifies "the storage of an electronic document through the use of technologies based on distributed records produces the legal effects of electronic time stamping referred to in Article 41 of Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014"[11].

18 Camacho López[12] explains that in the United States of America, the first State to legislate on SCs was Arizona in 2017, by introducing Article 5 which defines Blockchain technology and smart contract Title 44 of its statutes, dedicated to exchanges and

5 Kristel G Loayza, 'Legal Analysis of Smart Contracts Based on Blockchain Technology in Business-to-Consumer (B2C) E-Commerce' (thesis, Universidad Católica San Pablo 2019). Juan José Almonacid Sierra and Yeison Coronel Ávila, 'Aplicabilidad de la inteligencia artificial y la tecnología bloc chain en el derecho contractual privado' (2020) 38 *Revista de Derecho Privado* 119, 119–42.

6 Rafael J Centeno, 'Introducción al blockchain, contratos inteligentes y la aplicabilidad del arbitraje a esta tecnología' (2020) 1 *Anuario Venezolano de Arbitraje Nacional e Internacional* 483 https://avarbitraje.com/wp-content/uploads/2021/03/ANAVI-No1-A22-pp-483-500.pdf accessed 13 September 2026.

7 ibid.

8 EU Blockchain Observatory and Forum, 'Legal and Regulatory Framework of Blockchains and Smart Contracts' (Thematic Report, 2019) https://www.eublockchainforum.eu/sites/default/files/reports/report_legal_v1.0.pdf accessed 26 August 2024.

9 ibid. p. 6.

10 Decreto-legge 14 dicembre 2018, n 135, 'Disposizioni urgenti in materia di sostegno e semplificazione per le imprese e per la pubblica amministrazione' [2018] GU Serie Generale No 290.

11 ibid.

12 María Elisa Camacho López, 'Estudio de los contratos legales inteligentes y de sus normas aplicables en el ámbito del derecho del consumo en Colombia' (master's thesis, Universidad Externado de Colombia 2021) 35 https://doi.org/10.57998/bdigital.handle.001.3984 accessed 13 September 2026.

commerce, on electronic transactions. Indeed, Article 5 item e indicates that:

> Smart contract means an event-driven, stateful program running on a distributed, decentralized, shared and replicated ledger that can assume custody and provide instructions for the transfer of assets in that ledger.[13]

19 Thus, SCs are characterized as computer programs, based on a system of logical rules designed to execute the services of a contract, which depend on a sequence of programmed instructions, generating rights and obligations, as well as legal effects. Finally, the obligations are self-executing once they are activated and compliance with the established conditions is verified[14].

20 On the other hand, *Blockchain* technology according to Centeno[15] is an autonomous system, based on the operation of nodes working under a computerized protocol, i.e., it does not need a centralized system, Therefore, the trust, validity, and security of transactions does not come from any intermediary but from the network, thus representing a more transparent system, in which all nodes can participate. However, this system does so without altering the stored information because it is protected by cryptographic algorithms, and given its distributed nature, it makes a possible cyber crime more difficult, since all the copies stored in the network must be attacked at the same time[16].

21 In practice, SC as automatic mechanisms for the execution of computer instructions, in this case contract execution, became a reality from the development of *Blockchain* -- specifically since 2014-- with the creation of Ethereum (Domínguez, 2022).

22 An SC takes the form of a computer code (through a programming language, e.g. *Solidity*) in a distributed ledger technology ,which can be a *Blockchain*, and is executed when it receives electronic data inputs. It is an agreement between two parties whose purpose is to execute previously agreed upon obligations, replacing physical paper with a predetermined computer code with the will of the parties.

23 The use of SC involves a high-level language (compiled binary code), and the use of special software, which transforms a language such as Solidity into machine instructions expressed in byte code[17], operates according to the algorithm written in it, without human intervention, thus eliminating intermediaries and maximizing the efficiency of business processes.

24 It is worth mentioning, that in the period between 2018 and 2023, the use of SC has experienced exponential growth, mainly in the finance, insurance, real-estate, and supply chain sectors. However, tests have also begun to use smart contracts and blockchain in judicial case management, proposing a private-to-public blockchain approach in Pakistan[18]. According to research by Asmakov[19], the number of SCs deployed on the Ethereum network grew by 293 percent in 2022 and almost 300 percent in 2023, highlighting their growing application in areas such as decentralized finance (DeFi) and non-fungible tokens (NFTs).

25 While the positive results are undeniable in terms of trust provided by the use of *Ethereum* blockchain technology because of the highly specialized programming languages it uses such as *Solidity*, it is important to note that despite its technological efficiency it is difficult for ordinary users to design a usable SC based on their own demands[20]. A consequence of this difficulty that ordinary users have is revealed by Hoffman's study, who points out that how before 2021, *Ethereum* reveal confidential information from blockchains, which should not be exposed to the public and could potentially include identities, contract data as well as legal data[21].

26 This is how the intervention of the State is timely, by improvements in legislation and regulation around *blockchain* and SC technology providing a clearer legal framework that supports and drives its use.

27 Some scholars have proposed the use of Ricardian contracts or hybrid contracts as a mechanism to reinforce the accreditation of free consent and as a full expression of will in smart contracts. In this configuration, a text in natural language describing the terms and conditions to be automated is incorporated into the executable logic of the smart contract, which increases contractual transparency and makes it easier for a non-technical party to

13 ibid.

14 Tahir Alyas and others, 'Multi-Blockchain Architecture for Judicial Case Management Using Smart Contracts' (2025) 15 *Scientific Reports* 8471 https://doi.org/10.1038/s41598-025-92842-8 accessed 13 September 2026.

15 ibid. p. 6.

17 Agnieszka Kubiak-Cyrul, 'Challenges of Smart Contracts in Contract Law: Do Algorithmic Tools Undermine Human Autonomy?' in L Miraut Martín and M Załucki (eds), Artificial Intelligence and Human Rights (Dykinson 2021) 327.

18 Ibid. p. 8.

19 Andrew Asmakov, 'Smart Contracts Deployment on Ethereum Grows 293% in 2022' (Decrypt, 17 January 2023) https://decrypt.co/es/119381/despliegue-de-smart-contracts-en-ethereum-crecio-293-en-2022 accessed 13 September 2026.

20 Dianhui Mao and others, 'Visual and User-Defined Smart Contract Designing System Based on Automatic Coding' (2019) 7 IEEE Access 73131 <https://doi.org/10.1109/ACCESS.2019.2920776> accessed 13 September 2026.

21 Hofmann and others (n 17) 354.

understand the scope of the obligation before expressing their acceptance. Consequently, this scheme contributes to the formation of legally valid consent, insofar as it reduces information asymmetries and more precisely defines the content of the agreement.

D. Arbitration Clauses in SC

28 Arbitration as a mechanism for resolving private disputes (civil or commercial) has proven its effectiveness compared to state justice, especially when it comes to conflicts arising in the international arena, as a digitized and decentralized dispute resolution process becomes necessary[22].

29 Since the contract is self-executing, it is generally not necessary to know the identity of the counterparty, since compliance and execution are guaranteed by the automation of the code. However, it is necessary to determine how the parties, especially when they do not know each other, will resolve differences that arise after the automatic execution of a SC[23] due to disputes of various kinds, for example, programming errors, payment defaults, among others.

I. Aspects of the validity of international arbitration clauses

30 The validity of the arbitration clause expressed in codified language can be based on the content of international instruments and national legislation on arbitration and electronic commerce, such as the Convention on the Recognition and Enforcement of Foreign Arbitral Awards of the United Nations Commission on International Trade Law[24], Article 2:

> Each Contracting States hall recognize an agreement in writing under which the parties undertake to submit to arbitration all or any differences which have arisen or which may arise between them in respect of a defined legal relationship, whether contractual or not, concerning a subject matter capable of settlement by arbitration.

31 On the other hand, the Model Law on International Commercial Arbitration of the United Nations Commission on International Trade Law[25], as amended in 2006, establishes in Article 7:

32 (1)"Arbitration agreement" is an agreement by the parties to submit to arbitration all or certain disputes which have arisen or which may arise between the min respect of a defined legal relationship, whether contractual or not. An arbitration agreement may be in the form of an arbitration clause in a contract or in the form of a separate agreement. (2) The arbitration agreement shall be in writing. (3)An arbitration agreement is in writing if its content is recorded in any form, whether or not the arbitration agreement or contract has been concluded orally, by conduct, or by other means. (4) The requirement that an arbitration agreement be in writing is met by an electronic communication if the information contained there in is accessible so as to because able for subsequent reference; "electronic communication" means any communication that the parties make by means of data messages; "datamessage (?)" means information generated, sent, received or stored by electronic, magnetic, optical or similar means, including, but not limited to, electronic data interchange (EDI), electronic mail, telegram, telex or telecopy.(5)Furthermore, an arbitration agreement is in writing if it is contained in an exchange of statement soft claim and defence in which the existence of an agreement is alleged by one party and not denied by the other. (6) The reference in a contract to any document containing an arbitration clause constitutes an arbitration agreement in writing, provided that the reference is such as to make that clause part of the contract. In fact, doctrine and jurisprudence have understood the requirement of writing an arbitration clause to be fulfilled through technological means, and this is also consistent with the electronic commerce

22 Michael Buchwald, 'Smart Contract Dispute Resolution: The Inescapable Flaws of Blockchain-Based Arbitration' (2020) 168 *University of Pennsylvania Law Review* 1369 https://scholarship.law.upenn.edu/penn_law_review/vol168/iss5/3 accessed 13 September 2026.

23 James Metzger, 'The Current Landscape of Blockchain-Based, Crowdsourced Arbitration' (2019) 19 *Macquarie Law Journal* 81 https://www.mq.edu.au/__data/assets/pdf_file/0010/866287/Blockchain-Based-Crowdsourced-Arbitration.pdf accessed 13 September 2026.

24 Convention on the Recognition and Enforcement of Foreign Arbitral Awards (adopted 10 June 1958, entered into force 7 June 1959) 330 UNTS 3 (New York Convention).

25 United Nations Commission on International Trade Law, *UNCITRAL Model Law on International Commercial Arbitration* (1985) <https://uncitral.un.org/sites/uncitral.un.org/files/media-documents/uncitral/en/19-09955_e_ebook.pdf>

legislation of those countries inspired by the Model Law on Electronic Commerce of the United Nations Commission on International Trade Law (UNCITRAL, 1996), which in its Article 5 recognizes the principle of functional equivalence, granting the same validity of a written document to data messages.

II. Decentralization of applicable law and the seat of the arbitral tribunal

33 In international contracting, operators often opt for arbitration because of its efficiency, neutrality, greater control over the proceedings, specialization of the arbitrators, and enforceability of the award. However, these advantages are weakened when the arbitration agreement is poorly drafted, as this opens the door to protracted and costly disputes over its existence, validity, effects, interpretation, or enforcement, often associated with an inadequate determination of the applicable law. Although it is often treated as an ancillary provision during negotiations, the governing law of the arbitration agreement is structural: it governs essential aspects such as its existence, validity, effectiveness, scope, interpretation, waiver, termination, and enforcement, and may condition both the commencement of the proceedings and the subsequent validity of the award.

34 The traditional solution tends to subject the clause to the law of the main contract, but in international commercial arbitration, several legal systems with applicability converge, maintaining uncertainty: the lex arbitri, which does not necessarily coincide with the law of the agreement and whose divergence can generate friction in matters of doubtful classification; the law of arbitrability, usually linked to the seat, although it is sometimes connected to the substantive law of the dispute; the law applicable to the merits, where controversial references such as lex mercatoria and ex aequo et bono decisions appear; and, finally, the law governing the validity of the arbitration agreement itself, the determination of which often poses the greatest difficulties because there is no single legal system that uniformly regulates all its aspects[26].

35 In this vein, determining the seat of arbitration takes on special relevance, and in blockchain-based environments, the decentralized architecture itself makes it substantially difficult to identify the place of arbitration and, correlatively, the applicable law. The transnational dispersion of the nodes that make up the registry makes it difficult to easily establish an unambiguous regulatory connection point and, consequently, to define the governing law of the arbitration agreement. If this precision is not incorporated into the design of the legal transaction, the risks of controversy increase and the costs of resolution rise, as the ex post determination of the competent legal framework is often complex and contentious[27].

36 Although arbitration is based on freedom of contract, its practical effectiveness presupposes the intervention of judicial bodies of a given legal system, both for support and control functions. For this reason, the designation of the seat is decisive, especially in international arbitration, as it identifies the competent state courts in the absence of supranational bodies with general powers. The seat is not equivalent to the physical place where hearings are held, evidence is taken, or deliberations are conducted, nor does it necessarily coincide, in administered arbitrations, with the domicile of the administering institution; it operates as a strictly legal point of connection. To that extent, the seat links the proceedings to the legislation and courts of a State (lex loci arbitri), so that its legal system supports the arbitration by defining the jurisdiction for support and control, determining the applicability of the arbitration rules (lex arbitri), and attributing "nationality" to the award.

37 In light of the above considerations, the decentralization of arbitration supported by SC may lead to the parties omitting to determine the seat and applicable law from the design or codification phase of the arbitration agreement. This risk is intensified in adhesion contract schemes, which are common in web environments, where the predisposing party sets the conditions and the adhering party has no real room for negotiation. The indeterminacy of these elements has adverse legal consequences, prevents the competent court from being identified with certainty for support and control functions and, in particular, for the recognition and enforcement of the award; it introduces uncertainty about the legal regime governing the arbitration and the law under which the dispute will be resolved; and it makes it difficult to specify the scope of arbitrability, with a consequent increase in the risk of challenge and possible annulment of the award.

38 In this area, since arbitration is a discretionary mechanism based on freedom of contract, it is advisable that the Ricardian contract include, in its

26 José Carlos Fernández Rozas, 'Ley aplicable al convenio arbitral en ausencia de elección expresa de las partes' (2024) 16(2) Cuadernos de Derecho Transnacional 646 doi:10.20318/cdt.2024.8936.

27 Cayetana Santaolalla Montoya, 'Los criptotribunales: un arbitraje en la cadena de bloques, con smart contracts y según la teoría de juegos' (2025) 17(2) Cuadernos de Derecho Transnacional 780–801 doi:10.20318/cdt.2025.9899.

natural language wording, an arbitration clause, a clause on the law applicable to the substance of the dispute, and the determination of the seat, in order to enable its subsequent codification and enforcement on the blockchain[28]. When this provision does not exist, it is necessary to resort to the rules of the conflict method of private international law. In such a case, the arbitral tribunal may define the law applicable to the merits by identifying the most relevant connecting factors of the legal transaction, or consider approaches such as the so-called cryptographic lex; However, in order to ensure predictability and reduce uncertainty, the preferred criterion is to apply the law of the State with the closest links to the contractual relationship.

39 This approach determines the law applicable to arbitration based on the legal system that has the "closest links" to the agreement. Unlike models that privilege a single connecting factor, this technique allows for a more nuanced weighing of factors and has been used frequently in recent decades, supported by formulations such as those in the Restatement (Second) of Conflict of Laws (1971) and Article 4 of the 1980 Rome Convention.

40 Even though the Rome I Regulation partially departed from this logic. In this context, the court must recognize the choice of law made by the parties, even when it is implicit; if that choice is not identified, it is necessary to determine the legal system with which the contract has the most real and close connection. The notion of "closest connections" is not limited to indicators of territorial proximity—such as the residence of the parties, the place of conclusion, the place of performance, the location of assets, or the market affected—because none of these has autonomous value or universal scope. The weight of each criterion varies according to the type of contract and its structure, depending on the intensity of the legal and economic connection in the specific case. Consequently, the decisive factor is not the geographical circumstance itself, but the degree to which the connecting factor satisfies the requirements of economic efficiency, certainty of transactions, and predictability, taking into account the economic purpose and legal function of the contract[29].

41 Once a solution has been identified for cases in which the SC does not provide for the applicable law, the analysis is also extended to the possible challenge of the award on grounds of nullity. In effect, Article 34 of the UNCITRAL Model Law concentrates judicial review of the award in a single channel, the petition for nullity, which excludes other means of challenge before state courts. This action is only admissible on specific grounds that the requesting party must prove, including the incapacity of a party or the invalidity of the arbitration agreement under the law chosen by the parties or, in the absence of a choice, under the law of the State of the seat.

42 Failure to notify or affect the right of defence; exceeding jurisdiction by ruling on matters not submitted to arbitration; and failure to comply with the agreement on the composition of the tribunal or the procedure, unless such agreement contradicts mandatory rules. In turn, the court may declare the award null and void when the matter is not arbitrable under the law of the State or when the award is contrary to public policy.

43 In technical and legal terms, determining the applicable law not only fulfills a substantive function with respect to the contract, but also allows for the reconstruction of the procedural framework surrounding arbitration and, in particular, judicial review of the award. When the governing legal system is identified—either by choice of the parties or by conflict criteria, such as the closest connection—a regulatory framework is obtained that facilitates: (i) determining the legally relevant seat or, at least, the State with authority to project effects on the arbitration; and (ii) establishing the statute applicable to the action for annulment, insofar as the review of the award is generally linked to the legal system of the place that serves as the institutional support for the arbitration.

44 In this regard, the applicable law serves as a starting point for identifying the competent court to hear the nullity claim, for two reasons. First, because judicial review of the award is normally structured in accordance with the lex arbitri of the State acting as the seat, that legal system defines which state court exercises the functions of support and control, the form of the challenge, and the requirements for admissibility. Second, because even when the seat is not expressly determined—a plausible assumption in decentralized architectures—the identification of the legal system most closely linked to the legal relationship allows the attribution of jurisdiction to be directed toward the courts of that State, as a final solution that reduces jurisdictional uncertainty and avoids the practical impossibility of challenging or defending the award.

45 That same law also defines the grounds for nullity and their prerequisites. In systems inspired by the UNCITRAL Model Law, for example, judicial review is channeled through an action for annulment with limited grounds, which include defects in the

28 Eduardo Calderon Marenco, 'Los elementos esenciales en los smart contract' in E Álvarez Miranda (ed), *Retos y paradigmas del derecho y la sociedad: Instituciones modernas, tecnología y justicia* (Fondo Editorial de la Universidad de San Martín de Porres 2025) <www.publicacionesusmp.com/2025/04/retos-y-paradigmas-del-derecho-y-la.html>

29 José Carlos Fernández Rozas, 'Ley aplicable al convenio arbitral en ausencia de elección expresa de las partes' (2024) 16(2) *Cuadernos de Derecho Transnacional* 646.

arbitration agreement (incapacity or invalidity under the chosen law or, failing that, the law of the State), violations of the right of defence, excess of jurisdiction, irregularities in the composition of the tribunal or in the proceedings, as well as substantive limits related to arbitrability and public policy. Thus, the determination of the applicable law makes it possible to anticipate not only which judge will hear the application for annulment, but also what standard of review will be used, what substantive and procedural limits are relevant, and under what parameters the validity of the award will be assessed.

46 Consequently, in arbitration supported by SC, establishing the applicable law (ideally together with the seat) helps to ensure the enforceability of the award and the predictability of the judicial review system. If that determination is lacking, the use of a conflict-of-laws criterion such as the closest connection not only guides the governing law of the contract but also provides an argumentative basis for anchoring judicial jurisdiction and the nullity regime in an identifiable legal system, which mitigates the risk of defencelessness, forum shopping, and subsequent challenges due to regulatory uncertainty.

47 In light of the above, decentralization and the transnational projection of blockchain technology strain the classic criteria for determining jurisdiction and applicable law, requiring regulatory frameworks capable of providing effective responses to disputes arising in digital environments. Similarly, the analysis of SCs requires the incorporation of an ethical and philosophical dimension associated with the transition toward the automation of contractual relationships. In particular, the debate on the desirable scope and feasibility of formalizing legal language through its codification raises a structural question about the relationship between technology, law, and society. The possibility of translating legal categories and reasoning into computer code ultimately raises substantive questions about justice, equity, and autonomy in the digital age[30]. Therefore, in the subsequent analyses of this research, structural proposals will be made to guarantee the legal security of users in decentralized environments.

48 Other authors, such as Rab[31], argue that it is advisable to qualify the scope of this solution somewhat. This is because, although the Ricardian contract can reduce information asymmetries in the consent formation phase, it does not eliminate the problems that arise ex post, when the execution of the code produces effects that diverge from the text in natural language. In such cases, the critical issue is not so much the clarity of the initial consent as the determination of which of the two languages—legal or computational—should prevail, especially in automated and transnational environments.

49 Platforms that provide the dispute resolution service

50 From the increasing interaction of SC users, platforms have been developed that provide the service of dispute resolution, such as arbitration, within the blockchain environment. The following are some of them.

30 Eduardo Andrés Calderón Marenco, Jimmy Enrique Garzón Solano, Romina M Sánchez Silveyra, Gabriel Ravelo-Franco and Guillermo O Sal, 'Retos jurídicos en la codificación de la cláusula de responsabilidad civil en contratos inteligentes' (2025) **14** *Revista Chilena de Derecho y Tecnología* e74870 https://doi.org/10.5354/0719-2584.2025.74870.

31 Ajar Rab, 'Smart Contracts and Blockchain: The Panacea to the Unequal Bargaining Power of Consumers?' (2021) 8(1) International Journal *on Consumer Law and Practice* https://repository.nls.ac.in/ijclp/vol8/iss1/3 accessed **19 July 2026**.

Table 3.Platforms that provide dispute resolution service

Platform	Description
Kleros	Kleros is a decentralized platform that uses blockchain technology to provide a transparent and fair arbitration system. It uses a jury selection mechanism based on game theory, where jurors are randomly selected to evaluate evidence and make decisions in disputes related to smart contracts and other Digital agreements (Aouidef et al., 2021)
Aragon Court	Aragon Court is a platform designed to resolve disputes with in the ecosystem of Aragon's Decentralized Autonomous Organizations (DAOs). It provides a dispute resolution mechanism that allows DAO users to submit their disputes to a virtual panel of judges, who evaluate the evidence resented and make a collective decision based on pre-established community rules (Aouidef et al., 2021).
Jur	Jur is an innovative platform that offers arbitration services for commercial and legal contracts, leveraging blockcha in technology to ensure fairness and transparency in the dispute resolution process. It focuses on providing an efficient and accessible solution for commercial dispute resolution, facilitating an environment of trust for all parties involved (Schmitz and Rule, 2019).
OpenLaw	OpenLaw is a powerful tool that facilitates the creation, execution and Management of legally supported smart contracts. It allows users to easily integrate arbitration clauses into their digital contracts, providing a framework for automated, blockchain-based dispute resolution. Open Law stands out for its flexibility and for offering a user-friendly interface for Drafting digital legal contracts (Schmitz and Rule, 2019).
Mattereum	Mattereum is a platform that seeks to bridge the gap between physical and digital assets through smart contracts, providing a robust legal infrastructure for trading these assets. It includes an arbitration service to handle disputes, ensuring that agreements are executed fairly and efficiently. Mattereum is pioneering the use of Blockchain technology to create an automated and Decentralized legal system (The Mattereum Team,2018).

Source: Created by authors from Aouidef et al. 2021; Schmitz and Rule, 2019; The Mattereum Team 2018.[323334]

51 The use of the platforms works perfectly for consumers conducting transactions from long distances, but within commercial business with small amounts. However, not everything is perfect since, after analyzing the above, local laws present difficulties in terms of constitutional law, as well as altering the principles of due process, since elements of subjective nature are not present and the absence of such elements may result in unfair solutions.

E. How Codes of Arbitration Clauses Can Build Trust in Arbitration Clauses

52 Although arbitration as an alternative dispute resolution mechanism in SCs is gaining popularity, there are still problems related to the correct codification of contractual clauses, including the arbitration clause. It is also true that when problems arise in the execution (bugs in the code, for example), a complaint system has to be activated, which can be difficult due to the fact that blockchains involve people from different jurisdictions.

53 In this regard, some States have legislation on blockchains such as SCs, with minimum parameters for developing the arbitration clause, with the intention of achieving consistency with national legal systems and the rights of the contracting parties, especially consumers[35].

54 It is important to remember that although arbitration based on blockchain technology is often presented as an ideal mechanism for strengthening trust between parties operating in transnational environments, such trust cannot be derived exclusively from the standardization of the code

32 Yann Aouidef, Federico Ast and Bruno Deffains, 'Decentralized Justice: A Comparative Analysis of Blockchain Online Dispute Resolution Projects' (2021) 4 Frontiers in Blockchain 564551 https://doi.org/10.3389/fbloc.2021.564551.

33 Amy J Schmitz and Colin Rule, 'Online Dispute Resolution for Smart Contracts' (2019) Journal of Dispute Resolution 103.

34 The Mattereum Team, Mattereum Protocol: Turning Code into Law (2018) https://mattereum.com/wp-content/uploads/2020/02/mattereum-summary_white_paper.pdf accessed 19 July 2026.

35 A Legerén, 'Retos jurídicos que plantea la tecnología de la cadena de bloques' (2019) 6(1) Revista de Derecho Civil 177 https://dialnet.unirioja.es/servlet/articulo?codigo=6876649 accessed 19 July 2026.

that activates the arbitration clause. In fact, the trust generated by the code is primarily technical, based on the predictability of automatic execution and the resistance to manipulation offered by the cryptographic infrastructure. However, as Buchwald[36] points out, from a legal-arbitration perspective, that trust does not automatically acquire the character of substantive law, understood as the reasonable expectation that the procedure will respect essential guarantees such as due process, equality of the parties, and the possibility of ex post control of the award.

55 Most of the current studies from the legal field on SC focus on the security of codification and the theorization of these instruments based on national legislations. However, there are no studies from computer science and law that analyze the lack of user-friendly tools for users to design specialized contract codification templates, and in particular the arbitration clause, in order to provide legal certainty in case of disputes.

56 As a first step it should be taken into account that SCs dispense with state intervention and its regulation due to the delocalized nature that characterizes it[37], therefore it is desirable that "the online dispute resolution mechanism that is developed is aided by this same technology to provide simplicity to the process, but at the same time security both for the parties and for the shared data smart contracts"[38], thus providing the same dispute resolution mechanisms with flexibility and adaptability to emerging controversies[39]. It is worth mentioning that arbitration has moved too the virtual scenarios and acquired new forms, such is the case of *Online Dispute Resolution*[40], which is now intended to be carried out through codified language, characteristic of SCs[41].

57 Follow this logic, the agreements between the parties developed by means of the code sequence can be executed autonomously, it is necessary that the orders that the parties introduce in it abide by Boolean logic; or in other words, they must have the structure *if/then/else*: if this circumstance is fulfilled (*if*), then this action is executed (*then*); if it is not fulfilled, another action also foreseen (*else*) is executed[42]. Understanding this is of vital importance to know the ways of drafting and coding a possible arbitration clause or agreement.

58 Despite this structure, the codification of the clause may be questioned, since it is not a language that can be understood by everyone, and therefore, the code may not translate the true will to submit the process to arbitration, or it may not correctly articulate the agreements surrounding the arbitration process agreed by the parties, especially since it must be written in Boolean logic.[43], reiterates that as the possibility exists for parties to codify their contracts, new perspectives for efficient arbitration procedures emerge, however, national and international legal systems must begin to contemplate the possibilities of indicating the minimum requirements to develop the codification of the clause and even the arbitration procedure adopted by the parties. However, this is a challenge that implies that the regulation of the code requires a relatively new type of expertise, which will probably not be evenly distributed among national legislators.

59 The likely result, therefore, would be that some legislators would take action, while others would not be as prepared or able to adapt their implementing laws to this new technological reality.

60 Although the elaboration of the arbitration clause or agreement is traditionally based on the autonomy of the will of the parties, it is of vital importance, especially for consumers who adhere to adhesion SC, that there is a determination of the content of the code to avoid asymmetries and imbalances for lack of information or understanding of the contractual terms[44]. For these reasons, regulation at the

36 Michael Buchwald, 'Smart Contract Dispute Resolution: The Inescapable Flaws of Blockchain-Based Arbitration' (2020) 168(5) University of Pennsylvania Law Review 1369 https://scholarship.law.upenn.edu/penn_law_review/vol168/iss5/3.

37 María Yépez, I María Vela and Bernarda Haro, 'Smart contracts y arbitraje: hacia un modelo de justicia deslocalizado' (2020) 7(1) USFQ Law Review 1 https://doi.org/10.18272/ulr.v7i1.1698.

38 G Hierro Viéitez, 'Introducción al Blockchain, los contratos inteligentes y su relación con el arbitraje' (2021) 79 Revista de Derecho 299, 302 https://dialnet.unirioja.es/servlet/articulo?codigo=8328966 accessed 19 July 2026.

39 Sonia Sánchez and Faisury Rodríguez, 'Eficiencia del arbitraje en la garantía del derecho de acceso a la administración de justicia' (2021) 16(2) Via Inveniendi et Iudicandi 277 https://doi.org/10.15332/19090528/6251.

40 EA Calderón Marenco and J Torres Buelvas, 'La autonomía de la voluntad y los contratos wraps' in C Reyes Díaz, R Martínez Regino and HA Ramírez Medina (eds), Derechos Humanos y Políticas Públicas ante los nuevos retos sociales en México (Universidad Nacional Autónoma de México 2021) http://derecho.posgrado.unam.mx/site_cpd/public/publis_cpd/derechoshumanos.pdf accessed 19 July 2026.

41 Ignacio Tasende, 'Blockchain y arbitraje: un nuevo enfoque en la resolución de disputas. Especial énfasis en smart contracts y criptodivisas' (2020) 22 Revista de Derecho 138 https://www.scielo.edu.uy/pdf/rd/n22/2393-6193-rd-22-138.pdf accessed 19 July 2026.

42 Kristel G Loayza, Legal Analysis of Smart Contracts Based on Blockchain Technology in Business-to-Consumer (B2C) E-commerce (thesis, Universidad Católica San Pablo 2019).

43 Pietro Ortolani, 'The Impact of Blockchain Technologies and Smart Contracts on Dispute Resolution: Arbitration and Court Litigation at the Crossroads' (2019) 24(2) Uniform Law Review 430 https://doi.org/10.1093/ulr/unz017.

44 C Correa Martínez, 'La autonomía de la voluntad y el análisis económico del derecho desde la perspectiva de Richard Posner' in JM Álvarez Zárate and K Fach Gómez (eds), Apuntes prácticos sobre la autonomía de la voluntad como punto de conexión en el derecho económico internacional (Universidad Externado de Colombia and Universidad de Zaragoza 2021).

transnational level seems particularly desirable, in order to ensure efficiency while avoiding excessive restrictions on technological developments. A significant first step could be the development of soft law guidelines that identify a list of practices that codifiers should avoid[45].

61 Elements for the design of the arbitration clause or agreement

62 It should be recalled that the objective of the comparison between Smart Contracts and distributed systems is to implement a scalable and economical solution with high performance in the use of software resources, using cluster systems and development languages for the development of *per-to-per* systems as a means of process integration. The sharing of resources that a computer system has allows the use of such resources in a reciprocal way and allows access to them at any time, i.e., the physical and logical infrastructure of a computer currently allows infrastructures that must be oversized for normal operation in the sharing of technological resources. However, for this to occur, the normal operation of the technology must be guaranteed and, in the event of over-saturation, the secondary surplus must be available due to the oversizing of said technology.

63 This oversizing leads to the scenario of computational architecture, which takes an evolving context associated with processors and *multi-core-architectures*. In addition to this, the implementation of networks articulated with the architecture allows the distributed enablement of processes and the increase in the processing power of packet in formation in several computers and subsequently returning it to a single packet without the need for the accompaniment or control of the internal clock of the machine. This will allow *Smart Contracts* to be accessed in real time without altering the structure of the contract depending on the user who is executing it. Which allows the arbitration clause to generate results in real time for each of the processes required regarding the use of smart contracts and the completion of these by the users who access it[46].

64 In this sense, this concept is also applied to high-speed networks, they are the most viable alternative solution for sharing resources, since they are presented as an economical and scalable solution, allowing collaboration and optimization of these networks and thus the contracts that are processed will deliver a much more truthful result in an reduced time. This can materialize as long as at both ends there are devices that allow the interaction of users on smart contracts, directly guaranteeing the sharing of there source and the diligence of the contract execution to obtain the final result expected from *Smart Contracts.* This will be achieved by applying standards that guarantee the interoperability of the different participants or called users that require a definitive and systematic solution generated from the *Smart Contract.*

65 The existence of recognized standards from the development of *Ethereum* as in the design of *Smart Contracts,* and the integration that they must have guarantee the development of particular *software* to achieve interoperability of the different possible solutions of these softwares and their scalability.

66 It is necessary to ensure that there is a balance between what the smart contract offers to the world and what it can offer legally, from the answers and decisions that can be programmed in the smart contract from the proposed arbitration clause. The design of the *Smart Contract* proposes the realization of diagrams on the arbitration clause that allow the interaction between the programmers and the users that will work on the *software* in which the arbitration clause will be designed, achieved through the UML language (unified modelling language), which allows the development of software from objects.

67 The UML modelling language is one of the most widely used standards to documentary system accurately. In this context, the user interface can be designed with the requirements of the Smart Contract applied to the arbitration clause, creating profiles that are a mechanism to understand the syntax and semantics in the design of the software, and the language used by the contract or to express the specific concepts to be implemented in the smart contract. Based on graphs, a system is modelled that allows the modelling of SC sand understanding the work of the arbitration clause in between the users and the arbitrators[47].

68 Now, the development of UML diagrams can generate an *ad-hoc* application that in the context of the SC will allow the correspondence with the concepts that the arbitration clause will have, managing to respect the standard UML model applied to the concepts worked by the MOF (Meta-Object Facility). After the design has been carried out on the unified modelling language or UML, the people directly involved in the design of the *Smart Contract* (programmers and users) will verify the elaboration and consistency of the *items* placed in the arbitration clause and the final result that the contract will generate. Thus, the multidisciplinary work performed by developers and

45 ibid. p. 20.

46 Harold Castro, 'Grid Computing: Promesas de los Sistemas Distribuidos' (2006) Sistemas 98, 47–55 https://acis.org.co/portal/Revista/98/4_Uno.pdf.

47 Lidia Fuentes-Fernández and Antonio Vallecillo-Moreno, 'An Introduction to UML Profiles' (2004) 5(2) *UPGRADE: The European Journal for the Informatics Professional* 6–13.

programmers, are articulated with the legal review that starts from a Ricardian contract, or only the will of the parties, which will integrate all the elements participating in the execution of the arbitration clause applied to the smart contract. With the component of semantic analysis, the results and the generation of conclusions will be much safer to understand and from the content of the agreement that rules the contract

69 Thus, the result that can generate a smart contract is articulated with *peer-to-peer* systems, firm and distributed, a stable network with the property of resolving internal conflicts automatically concerning events, predictable and secure will be able to deliver much more structured and understandable results for users who make use of the tool based on the arbitration clause[48].

70 To conclude, it should be noted that, for the coding of the arbitration clause, a programming language is required, which depends on the specific context in which the coding is to be performed. However, the most widely used programming language for arbitration clause encoding is the *Solidity* software running on the *Ethereum* platform (which develops a virtual machine called EVM [Ethereum Virtual Machine]),which has the ability to run block chain software. In the virtual machine, smart contract codes are processed which guarantees accurate and reliable results, allowing any user to become a developer user and be able to access the blockchain[49].

71 It should be added that, the *Solidity* programming language defines a structure called *„Arbitration Clause",* which includes the following fields: *arbitrator:* the address of the arbitrator or arbitration court designated in the clause; *arbitration Fee*: the amount of *ether* (Ethereum's native cryptocurrency) that the parties agree to pay upfront to cover the arbitration costs; *deadline:* the deadline for initiating arbitration after a dispute arises; *dispute Resolution:* a link to the detailed description of the dispute resolution process agreed by the parties; and *jurisdiction:* the applicable jurisdiction in case there is a dispute that cannot be resolved through arbitration.

72 Taking the above, and comparing the model arbitration clause offered by the Arbitration Rules of the United Nations Commission on International Trade Law[50]: "Any dispute, controversy or claim arising out of or relating to this contract, its breach, termination or invalidity, shall be settled by arbitration in accordance with the UNCITRAL Arbitration Rules"(31). If a dispute, controversy or claim arising out of or relating to this contract, its breach, termination or invalidity shall be settled by arbitration in accordance with the UNCITRAL Arbitration Rules"(31), it should be written in a Ricardian contract, following Boolean logic: "If a dispute, controversy or claim arising out of or relating to this intelligent contract shall be settled by the Tribunal (*arbitrator) within* 24 hours of its non-performance (*deadline*) in accordance with the UNCITRAL Arbitration Rules (*dispute Resolution).* Each party shall pay the amount of ... (as the parties may agree) *(arbitrationFee).* If the arbitral tribunal does not have jurisdiction, it shall be decided by the local *jurisdiction.*

73 However, there are other types of arbitration clauses offered by different arbitration centres with along track record in the field. The model clause provided by the American Arbitration Association[51] establishes the following model:

> Any controversy or claim arising out of or relating to this agreement, or the breach thereof, shall be settled by arbitration administered by the American Arbitration Association in accordance with its Commercial Arbitration Rules, and judgment on the award rendered by the arbitrator(s) may be entered in any court having jurisdiction thereof.

74 From the above model a model Boolean logic clause may be suggested: If (controversy or claim arising out of or relating to this contract) then, Resolve by arbitration administered by the American Arbitration Association in accordance with its Commercial Arbitration Rules. Register judgment on the award rendered by the arbitrator(s) in any court having jurisdiction thereof.

48 Juan P Valencia, 'Derecho, tecnología e innovación blockchain y contratos inteligentes' (2020) 8(16) *Revista de Investigación en Tecnologías de la Información* 46–56 https://dialnet.unirioja.es/servlet/articulo?codigo=7576622.

49 J Romero Solís, *Aplicaciones de Contratos Inteligentes en Ethereum* (thesis, Universidad Carlos III de Madrid 2019) https://e-archivo.uc3m.es/handle/10016/29653.

50 UNCITRAL, *UNCITRAL Arbitration Rules* (2021) https://uncitral.un.org/sites/uncitral.un.org/files/media-documents/uncitral/es/21-07999_ebook_s.pdf.

51 American Arbitration Association, 'Standard Arbitration Clause' (2024) https://www.adr.org/Clauses accessed 27 August 2024.

75 The London Centre for International Arbitration (LCIA) provides the following model:

> Any dispute arising out of or in connection with this Agreement, including any question as to its existence, validity or termination, shall be referred to and finally resolved by arbitration under the LCIA Rules, which redeemed to be incorporated by reference in this clause.

76 In view of the above, the following model is proposed: The number of referees shall be [one/three].

> The seat, or legal place, of arbitration shall be[City and/or Country].
>
> The language to be used in the arbitration proceedings shall be [ ].
>
> The applicable law of the contract shall be the substantive law of[ ].
>
> If(dispute arising out of or in connection with this contract, including any question as to its existence, validity or termination) then
>
> Resolve by arbitration under the LCIA Rules, which are deemed to be incorporated by reference into this clause.
>
> If(number of arbitrators is one) then Select a referee.
>
> If not,
>
> If(number of arbitrators is three) then Select three referees.
>
> Establish the seat of arbitration in [City and/or Country]. Determine the language to be used in arbitration proceedings. Apply the substantive law of the contract as specified.
>
> Finally, the International Chamber of Commerce[52] provides the model arbitration clause: "All disputes arising out of or in connection with this contract shall be finally settled in accordance with the Rules of Arbitration of the International Chamber of Commerce by one or more arbitrators appointed in accordance with these Rules".
>
> If(disputes arising out of or in connection with this agreement) then
>
> Make a final decision in accordance with the Rules of Arbitration of the International Chamber of Commerce.
>
> Appoint one or more arbitrators in accordance with these Rules.

F. Conclusions

77 The article concludes that, despite the automatic nature and the promise of reducing contractual non-compliance that characterize *Smart Contracts*, they are not exempt from generating conflicts between the parties, highlighting the importance of arbitration as the ideal mechanism for the resolution of such disputes. The need to include properly codified arbitration clauses or agreements in *Smart Contracts* is justified, facing the challenge of ensuring that such codification faithfully reflects the will of the parties and maintains its legal validity. This is especially relevant in the absence of a parallel contract drafted in common language (Ricardian contract), which may compromise the effectiveness of the arbitration clause.

78 The research emphasizes the need to standardize the coding of the arbitration clause, using as a reference the Unified Modelling Language (UML) standards to design user interfaces that accurately reflect the requirement soft the SCs and, therefore, facilitate the understanding and execution of the arbitration clauses by the endusers. This involves a multidisciplinary process that integrates the technical expertise of software developers with legal expertise, ensuring that contractual terms are clear, fair and enforceable, minimizing informational asymmetries and contractual imbalances, especially for consumers adhering to adhesion SCs.

79 Finally, a codified arbitration clause model is proposed that contemplates critical aspects such as the appointment of the arbitrator, arbitration costs, time limits for initiating arbitration proceedings, description of the process and applicable jurisdiction. This model is based on Boolean logic to ensure automatic enforcement under predefined conditions, providing a framework for the effective implementation of arbitration clauses in the *Smart Contracts* environment. Standardization and transnational regulation are presented as imperatives to ensure the effectiveness of *Smart Contracts* in arbitration, avoiding excessive restrictions that may inhibit technological innovation.

Eduardo Andrés Calderón-Marenco
Universidad Cooperativa de Colombia
eduardo.calderon@campusucc.edu.co

Tania Rodríguez Palacios

52 International Chamber of Commerce, 'Arbitration Clauses' (ICC Spain 2024) https://www.iccspain.org/wp-content/uploads/2024/02/Arbitration_clause_SPA.pdf accessed 27 August 2024.

Universidad Americana
tania.rodriguez@uam.edu.ni

Jaime Torres Buelvas
Universidad Cooperativa de Colombia
jaime.torresb@campusucc.edu.co

Piedad Barreto-Granada
Universidad Cooperativa de Colombia
piedad.barreto@campusucc.edu.co

Social media bans and the Digital Services Act: A critical analysis of Statement of Reasons and Online Dispute Resolution

by **A.R. Lodder and J. Zeleznikow ***

Abstract: The Digital Services Act introduces specific rules regarding social media bans in the context of content moderation, focusing on both what is known as the Statement of Reasons and user redress mechanisms. The Statement of Reasons should provide an elaborated explanation for platform decisions regarding content moderation. In this paper, we critically analyse to what extent the Digital Services Act (DSA) provides users with the necessary tools to effectively challenge a ban.

Keywords: social media bans, negotiation, Online Dispute Resolution

© 2026 A.R. Lodder and J. Zeleznikow

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: A.R. Lodder and J. Zeleznikow, Social media bans and the Digital Services Act: A critical analysis of Statement of Reasons and Online Dispute Resolution, 17 (2026) JIPITEC 192 para 1.

A. Introduction

1 Social media platforms such as Facebook, Twitter, and TikTok serve as essential communication channels for individuals and businesses. Consequently, being banned from these platforms can have significant repercussions, including damage to reputation, lost friendships, and financial loss. In the fall of 2021 at a conference in Porto, Portugal, Lodder & Zeleznikow explored how Online Dispute Resolution (ODR) might be used to address social media bans.[1] Our proposal focused firstly on requiring platforms to provide a reason when banning users from their services. At the time, we observed that users were often excluded from social media platforms without any explanation. This lack of transparency was exemplified by a common clause in many terms of service stating that access could be terminated "for any reason or no reason at all" (Benoliel & Becher 2022). Secondly, we argued that once users know why the platform decided to ban them, they should be provided with an ODR procedure to quickly resolve disputes about social media bans. From a legal and societal point of view, we consider it crucial that users understand the reasons behind a temporary or permanent suspension of their accounts. Also, they should be offered means to contest such a ban.

2 Since our initial discussion, the Digital Services Act (DSA) was accepted by the European Parliament on 5 July 2022,[2] and is fully applicable since 17

* Arno R. Lodder, ALTI - Amsterdam Law & Technology Institute, Vrije Universiteit, a.r.lodder@vu.nl
John Zeleznikow, School of Information Systems, Victoria University, John.Zeleznikow@vu.edu.au

1 Arno R Lodder and John Zeleznikow, 'How Online Dispute Resolution Can Be Used to Negotiate Social Media Bans' (ODR Conference, Porto, 2021).

2 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1 ('DSA').

February 2024.[3] The DSA introduces specific rules regarding social media bans in the context of content moderation, focusing on both what is known as the Statement of Reasons and user redress mechanisms. The Statement of Reasons should provide an elaborated explanation for platform decisions regarding content moderation. Interestingly, at the time of our presentation in Porto, the proposed DSA (Savin 2021) covered in that Article on Statement of Reasons only measures related to the removal or disabling of access to information. In the final version, however, the scope was expanded to include, among other things, the suspension and termination of user accounts. In this paper, we assess to what extent the Digital Services Act (DSA) provides users with the necessary tools to effectively challenge a social media ban.

3 Two decades ago, the concept of being banned from a virtual world was a relatively new phenomenon. At its core, it meant a digital exile, a loss of access to a community and a persona, either for a set period or indefinitely (Castronova 2004). The social media bans examined in this paper are comparable, as they all involve the temporary or permanent loss of access to a user's social media account. Our analysis excludes other restrictions, such as demonetization or ad-account limitations, focusing instead on the (indefinite) suspension of account availability.

4 The remainder of the paper is structured as follows. First, we provide context with a discussion of content moderation against the background of freedom of speech, the liability exemptions and ODR. We continue with an analysis of the Statement of Reasons (Article 17 DSA). Subsequently, we discuss the internal complaint-handling system (Article 20 DSA) and Out-of-Court settlement (Article 21 DSA). We conclude with a critical reflection.

B. Freedom of Speech, DSA Liability Exemptions and ODR

5 Before discussing Articles 17, 20 and 21 DSA, we sketch some background in the context of social media in three areas that are relevant for this paper: Freedom of Speech, the DSA liability exemptions, and Online Dispute Resolution. This is not meant as a thorough overview of any of those topics, but covers the essentials.

I. Freedom of Speech

6 Most tech companies, including social media platforms, are US companies, with the notable exception of China's TikTok. This American background obviously shapes the activities of those platforms, deeply rooted in the First Amendment tradition (Wu 2018). An early example, from before the days of social media, is the Yahoo France case, where Yahoo did not want to take down content violating French law, viz. the selling of Nazi paraphernalia. Their rigid approach not to comply with local law changed soon afterwards. In *Who Controls the Internet?* (2006), Goldsmith & Wu noted the total turnaround of Yahoo compared to its earlier insistence on First Amendment protections in the French paraphernalia case, when, some years later, Yahoo acknowledged – unlike before – the importance of adhering to local law, in this case Chinese censorship law.

7 These days social media companies often comply with national laws. Also, they moved from being passive carriers of content to active gatekeepers of public discourse. Sometimes they are even considered authorities of their own, or as De Keulenaar *et al.* (2023) call Twitter in the context of the discourse on COVID: accidental authorities. Still, at their core, social media platforms prefer to stay away from content moderation, often hiding behind freedom of speech arguments. Despite calls for US legislation that would "narrowly dictate First Amendment protection for users of social media" (Everett 2018), no such federal framework has yet been enacted.

8 Freedom of speech, as defined in the First Amendment to the US Constitution, protects citizens from government censorship. Social media platforms are not governments but privately owned entities. They are not obliged to serve everybody. As businesses, they have the authority to establish their own policies and determine the boundaries of permissible online speech, and allow users and uploads based on that policy.

9 Over the past decade, social media platforms have increasingly engaged in content moderation. Events such as Brexit and the election of Donald Trump as US president in 2016 prompted debates on whether social media companies bore responsibility for certain societal developments and what actions they could take to prevent similar occurrences in the future. In response, social media platforms adopted

3 European Commission, 'The Digital Services Act Package' (*Shaping Europe's Digital Future*) <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package> accessed 13 September 2026.

a *Code of Practice on Disinformation*.[4] The COVID-19 pandemic further intensified content moderation efforts, particularly concerning the removal of so-called disinformation. In 2022 the *Strengthened Code of Practice on Disinformation*[5] appeared and on 13 February 2025 the Commission and the European Board for Digital Services endorsed the official integration of the voluntary Code of Practice on Disinformation into the framework of the DSA (Eskens 2025).[6] The balance between free speech and platform moderation remains a widely discussed and complex issue (Riemer & Peter, 2021; Ali *et al.*, 2021).[7]

10 In Europe, the EU Charter of Fundamental Rights and the European Convention on Human Rights impose a positive obligation on states to ensure effective enjoyment of free speech rights. In the context of social media, this implies that citizens should not be arbitrarily silenced (Enarsson 2024). However, social media platforms retain the right to moderate content based on their own policies. While the broader implications of free speech on social media are beyond the scope of this paper (Brannon 2019), we focus specifically on the consequences of content moderation, namely, user bans, and explore potential dispute resolution mechanisms to address these challenges. For many people and companies, social media platforms such as Facebook, Twitter, and TikTok are important communication channels (Barroso & Van Brussel Barroso 2023). Being banned from a social media platform can have a major impact, as noted, in terms of reputation, relationships and even financially. Also, it deprives people of a means to express themselves.

11 Freedom of speech is fundamentally about the right to express one's opinion, not about the specific platform or location from which that expression takes place. In the DSA freedom of expression is mentioned several times, mostly in the context of risk assessment (Eder 2024). For instance, under Articles 34 and 35, Very Large Online Platforms (VLOPs) must assess how their moderation practices impact freedom of expression as a "systemic risk." Other notable mentions are about a safe and trustworthy online environment that allows users to "exercise their fundamental rights (...) in particular the freedom of expression and of information" (Recital 3), notice and action mechanisms (recitals 52 and 53) and on misuse (recital 63). It is also mentioned in the context of taking down of content (recitals 22 and 51) and demonetisation (recital 54), but not in the context of social media bans (suspension of service or account)

12 This paper does not focus on content removal. Instead, it examines social media bans, namely resulting in users having temporarily or permanently no access to a social media account. One might argue that this does not infringe on freedom of speech, as users are merely excluded from a particular platform and can theoretically move to another to express their views. However, in practice, there are only a few dominant social media platforms, and most people are active on just those platforms. The significance of having access to these platforms cannot be overlooked.

II. The Digital Services Act, Notice and Action

13 The Digital Services Act (DSA), as its full name also indicates, amends Directive 2000/31/EC on e-commerce.[8] At its core, the DSA aims to create a more responsible digital environment by regulating platform operations (Savin 2021), ultimately striving toward the broader objective of a safer internet (Law 2024). Chapter III of the DSA is on *Due diligence obligations for a transparent and safe online environment*. Sections 2 and 3 of this Chapter are relevant for our paper: especially Article 17 on *Statement of Reasons*, Article 20 on *Internal complaint handling systems* and Article 21 on *Out of Court Dispute Settlement*.

14 EU Directive 2000/31/EC on e-commerce regulated ISP liability in Articles 12-14 (Article 15 is on no general obligation to monitor). Whereas the US Digital Millennium Copyright Act (DMCA), which the text of Articles 12–14 was copied almost verbatim from, had a notice-and-take-down procedure, the e-commerce directive did not. Times have been changing. The main idea 25 years ago was to exempt providers from liability to enable the growth of e-commerce. Now, the focus has shifted more to moderation, so active involvement (Van de Kerkhof

4 European Commission, '2018 Code of Practice on Disinformation' (*Shaping Europe's Digital Future*, 16 June 2022) <https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation> accessed 13 September 2026.

5 European Commission, '2022 Strengthened Code of Practice on Disinformation' (*Shaping Europe's Digital Future*, 16 June 2022) <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation> accessed 13 September 2026.

6 European Commission, 'Code of Conduct on Disinformation' (*Shaping Europe's Digital Future*) <https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation> accessed 13 September 2026.

7 Martin Husovec and others, 'Re: Hearing on "Europe's Threat to American Speech and Innovation"' (Letter to Rep Jim Jordan, 3 September 2025) 1 <https://husovec.eu/wp-content/uploads/2025/09/US-Academic-Letter-DSA-Censorship.pdf> accessed 16 March 2026.

8 European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC' COM (2020) 825 final.

2025). Reflecting this change in approach, Article 7 of the Digital Services Act introduces the so-called 'Good Samaritan' rule. It always has been problematic that providers who "in good faith and in a diligent manner, carry out voluntary (...) measures aimed at detecting, identifying and removing, or disabling access to, illegal content" lost their exemption from liability. Now intermediaries such as social media platforms no longer do so if they actively moderate content in good faith.

15 Another change is that providers have to enact an easy notice and take down system, called a notice and action mechanism (Article 16(1) DSA). The notices only concern illegal content, not content violating the terms of service. In this context the terms unlawful (illegal) and awful (harmful) are often used (Turillazzi *et al.* 2023), a distinction also relevant in the case of misinformation where the content not necessarily is unlawful. This notice and action is only indirectly relevant for social media bans and redress, the focus of this paper. Normally, a series of take downs of information can lead to a temporary or permanent ban. So, the user should be offered the option to object against notices, and fight take down of information based on those notices. Notifications should indicate this: under Article 16(2)(a) what is required is "a sufficiently substantiated explanation of the reasons why the individual or entity alleges the information in question to be illegal content". This is comparable to the ground and explanation required under Statements of Reason discussed in Section C. The information is helpful to the person who uploaded the information, for it makes clear the reason why the material is considered illegal. The notice is, under Article 16(3), "considered to give rise to actual knowledge or awareness for the purposes of Article 6". So, after a notice the provider can no longer be exempt from liability if it does not act and the information indeed appears to be illegal.

16 Article 16 only refers to the person who notified, not the person whose information is considered illegal. One might expect that the notice and action procedure would also include a notification to the one whose information it concerns, but this is covered in Article 17 on Statement of Reasons to be discussed in Section C. It makes sense not to inform the uploader about void notices. However, under circumstances, e.g. if there are a lot of unsubstantiated notices, it could be interesting for the uploader to be informed. Providers might do that voluntarily, but the DSA does not require this.

III. Online Dispute Resolution

17 Our focus is on one of the most severe consequences of moderation: social media bans. While such bans often relate to alleged disinformation, the question remains: why are users occasionally prohibited from accessing these platforms? A prominent example is the banning of Donald Trump from Twitter and Facebook following the January 6, 2021, attack on the U.S. Capitol. On January 8, 2021, Twitter announced Trump's permanent ban but did not clarify whether he had any recourse to challenge or appeal the decision.[9] Although we do not support Trump's claims regarding the 2020 U.S. presidential election, nor do we dispute the platforms' decisions to ban him, we argue that transparent justifications and an avenue for redress should be integral to such actions. ODR could serve as a mechanism for contesting bans.

18 In the first decade of this century, we considered the possibilities of ODR to be enormous. Over the years the take-up has been somewhat disappointing. Cortes & Lodder (2014) expected the EU ODR platform to create a momentum for ODR. As Netjes & Lodder (2019) indicated, at that time they were still waiting for the EU ODR platform to become a success. Five years later, in December 2024, the EU announced in Regulation 2024/3228 [10] the disbandment of the ODR platform due to its limited effectiveness, with operations ceasing as of 20 July 2025.

19 Despite the lack of success of the ODR platform, the European Union keeps on pushing the use of ODR in several legislative initiatives. Examples are the PSD2 Directive (Lodder 2021), Article 25 of Directive 2018/1972 (European Electronic Communications Code), and Article 33 of Regulation (EU) 2021/782 on rail passenger rights, etc. We began our work on social media bans and ODR based on the assumption that it would be a domain where ODR could be meaningfully used.

20 As Wetzstein and Leitner (2013) noted at the International Conference on e-Society, the combination of social media and ODR presents

9 Twitter Inc, 'Permanent Suspension of @realDonaldTrump' (*Twitter Blog*, 8 January 2021) <https://blog.twitter.com/en_us/topics/company/2020/suspension> accessed 13 September 2026.

10 Regulation (EU) 2024/3228 of the European Parliament and of the Council of 19 December 2024 repealing Regulation (EU) No 524/2013 and amending Regulations (EU) 2017/2394 and (EU) 2018/1724 with regard to the discontinuation of the European Online Dispute Resolution Platform [2024] OJ L, 2024/3228, 30.12.2024.

an innovative approach to resolving disputes. Similarly, Katsh and Rabinovich-Einy (2017) start their chapter on 'The Challenge of Social and Anti-Social Media' with a fictional example of a user being banned for using her everyday name on Facebook that is different from her official name. However, their chapter focuses on disputes between users. In contrast, our paper shifts attention to conflicts between users and platforms, in the context of bans and account removals.

C. Social Media Bans and Article 17 on Statement of Reasons

21 Article 17 is about the Statement of Reasons for the eventual take-down of an account. The drafted text of the Article appears somewhat unbalanced. In Article 15 of the Proposal of the DSA the Statement of Reasons only concerned the taking down of content. The Statement of Reasons in what has now become Article 17 of the DSA deals also with restrictions concerning demonetisation as well as the suspension or termination of either the service or the account. So, now the user should be informed in the Statement of Reasons about either removal of content and/or restrictions regarding account, service or monetization. This extension of the possible restrictions imposed on the user seems not fully reflected in the text of Article 17 (Husovec 2024).

I. When to Be Informed and About What? (Sub 1-2)

22 The initial Article 15 on statements of reason solely mentioned "remove or disable access to specific items of information provided by the recipients of the service". In Article 17(1)(a) this restriction is formulated as "any restrictions of the visibility of specific items of information provided by the recipient of the service, including removal of content, disabling access to content, or demoting content". Husovec (2024) warns "that applying Article 17(1)(a) too broadly can seriously strain platform development and design, while using it too restrictively can undermine the goal of the legislature—to ensure the explainability of decisions". It remains to be seen how the courts will strike this delicate balance in practice. What is new is the demonetization in Article 17(1)(b): "suspension, termination or other restriction of monetary payments". An example of such a measure is mentioned in recital 55 "The monetization via advertising revenue (...) can be restricted". Then there are two restrictions relevant for social media bans:

> suspension or termination of the provision of
>
> 1(c) the service in whole or in part,
>
> 1(d) the recipient of the service's account.

23 The recitals do not discuss the termination or suspension of accounts, but recital 64 does address the suspension of the service. The provider should send a prior warning before deciding on the suspension. That warning should include the reasons for the possible suspension and the means of redress. The actual suspension should be stated in a statement of reason. There is no similar consideration in the recitals about the account. A social media ban, as we discuss it, is related both to termination or suspension of account (prior warning is not discussed in the recitals) and the suspension or termination of the service. As discussed below, the warning is mentioned in Article 23 about suspension of the provision of the services in case a user frequently provides manifestly illegal content.

24 In case the provider decides to impose a restriction, the recipient should be informed. Ideally, this notification is given before the restriction is imposed. Hence, one would expect that the user is informed immediately after the provider has decided on a restriction. Article 17(2) specifies "at the latest from the date that the restriction is imposed, regardless of why or how it was imposed". It is likely that the decision and execution of the restriction will coincide, but whether it does may depend on the severity of the violation. For instance, if the nature of the content allows, the uploader should have a chance to object before the information is taken down. In practice, however, the term 'at the latest' most likely refers to the actual moment of communication, implying that the user is only able to respond once the content has already been removed. In the DSA proposal, it was indicated that the user should be informed 'at the latest at the time of the removal or disabling of access', because the restrictions on monetization, account and service were not included.

25 It seems that the final text of Article 17(2) has not been adequately adapted in order to take into account the three new restrictions. Hence, one would expect that the timing of those new restrictions is different from the mere taking down of content. Notably, social media bans and demonetization are generally imposed only following a sequence of prior events or violations. One would expect prior warnings. Article 23(1) provides details about the suspension of service provision in cases where users "frequently provide manifestly illegal content".

In Article 23 it is indicated that the suspension of the service takes place "after having issued a prior warning". This sequence of communications is not accounted for in Article 17. One could argue that the Statement of Reasons in case of a social media ban occurs after the warnings indicated in Article 23. Besides the fact that these warnings are not part of the actual Statement of Reasons cf. Article 17, it should also be noted that Article 23 does not concern violations of terms of service, but only cases in which content is considered illegal. Moreover, Article 23 addresses only suspension of service provision (sub c), and not termination of services (sub c), suspension/termination of account (sub d) or demonetisation (sub b). So, it covers only one part of the three new restrictions. In sum, this part of Article 17 is ill-drafted.

II. Clear and Specific Statement (Sub 3(a))

26 Article 17 requires providers to give a clear and specific Statement of Reasons for their decision. This means the user must be properly informed, rather than only receiving a vague reference to the terms, a practice that is all too common. As Rocha (2026) observes, the Statement of Reasons resembles, in important respects, requirements imposed on public authorities under principles of the rule of law and due process. However, this analogy should not be overstated. While Article 17 aims to enhance transparency and accountability, the quality and clarity of the reasons provided remain crucial. Where statements of reasons are overly generic or insufficiently precise, they may fail to meaningfully inform users and thereby undermine their effectiveness. An interesting suggestion for improvement is described by Aspromonte *et al.* (2024), who suggest using LLMs to link Statement of Reasons to specific clauses in the Terms of Service. However, given the abstract nature of the terms there could still be many situations in which this explanation does not help that much.

27 Article 17(3) sub a-f describes what should be included in the Statement of Reasons. As there are four restrictions mentioned in Article 17(1), one would expect these same four to be addressed. The phrasing of the Article, however, echoes the original language of the DSA proposal, hence its sole restriction concerning the removal of content.

28 Sub (3)(a) mentions to some extent all four restrictions. It clearly covers visibility (Article 17(1)(a)) and monetary repercussions (Article 17(1)(b)). The other two are particularly relevant in the context of social media bans, hence restrictions regarding termination/suspension of the service or the account. There is no specific reference to taking such action in Article 17(3)(a), but the phrase that refers to these other two restrictions of Article 17(1) reads "or imposes other measures referred to in paragraph 1 with regard to the information". This lack of clarity, similar to what occurred with Article 17(1), was likely caused by amendments made to the Article after the proposed DSA. As noted, only in the final text were the possibilities of suspending or terminating either accounts or services explicitly included. Obviously, a decision regarding a social media ban is not primarily linked to the information being taken down, so Article 17(3)(a) does not accurately reflect what should be communicated regarding, as we call them, social media bans. Rather, normally there will be a series of events that led to the decision to suspend an account or service.

29 Similarly, in the context of the Transparency database (Article 24(5) DSA), recital 66 refers to "the decisions and statements of reasons of the providers of online platforms when they remove or otherwise restrict availability of and access to information." While the Transparency Database has failed to fully deliver on its initial promise (Kaushal *et al.* 2024, Leijten 2026), its current form arguably mirrors the 'Statement of Reasons' found in the DSA proposal, rather than the final text which includes e.g. social media bans. Although account-level actions do restrict access to information, this is not made explicit.

III. Intermezzo: Visibility

30 It is interesting to note that Article 17 contains a logical inconsistency regarding visibility. Article 17(1) refers to any restrictions of the visibility of specific items of information provided by the recipient of the service, *including* removal of content, disabling access to content, or demoting content. So, visibility is the general category, and examples of measures concerning visibility are "removal of content, disabling access to content, or demoting content". In sub (3)(a) visibility is no longer the general category, but just one of four possible measures: "the removal of, the disabling of access to, the demotion of *or the restriction of the visibility of the information*". In the DSA proposal the statements of reasons were restricted in Article 15(1) and 15(2)(a) to just "removal of, or the disabling of access". So, demotion and visibility have been added.

31 Clearly, the DSA designers did not clearly figure out the status of visibility as either the general overarching category or just one of the four categories. The definition of content moderation has yet another ordering. The effects of content moderation are generally described as concerning availability, visibility, and accessibility, so, next to visibility, accessibility and availability are also mentioned. The examples of this trio used in the

definition of content moderation are "demotion, demonetisation, disabling of access to, or removal thereof". So, here, suddenly demonetisation becomes one of the examples of visibility. While Article 17(1) lists demonetisation as one of four distinct categories of restrictions, separate from visibility, the definition of moderation includes demonetisation merely as an example of visibility. This categorisation of demonetisation seems incorrect. One might argue that demonetisation can have indirect effects on visibility, or even to a lesser extent on availability or accessibility, but as an economic measure it is conceptually distinct from both. Therefore, treating it as a separate category, as is done in Article 17(1), seems more appropriate.

IV. Facts and Circumstances, and Use of Automated Means (Sub 3(b-c))

32 Most content is likely removed automatically, and knowing this may help users in contesting such decisions. We do not expect, however, that the quality of justifications provided has improved since the human rights analysis conducted by Dias Oliva (2020). Understanding this systemic limitation may help users seeking to contest these decisions. For the person who is facing a social media ban it does not seem particularly valuable to know whether automated means were used in making the decision (sub c). The facts and circumstances relied on in taking the decision (sub b), could, however, entail crucial information. The social media platform should explain what behavior of the user caused the decision to either suspend (or terminate) the account or service. This information is essential to fight a social media ban. Already in the DSA proposal, the same was stated regarding the removal of information, and it was added that information should be included about whether the decision was taken pursuant to a notice. In the final version of the DSA, "where relevant" has been added, which will obviously never be the case with social media bans, since notices always concern specific information posted.

V. V. The Ground and Explanation (Sub 3(d-e))

33 The grounds for the decision and the justification are solely relevant for restrictions on visibility (sub 1(a)). In sub d it is indicated that in case of illegal information the legal grounds should be included, and explanations given why the information is illegal. In sub e it is stated in a similar fashion that any incompatibility of the terms should be specified, with reference to specific terms and why they have been violated. For visibility this is likely the most relevant information. For the other three restrictions, viz. regarding demonetization, services and accounts, a reference to the terms of service could be informative for the recipient. Compared to the proposal of the DSA, the phrasing of the violation of either the terms of service or the law has not been amended in the DSA.

34 The final text cannot be interpreted in a way that includes the other three types of restrictions beyond visibility. It refers, respectively, to 'why the information is considered to be illegal content on that ground' and to 'explanations as to why the information is considered to be incompatible with that ground.' Both formulations clearly apply only to decisions affecting the visibility or availability of content, and not to other types of restrictions such as demonetization, account suspension, or service termination. This is at least the case for social media bans. One could argue that demonetisation does apply to information, cf. recital 55 "The monetisation via advertising revenue of information provided by the recipient of the service can be restricted by suspending or terminating the monetary payment or revenue associated to that information." Still, the explanation focuses on why the information is illegal or violates the terms, rather than on the reasons for the application of demonetisation (if addressed at all, it is only indirectly).

VI. Redress (Sub 3(f)) and Article 17(4): Clear, Clearer, ...

35 Finally, under Article 17(3)(f), the provider should refer to redress possibilities, such as internal complaint procedures, ADR and litigation. How this information should be communicated is formulated as "clear and user-friendly information on the possibilities for redress available to the recipient of the service in respect of the decision". The redress options should be communicated for all decisions, including the ones related to suspension of service or account: social media bans.

36 The information regarding redress should be not only clear, but also user-friendly. This seems an indication that the information cannot pass unnoticed by the user. Article 17(1) requires "a clear and specific Statement of Reasons". Apparently, "clear and user-friendly" and "clear and specific" were not considered sufficient to delineate how information regarding redress should be communicated. In Article 17(4) the clear and specific purpose of Article 17(1) is further elaborated upon as follows "clear and easily comprehensible and as precise and specific as reasonably possible under the given circumstances." One could argue that "given the circumstances" weakens the information requirement. Tech companies could use it as an argument to provide

less information, as Leijten (2026) argues that the open norm in Article 17(4) may dilute the specificity requirement stipulated in the provision's first paragraph. Although companies may attempt to do so, the following sentence seems to establish a clear minimum standard for what the information must contain: "The information shall, in particular, be such as to reasonably allow the recipient of the service concerned to effectively exercise the possibilities for redress". So, the communication has to, in particular according to the last phrase, contain at least information essential to allowing meaningful action against restrictions imposed. For social media bans, this means that users must be clearly informed about why the ban was imposed and what they can do to have it lifted. That being said, one can question why Article 17(4) is needed in the first place.

37 Finally, on top of all this abundance of clarity, Article 14 indicates that the terms of service should include the rules of procedure of the internal complaint handling system, to be discussed shortly. This information has to be provided in "clear, plain, intelligible, user-friendly and unambiguous language". So, this adds to clear, easily comprehensible, user-friendly, unambiguous and plain information and implies a high threshold obligation (Quintais, Appelman & Ó Fathaigh 2023).

D. Internal Complaint-Handling System, Article 20

38 The reference to redress is undeniably clear, but what does the DSA say about the complaint procedure? Article 20 of the DSA deals with the provision of an internal complaint-handling system. Article 20(1) requires an effective internal complaint-handling system, that enables complaints to "be lodged electronically and free of charge". The decisions are the same as those mentioned in Article 17(1), so it is not especially clear why they are repeated here. And they are worded slightly differently. The visibility is again, unlike in Article 17(1), mentioned as a separate category, not the overarching one: "decisions whether or not to remove or disable access to or restrict visibility of the information."

39 The provision on demonetization has been relocated, moving from subparagraph (b) in Article 17(1) to subparagraph (d) in Article 20(1). Also, it has been reworded and now is phrased as 'suspension, termination or other restriction of monetary payments.' What was previously referred to as 'the ability to monetize information' is now simply described as 'monetary payments.' For the purposes of this analysis, subparagraphs (b) and (c) are most relevant, as they address the suspension and termination of, respectively, the service and the user account. The reason for the shift of the demonetization clause from subparagraph (b) in Article 17(1) DSA to subparagraph (d) in Article 20(1) appears to be that, unlike the Article on statements of reasons in the proposed DSA, the Article on complaint handling already explicitly included references to social media bans—under subparagraph (b) for services and subparagraph (c) for accounts. So social media bans were already partially covered in the proposed DSA. Probably, while they thought about social media bans in the context of redress, only later did they realize that the redress only is meaningful if it is clear what exactly the reason has been to impose the sanction, viz. the social media ban.

40 The time given to start the procedure is at least six months, starting on the day the recipient is informed (Article 20(2)). This is quite a long period. It is likely most users will start a procedure immediately after the ban or at least shortly afterwards.

41 The complaint procedures should be easy to access, according to Article 20(3). Also, the procedure has to be "user-friendly and enable and facilitate the submission of sufficiently precise and adequately substantiated complaints." This provides the opportunity for users of social media to file a complaint in which they can clearly express why they consider the ban to be unjustified.

42 Article 20(4) defines some quality standards for the complaint handling procedure. It says that complaints should be handled in a "timely, non-discriminatory, diligent and non-arbitrary manner." In our opinion it is crucial that the procedure is carried out quickly. Due to the important role that social media accounts play for both individuals and businesses, any restriction or suspension should be subject to a resolution process that operates within hours, rather than days.

43 The requirement that complaints are handled in a 'timely' manner introduces a degree of flexibility for platforms. While it may be reasonable to allow smaller platforms up to 24 or even 48 hours to respond, 'timely' for large social media platforms should imply a much shorter timeframe, ideally, within a matter of hours. These platforms have the technical and human resources to act quickly, especially given that the majority of bans likely concern fraudulent or scam accounts. Such accounts rarely contest the decision, and if they do, the rejection of their complaint is not that complicated. The same is true for those users who have a serious objection. In case of doubt the platform could even decide to reinstate the account temporarily while further assessing the case. After all, the fact that there is doubt is an indication that the arguments for the ban are not overwhelming.

44 After this procedure, in case the complaint of the recipient is accepted because it contains information indicating that the complainant's conduct does not warrant the measure taken, the provider should "reverse its decision (...) without undue delay." It is important that the consequence should take effect immediately, which is basically what 'without undue delay' means.

45 Article 20(5) also requires the platform to communicate this decision quickly: "inform complainants without undue delay of their reasoned decision in respect of the information to which the complaint relates and of the possibility of out-of-court dispute settlement provided for in Article 21 and other available possibilities for redress." So, in case the decision is not beneficial to the user, (s)he should be referred to ADR/ODR.

46 The DSA also requires human intervention, that is, decisions solely based on automated means are not allowed (Article 20(6)): "are taken under the supervision of appropriately qualified staff, and not solely on the basis of automated means." It is tempting to use automated means such as AI. Human review in case of an appeal, in particular if it is against an automated decision, is a necessary safeguard. For smaller platforms this may be more problematic than for the big ones, although the number of decisions is likely lower as well. What has been said about the social media bans is also likely to be true for the other restrictions, including the taking down of content. Most decisions are clear and not contestable, but the small number that are not likely need serious attention. Humans have the capacity to assess the plausibility of claims against the decisions made by social media platforms.

E. Out of Court Dispute Settlement, Article 21

47 Article 21 refers to ADR, or as the EU calls it out-of-court dispute settlement. The bodies providing ADR, or likely and preferably ODR, should be impartial and independent; the procedure should be accessible through electronic means, and the dispute should be settled in a swift, efficient and cost-effective manner. This requirement is redundant in the light of other EU law, such as the ADR Directive. Article 21(9) DSA does refer to this Directive: "This Article is without prejudice to Directive 2013/11/EU and alternative dispute resolution procedures and entities for consumers established under that Directive."

48 The explicit reference in Article 21(3)(b) to expertise is: "it has the necessary expertise in relation to the issues arising in one or more particular areas of illegal content, or in relation to the application and enforcement of terms and conditions of one or more types of online platform[s], allowing the body to contribute effectively to the settlement of a dispute".

49 The quote does not reflect all restrictions, but instead mentions areas of illegal content and refers in general to terms and conditions. Social media bans are thus not explicitly referred to. Notably, unlike in the DSA proposal, there is a missing S: online platformS. What is more interesting is that this means that not every ODR provider can handle these claims. The provider needs to be familiar with, and have expertise in, disputes related to the taking down of material and social media bans.

50 We decided not to discuss further all nine subparagraphs, but to stick to some general observations. First, we believe the role of an appeal body does not fit ODR well. We believe it is too little, too late, when ODR is used as a procedure that starts once the internal complaint handling procedure fails. In our original thoughts about ODR and social media bans we envisioned the role for ODR to be that which the internal complaint handling procedure currently has. This is also, at its core, how the two most successful ODR applications to date, WIPO domain name arbitration and eBay dispute resolution, operated. We see a role for specialized ODR providers, as expressed very well in the just-mentioned Article 21(3)(b). These ODR providers could provide decisions immediately, or at a very short notice, as we discussed under the internal complaint handling procedures.

51 Second, the process described in Article 21 to appoint ODR bodies reminds us vividly of the failed ADR directive and ODR regulation (see above). This proposal in the DSA seems to be history repeating itself. Illustrative of this is that in July 2025, roughly eighteen months after the DSA became applicable, there were only six providers listed.[11] And the last addition of a provider was in December 2024.

F. Some Concluding Reflections

52 Our discussion of Article 21 on ODR was not very optimistic. Some authors are more optimistic. For instance, at the end of 2024 Holznagel said: "As an overall impression, Art. 21 DSA is seemingly delivering a good start."[12] Around the same time

11 European Commission, 'Out-of-Court Dispute Settlement Bodies under the Digital Services Act (DSA)' (*Shaping Europe's Digital Future*) <https://digital-strategy.ec.europa.eu/en/policies/dsa-out-court-dispute-settlement> accessed 13 September 2026.

12 D Holznagel, 'Art 21 DSA Has Come to Life' (*Verfassungsblog*, 5 November 2024) <https://verfassungsblog.de/art-21-dsa-fundamental-rights-certification/> accessed 13 September 2026.

Shipp *et al.* claimed "Article 21 DSA's out-of-court dispute settlement (ODS) solution is one unique tool to achieve these objectives [to improve the safety of EU internet users whilst protecting fundamental rights] with potentially far-reaching implications for users." The latter authors discuss both opportunities and challenges.[13] Ortolani (2023) is less optimistic: "some of the new provisions (e.g. concerning out-of-court dispute settlement) may prove too vague to have a tangible impact." Filatova-Bilous *et al.* (2024) criticize Article 21 DSA in a similar vein by claiming that it sets up the concept of out-of-court dispute settlement, but it does not give enough detail to understand how it would work in practice. Our sentiment largely reflects both the optimistic and pessimistic observations in the literature.

53 Obviously, the process should take place online. When being banned from a social media platform it does not make sense to fight the ban other than with online communication tools. First, because speed in communication is needed and that is what online tools provide. To give an extreme example, around 2022 Instagram mentioned in their terms they were required based on the "European Electronic Communications Code (Directive (EU) 2018/1972) (the "EECC") under the heading 'Information on handling disputes' that "If you are unhappy with our messaging services, you can raise a complaint by writing to us at[14]:

> Facebook Ireland Limited
> Attn: SERVICE COMPLAINTS
> (...) Ireland

54 In general, we believe in the value of ODR. After our initial discussion in 2021 on ODR and social media bans, the DSA has provided some laudable building blocks for fighting social media bans. The statements of reason in Article 17, although poorly drafted from the perspective of social media bans, provide valuable information requirements in the context of social media bans. After all, a social media ban can only be contested if it is clear why it was imposed. As a next step we would like to analyse empirical data, as vetted researchers under Article 40(4) DSA. That would give us the opportunity to analyse what exactly is communicated to the user in case of social media bans.

55 As a first warm-up exercise, on 4 December 2025 we selected only 'total termination of provision of service' for Facebook, Instagram, LinkedIn, TikTok, and X. We only received hits concerning X, just 159 cases. Then, we searched for the same platforms on 'total termination of account'. Much to our surprise, X had again 159 cases, which seems to suggest they are identical to the ones regarding the service. The number of cases we retrieved from other platforms were overwhelming, a little over 40 million: 40,126,801, distributed as follows:

- Instagram 39,377,603
- TikTok 499,112
- Facebook 249,927
- X 159
- LinkedIn 0

56 Without going into detail, we can conclude that for sure both X (159 social media bans in 6 months) and LinkedIn (not a single ban) are underreported. Another conclusion is that the termination or suspension of service was not used (or redundantly, as in the case of X), although users can have the service of Instagram and Facebook via the same Meta account. This seems to suggest the restriction regarding service is redundant. However, when you look at the most obvious platform in this context, YouTube (linked to a Google account), you find a little over 300,000 restrictions related to service and not a single one related to the account.

57 While digging a bit deeper into by far the largest group of social media bans by Instagram, we identified that 97% of the bans were detected and decided automatically. That seems quite a high percentage for social media bans. For termination of scam and fraud accounts (27 million bans related to this) the percentage of automated means was even higher (99.7%), but for those cases it makes sense. The other high numbers were related to both data protection and privacy violations (little over 5 million) and 'other terms' (little over 6 million). Several categories were never applied, of which cyber violence against women is an intriguing one because there were 26,615 terminations due to cyber violence. Obviously, among those there will be ones targeted at women. In order to conduct a proper analysis, we need more detailed data, which we hope to obtain as vetted researchers.

G. Conclusion

58 In our presentation in Fall 2021 in Porto we were hopeful that ODR could help social media platform users faced with a ban. We sketched a technical

13 J Shipp, R Harrison and A Curtis, 'Resolving Content Disputes Outside the Courtroom Using the Digital Services Act' (*Media@LSE*, 7 October 2024) <https://blogs.lse.ac.uk/medialse/2024/10/07/resolving-content-disputes-outside-the-courtroom-using-the-digital-services-act/> accessed 13 September 2026.

14 Instagram, 'Terms of Use' (*Instagram Help Centre*) <https://www.facebook.com/help/instagram/termsofuse> accessed 23 February 2022.

framework with different tools helpful in such a procedure, and emphasized the process should be quick and that the platforms should provide the reasons underlying their actions. Article 17 DSA forces social media platforms to communicate clearly about their actions to users, in case restrictions are imposed. This was needed, because the platforms took no or insufficient action on a voluntary basis. We also claimed that to fight a ban ODR techniques, preferably offered as a complaint handling procedure by a specialized third party, should be available. The DSA has an internal complaint handling procedure followed by ODR. We believe that this takes too much time and is therefore not efficient. Unless the internal complaint is dealt with almost instantly, the procedure is poor. But the DSA provides six months.

59 We discussed Articles 17, 20, and 21 of the Digital Services Act (DSA) which jointly establish the framework available to users who wish to challenge content moderation decisions, and relevant for this paper: social media bans. Article 17 is valuable, obliging platforms to provide clear statements of reasons for social media bans, allowing users to contest effectively the platform's decisions. Article 20 requires platforms to offer an internal complaint-handling system, but does not impose a sufficiently strict deadline for responses, which can undermine the effectiveness of the remedy, particularly because of the impact social media bans may have. We would prefer to see the internal complaint procedure replaced entirely by a swift external online dispute resolution (ODR) process. An impartial third-party mechanism could better guarantee fairness, neutrality, and timely resolution. In case internal complaint procedures remain, the DSA should require platforms to respond within a maximum of 24 hours. Article 21 is as yet too vague, but we still do believe ODR to be helpful for users facing a social media ban.

60 To assess whether our informed opinions reflect reality, we need access to detailed data on social media bans and the procedures through which users contest such decisions. In our next paper we hope to report findings based on data we obtain as vetted researchers under Article 40(4) DSA.

The Concept of Lawful Access and Its Implications for User Freedoms in the EU Copyright Acquis: Opinion of the European Copyright Society

by **Thomas Margoni; Martin Kretschmer; Axel Metzger; João Pedro Quintais; Tatiana Synodinou ***

Executive Summary: The advent of digital technologies and connected patterns of production and consumption has progressively displaced ownership in favour of access as the predominant mode of enjoyment of works of authorship and other information goods. This transition from acquiring physical copies to subscribing to digital services has become prevalent across various sectors and has even enabled new commercial, cultural and scientific practices. However, it also marks a major shift in the theory and function of copyright: it transforms a right primarily concerned with regulating certain exploitations and circulations of works into one capable of controlling individual acts of access and use.

EU copyright law increasingly reflects this shift through conditions of lawfulness. These appear across the acquis in different forms: the "lawful acquirer" in the Software Directive, the "lawful user" in the Database Directive, "lawful use" and "lawful source" in the Information Society Directive and CJEU case law, and "lawful access" in the Text and Data Mining (TDM) provisions of the Copyright in the Digital Single Market Directive (CDSMD). These expressions are not merely terminological variants. They perform a boundary-setting function by determining when exceptions and limitations may be relied upon, and therefore when certain acts may take place without right holders' authorization.

This function is especially important for text and data mining (TDM), machine learning and the training of AI models. Articles 3 and 4 CDSMD make lawful access a condition for the TDM exceptions, but do not define the concept. Recital 14 CDSMD refers to access based on open access policies, contractual arrangements, other lawful means and content freely available online. Developments under the AI Act and the General-Purpose AI Code of Practice further support the view that lawful access cannot be reduced to prior authorization for each individual work. It must also be assessed by reference to lawful conduct, due diligence and the relevant regulatory context.

The principal risk, surfaced in decisions such as ACI Adam or VOB, is that lawful access, lawful use or lawful source might be equated with right holders' authorization. An interpretation of this kind would undermine the nature and function of exceptions and limitations. Since an exception presupposes that authorization is not required, making authorization the sole condition of lawfulness would create circularity and deprive exceptions of practical effect.

At the same time, lawfulness cannot be understood as an unlimited permission to access or use protected works. It must operate within identifiable legal boundaries, including effective technological protection measures, valid and enforceable contractual restrictions where not overridden by mandatory law, and other legally binding limits. Its application must also be guided by contextual principles, including the meaning of content "freely available online", the knowledge and role of the user, the purpose and commercial or non-commercial character of the activity, the applicable fundamental rights framework, and the three-step test.

Accordingly, this Opinion seeks to provide normative and interpretative guidance for courts and legislators by developing a concept of lawfulness that is legally coherent, context-sensitive and capable of preserving the effectiveness of copyright exceptions. The analysis leads to the following five recommendations.

Recommendations:

1. Lawfulness cannot be reduced to right holders' authorization. This applies in particular to exceptions and limitations conditioned on lawful access, lawful use or lawful source. The text, structure and function of EU copyright law support this conclusion. Since an exception to an exclusive right presupposes that authorization is not required, making authorization the sole condition of lawfulness would create circularity and would risk rendering exceptions ineffective.

2. Beyond authorization by right holders, there are other clear cases in which lawfulness is undisputed. These include situations where the law authorizes the relevant activity, including through exceptions, limitations or other free uses under copyright law or other legal sources; situations that fall outside the scope of an exclusive right; and situations where

the relevant exclusive right has been exhausted.

3. Situations not expressly regulated by law require a context-specific assessment. That assessment must respect the hard boundaries of lawfulness, including effective technological protection measures, valid and enforceable contractual provisions or rights reservations, and other legally binding restrictions. It must also be guided by softer interpretative criteria, including whether content is freely available online, the user's knowledge, the nature and purpose of the activity, the relevant fundamental rights framework and, where applicable, the three-step test. The interpretation adopted must preserve the policy objectives and effectiveness of the rule at issue.

4. Scientific research deserves special treatment within this framework. Article 3 CDSMD reflects a deliberate legislative choice to privilege scientific research through a mandatory exception limited to specific beneficiaries and purposes. In light of the public-interest function of research and its connection to academic freedom under EU law, lawful access should be interpreted in a manner that preserves the effectiveness of that exception and avoids disproportionate source-verification obligations for research organisations and cultural heritage institutions.

5. Other special situations have been identified and further ones may emerge. These include general-purpose AI, open-source software and AI models, and other contexts in which access, use and lawfulness interact with public-interest objectives, open licensing practices or technological innovation. Such situations should be assessed under the framework proposed in this Opinion. As a general principle, where a new situation warrants treatment comparable to that granted to scientific research, the conditions of lawfulness should be interpreted in a manner that favours the relevant activity, subject to the applicable legal safeguards.

© 2026 Thomas Margoni; Martin Kretschmer; Axel Metzger; João Pedro Quintais; Tatiana Synodinou

The European Copyright Society (ECS) was founded in 2012 with the aim of creating a platform for critical and independent scholarly thinking on European Copyright Law and policy. Its members are scholars and academics from various countries of Europe, seeking to articulate and promote their views of the overall public interest on all topics in the field of authors rights, neighbouring rights and related matters. The ECS is neither funded nor instructed by any particular stakeholders. Its Opinions represent the independent views of a majority of ECS members.

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: Thomas Margoni, Martin Kretschmer, Axel Metzger, João Pedro Quintais and Tatiana Synodinou, The Concept of Lawful Access and Its Implications for User Freedoms in the EU Copyright Acquis: Opinion of the European Copyright Society, 17 (2026) JIPITEC 203 para 1.

A. Introduction

1 With the advent of digital technologies and ensuing patterns of production and consumption, access has progressively substituted ownership as the default mode for enjoying information goods and creative works. This transition from acquiring physical copies to subscribing to digital services has become prevalent in many sectors and has even given rise to new forms of business models and creative practices.

2 Within the EU copyright *acquis*, this shift from ownership to access represents a major change in the theory and function of copyright. It transforms a right originally conceived to regulate copies of works into a form of regulation of individualised uses of works with negative impact on individual freedoms, authors autonomy and creative agency. It is undisputable, however, that the evolution of EU copyright law has largely, and perhaps unconsciously, embraced this shift, as shown by the increasing use of concepts such as lawful use or lawful access to condition the validity of copyright exceptions or limitations.

3 The observation that access has emerged as the new regulatory paradigm for information markets and cultural good demands a deeper reflection. The way in which access is defined or qualified carries profound legal and policy implications. Within the European *acquis*, access has been increasingly subjected to a qualification of "lawfulness." Although seemingly reasonable and logical, the precise meaning of this qualification may have far-reaching consequences.

4 The recent emergence of the "lawful access" concept in the Copyright in the Digital Single Market Directive (CDSMD[1]) is illustrative. While operating as a key condition for the text and data mining (TDM) exceptions, lawful access does not receive an unequivocal definition in the Directive. Nor does the case law developed by the Court of Justice of the European Union (CJEU or the Court) offer a stable or uniform conceptualization of the term.

5 This uncertainty is striking. The concept now plays a central role in regulating access to online content for TDM purposes and, by extension, for machine learning and the training of artificial intelligence (AI) models. We argue that the concept of "lawful access" has been elevated to a sort of (loosely defined) regulatory interface between the protection of private (e.g., intellectual property rights, private economic entitlements) and public (e.g., fundamental rights and freedoms, market regulation, cultural and educational policies) interests. The concept therefore performs an important boundary-setting function. It seeks to enable legitimate (re)uses and forms of innovation while also preventing the misappropriation of human creativity. While the notion of lawfulness is present in many other contexts, including Article 5(3)(d) of the Information Society Directive (ISD)[2] or Article 2 of Regulation (EU) 2017/1563,[3] our focus is on creativity and technological innovation.

6 Accordingly, this opinion explores the theoretical underpinnings and the practical implications of allocating such a crucial public-private regulatory function to the underdeveloped concept of lawfulness in access to and use of creative content. It seeks to provide normative and interpretative guidance on the proper conceptualization of this expression by developing a notion of lawful access and use informed by legal, policy and technological principles. The objective is to offer courts and legislators a framework for interpreting "conditions of lawfulness" when future litigation or legislative reform arise.

B. The Historical Development of Lawfulness

7 This section offers a succinct chronological overview of the different appearances of the concept of "lawfulness" in EU law in relation to the acquisition, use or access of works.

I. The Software Directive: The Lawful Acquirer

8 The first use of the concept in the form of "lawful acquirer" can be found in the Computer Programs Directive of 1991 (Software Directive),[4] the oldest directive in European copyright law. This directive applied copyright for the first time to the then novel challenge of digitization. One of the most significant decisions was to include the acts of reproduction

1 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 (CDSMD).

2 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 (ISD).

3 Regulation (EU) 2017/1563 of the European Parliament and of the Council of 13 September 2017 on the cross-border exchange between the Union and third countries of accessible format copies of certain works and other subject matter protected by copyright and related rights for the benefit of persons who are blind, visually impaired or otherwise print-disabled.

4 Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version).

and adaptation necessary for the intended use of computer programs within the scope of copyright protection in Article 4 (a) and (b). This provided software producers with a legal basis for linking the use of programs by end-users to the conclusion of a licence agreement. This extension of exclusive rights was at least partially offset by the limitation provision in Article 5, which granted certain, albeit rather limited, minimum rights to the "lawful acquirer" of a copy of the program.

9 According to the Directive, the "lawful acquirer" shall not require authorization from the right holder for the acts referred to in Article 4 (a) and (b) where they are necessary for the use of the computer program in accordance with its intended purpose, including for error correction. These minimum rights apply irrespective of whether the "lawful acquirer" has concluded a licence agreement with the right holder. However, Article 5(1) does not prohibit end-user licence agreements since it starts with the words "in the absence of specific contractual provisions". This compromise preserved the possibility of end-user licence agreements on the one hand but also granted the "lawful acquirer" a minimum level of user rights without concluding a licence agreement on the other.

10 It is also generally accepted that the core user rights set forth in Article 5(1) cannot be waived by contract. According to Article 5 (1), (2) and (3), these minimum rights also include error correction, the creation of a backup copy, and the observation, examination, and testing of the functioning computer program. In *Top System SA v État belge* (Case C-13/20), the CJEU ruled that a "lawful acquirer" may decompile or reverse engineer the program to correct errors affecting its operation without the right holder's consent, provided it is necessary and not restricted by contract. The concept of the "lawful acquirer" (or "person having a right to use" in Article 5(2) and (3)) thus functions as a form of market regulation, offering the user legal certainty regarding the use of the computer program even in the absence of a contractual relationship.

II. The Database Directive: Lawful Users

11 The 1996 Database Directive[5], as the second instrument on digital copyright in the EU, follows the model of the Software Directive and develops it further in various respects. Article 5(a) includes temporary reproductions within the scope of copyright protection for database works. Article 6(1) derogates from this by stating that "the lawful user of a database or of a copy thereof" may perform "any of the acts listed in Article 5 which is necessary for the purposes of access to the contents of the databases and its normal use." Such a use shall not require the authorization of the right holder. For the sui generis database producer's right, Article 8(1) mirrors the privilege for "lawful users" of the database and combines it with duties of such users vis-à-vis the maker of the database. Unlike the Software Directive, Article 6(1) Database Directive does not provide for the possibility of deviating contractual conditions. Quite the opposite, according to Article 15, any contractual provision contrary to Articles 6 (1) and 8 shall be null and void.

III. The Information Society Directive (ISD): Temporary Copies and Lawful Use

12 The concept of lawful use is also present in the 2001 ISD. Article 5(1) ISD created a mandatory exception for temporary acts of reproduction when five cumulative conditions are met, namely the reproduction must be: (1) temporary, (2) transient or incidental, (3) an integral and essential part of a technological process, (4) the sole purpose of which is to enable … a *lawful use* of a work, and (5) the act has no independent economic significance.

13 In *Infopaq I* and *Infopaq II* the CJEU offered valuable interpretative guidance.[6] With regard to the requirement of lawful use the Court re-stated that excused temporary acts of reproduction must pursue the purpose to "enable [...] the lawful use of a protected work, which is in turn fulfilled when such use is authorised by the right holder or where it is not restricted by the applicable legislation".[7] In relation to the latter requirement, the Court held that acts such as the preparation of genuinely independent news summaries,[8] on-screen and cached copies,[9] and ephemeral satellite decoder reproductions,[10]

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

6 Respectively, Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* ECLI:EU:C:2009:465 (*Infopaq I*); Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* ECLI:EU:C:2012:16 (*Infopaq II*).

7 *Infopaq II* (n 6) para 42; see also Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd and Others v QC Leisure and Others and Karen Murphy v Media Protection Services Ltd* ECLI:EU:C:2011:631 (*FAPL*) para 168ff; Case C-527/15 *Stichting Brein v Wullems* ECLI:EU:C:2017:300 (*Filmspeler*) para 64.

8 *Infopaq I* (n 6) para 23.

9 Case C-360/13 *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others* ECLI:EU:C:2014:1195 (*NLA*) para 25.

10 *FAPL* (n 7) para 167.

are examples of lawful uses. Conversely, it excluded lawful use in the case of the temporary reproductions made during the streaming of a copyright-protected work from a third-party website without the right holder's authorisation.[11]

14 Importantly, the CJEU, in interpreting the conditions of Article 5(1), including the concept of lawful use, consistently stated that the function of Article 5(1) is to 'allow and ensure the development and operation of new technologies, and safeguard a fair balance between the rights and interests of rights holders and of users of protected works who wish to avail themselves of those technologies'.[12]

IV. The Information Society Directive (ISD): Private Copies and Lawful Source

15 The concept of lawfulness reappears in another form (that of "lawful source") in the CJEU's case law in relation to the private copy exception. First in *ACI Adam*[13] and then in *Copydan*[14], the CJEU has taken on the task of implicitly expanding and further elaborating the concept of user lawfulness by introducing it as a prerequisite for the enjoyment of the private copy exception when affirming that the exception concerns only reproductions made from "lawful sources".

16 In *ACI Adam*, the CJEU does not provide a definition of what constitutes an "unlawful source", but it bases its argumentation on a principle of strict interpretation of copyright exceptions and on the three-step test.[15] In *Copydan*[16], the CJEU was more explicit regarding the conditions governing "lawful source" and stated that reproductions made using unlawful sources are those which are made from protected works that are made available to the public without the right holder's consent. By doing so, the CJEU appears to embody in its reasoning a logic of exclusive control of the uses of copyright-protected works, framing copyright exceptions as dependant on private ordering.

V. The Rental and Lending Directive: Lawful Sources

17 In the 2016 *VOB* case[17] the CJEU clarified that the lending of digital copies of books, or e-lending, falls within the scope of the Rental and Lending Directive (RLD)[18]. This was an important clarification as it entails that, also for e-lending, Member States have the possibility to resort to the derogation contained in Article 6 (a remuneration scheme that benefits "at least authors") as an alternative to the exclusive right of Article 3.[19] However, the CJEU, relying on its *ACI Adam* and *Copydan* doctrine, added that this is only possible when digital copies are obtained from a lawful source.[20] In the opinion of the Court, "to accept that a copy lent out by a public library may be obtained from an unlawful source would amount to tolerating, or even encouraging, the circulation of counterfeit or pirated works".[21]

18 While it is imperative for copyright law to prevent the circulation of counterfeit or pirated works, the effect of requiring a "lawful source", especially as defined in *Copydan*, is problematic if the only permitted source becomes a licensing agreement. It is doubtful whether libraries can lawfully digitize for the scope of e-lending their (already lawfully acquired) physical collections.[22] In addition, the business model characterizing the evolving e-Book market makes it practically impossible to "buy" e-Books. Right holders can control each independent act of communication to the public, which extends to all forms of access to e-Books: sharing, renting, certain forms of lending, and even selling or re-selling an e-Book (see *Tom Kabinet*).[23]

11 *Filmspeler* (n 7) para 67.

12 See *NLA* (n 9) para 24; *FAPL* (n 7) para 164.

13 Case C-435/12 *ACI Adam BV and Others v Stichting de Thuiskopie and Stichting Onderhandelingen Thuiskopie vergoeding* ECLI:EU:C:2014:254.

14 Case C-463/12 *Copydan Båndkopi v Nokia Danmark A/S* ECLI:EU:C:2015:144.

15 *ACI Adam* (n 13) paras 31–37.

16 *Copydan* (n 14) paras 75-79.

17 Case C-174/15 *Vereniging Openbare Bibliotheken (VOB) v Stichting Leenrecht* ECLI:EU:C:2016:856.

18 Directive 2006/115/EC of the European Parliament and of the Council of 12 December 2006 on rental right and lending right and on certain rights related to copyright in the field of intellectual property. The original Directive was enacted in 1992.

19 M. Frigeri, M. Kretschmer, P. Mezei (2024) Copyright and eLending in public libraries: an incomplete revolution?, JIPITEC, 15, 156, para 1.

20 *VOB* (n 17) para 72.

21 *VOB* (n 17) para 68.

22 *VOB* (n 17) paras 66–72; but see Case C-117/13 *Technische Universität Darmstadt v Eugen Ulmer KG* ECLI:EU:C:2014:2196, paras 36–49.

23 Case C-263/18 *Nederlands Uitgeversverbond and Groep Algemene Uitgevers v Tom Kabinet Internet BV* (*Tom Kabinet*) ECLI:EU:C:2019:1111; Frigeri and others (n 19) para 1.

19 Accordingly, applying the *ACI Adam* and *Copydan* concept of "lawful source" to the case of e-lending has the paradoxical effect of nullifying the important point made by the Court in *VOB* that e-lending is a form of lending – and not of communication to the public – benefitting from the derogation of Article 6 RLD. This incongruency shows, once again, how a *prima facie* legitimate condition (lawfulness of the source) needs to be defined and interpreted in the light of the broader legal and policy framework in order to give proper effect to the important principles established in the law and avoid unwanted, even contradictory, outcomes.

VI. The CDSMD: Lawful Access and the TDM Provisions

20 Both Arts. 3 and 4 of the CDSMD establish a requirement of "lawful access" for TDM purposes. Although this requirement is not defined in Article 2, the preamble of the Directive offers important guidance. Recital 14 identifies four primary types of lawful access related to Article 3: access based on open access policies, access through contractual arrangements between right holders and research organizations (such as subscriptions), access through other lawful means, and access to content that is freely available online. On the other hand, Recital 18 provides one main example ("including") of lawful access in relation to Article 4: access to content made available to the public online.

21 Regarding Article 3, the first three instances of Recital 14 appear quite straightforward. The fourth (content freely available online) introduces a concept which seems to be new in the legislative instruments of the EU copyright *acquis*.[24] The literature has debated the exact meaning of this requirement in an attempt to determine whether it should be interpreted within the framework developed by the CJEU as summarised in Sections B.I – B.V above, or whether the terminological inconsistency of the framework and the specific choice of wording in the CDSMD should be read as a legal basis and a legislative intention, respectively, to establish a new category in need of autonomous treatment.[25]

24 The expression "freely available online" has been used by the Court in cases such as *CV-Online Latvia* or *VG Bild-Kunst*, but without developing it into a clear legal category.

25 T. Synodinou, G. Vrakas, Lawful Access as a Gatekeeper for TDM in the EU: An Analysis of the Scope of the Lawful Access Requirement under EU Copyright Exceptions, Verfassungsblog, 17 November 2025, https://verfassungsblog.de/lawful-access-gatekeeper/; M. Senftleben (2026) TDM, GenAI and the Copyright Three-Step Test, International Review of Intellectual Property and Competition Law (IIC), 57, 67–107; T. Margoni (2024) TDM and Generative AI: Lawful Access and Opt-Outs, Auteurs & Media, 172; K. Szkalej (2025) The Paradox of Lawful Text and Data Mining? Some Experiences from the Research Sector and Where We (Should) Go from Here, GRUR International, 74(4), 307–319, https://doi.org/10.1093/grurint/ikaf029.

26 See Article8 of Directive 2009/24/EC (codified version of Directive 91/250); Article15 of Directive 96/9/EC.

27 *Infopaq II* (n 6) paras 44 and 45.

28 *FAPL* (n 7) paras 170–172.

29 *Filmspeler* (n 7) paras 65-70.

C. Lawfulness as a Condition for Exceptions in EU Law: Terminological Consistency or a Versatile Taxonomy?

I. Similarities and Differences in the Conceptions of Lawfulness

22 The concept of lawfulness in relation to user acts is pluriform and has appeared in EU copyright legislation or was introduced by the CJEU's case law in relation to exceptions established for specific protected subject matter (computer programs, databases), and exceptions for all categories of copyright protected works (temporary copies, private copies, public lending, or TDM exceptions).

23 Both the terminology and the object of the specific expressions of user lawfulness point towards a versatile taxonomy. In the case of specific subject matter, the emphasis is on the user (lawful acquirer, lawful user), while the recognition of the lawful user as beneficiary of the exceptions marks also a paradigm shift towards the recognition of certain of these exceptions as mandatory, both in the sense that Member States shall provide for those exceptions and, more significantly, in the sense that these exceptions cannot be overridden by contractual terms.[26] In relation to the condition of lawful use in the context of the temporary copy exception, the focus has been given mainly to a flexible understanding of lawfulness in compliance with Recital 33 ISD (*Infopaq II*[27], *FAPL*[28]), while in *Filmspeler*[29] lawfulness was conceived more holistically by taking into account the status of the user's knowledge in relation to the legality of the source of the copy of the work (the *mens rea* of kodi box users), who deliberately access a free and unauthorised database of protected works. In contrast, in the private copy cases the assessment of "lawfulness" is strictly linked to the source of the copy and does not take into consideration the end-user's knowledge in relation to the unlawfulness of the source of the copy. Significant differences also exist in how lawfulness is assessed by the CJEU. In some cases, the Court adopts a flexible approach. Access or use may be considered lawful because

no legal rule restricts it or because it is justified on other legal grounds. In other cases, the Court applies stricter conditions and requires authorization from the right holder.

24 The diversity of the concept may be ascribed to an evolutionary path that reflects technological, economic and social advancements and the corresponding mutations of the copyright ecosystem (from acquisition of a copy as a basis of lawful use to lawful access based on express or implied consent for content that is freely available online).[30] Furthermore, it is noteworthy that the diverse expressions of user lawfulness have been attributed, as far as now, to a necessarily restrictive function. While the terms may differ ("lawful user/acquirer", "lawful use", "lawful source", "lawful access") the underlying common normative axiom is that the protected work must have been accessed lawfully.

25 However, despite its inherent restrictive function, the concept of lawfulness is marked by a certain elasticity. That is to say, depending on its interpretation, the concept could enable a comprehensive assessment of the conditions of access and use of a copyright protected work, which might take into account general principles of law (such as good faith in contractual relations, extracontractual liability standards, the knowledge of the user) as well as fundamental rights.

II. Between Flexibility and Certainty: A Principled Interpretative Framework

26 In the previous section we showed that lawfulness is a flexible concept whose meaning depends on the use, technology, or situation to which it applies. The next step is to determine what counts as lawful within that flexible framework. This determination must be as precise as possible to ensure legal certainty, predictability, and consistency. These requirements follow not only from Recitals 11 and 18 CDSMD, but also from general principles of EU lawmaking.[31] To that end, the following paragraphs proceed in two stages. They first identify the main "hard" or horizontal limits within which any definition of lawfulness must operate. Then they set out a number of "soft" principles that may guide the application of that definition in specific contexts.

1. Technological Protection Measures (TPMs)

27 A first "hard" boundary of lawfulness appears to be provided by effective Technological Protection Measures (TPMs). It seems indisputable that circumventing an effective TPM would make access or use unlawful. This conclusion finds ample support in the letter of the law. In relation to TDM, the cryptic Article 7(2) second sentence of the CDSMD[32] points towards the ISD obligation on Member States to take appropriate measures to ensure that right holders make available to the beneficiary of an exception or limitation provided for in national law a non TPMed version of the work.[33] In *Nintendo* the CJEU has arguably limited the anti-circumvention protection of effective TPMs "only in the light of protecting that right holder against acts which require his authorisation".[34] In relation to the temporary copies of Article5(1) ISD the concept of lawful use defined as "not restricted by applicable legislation" performs a comparable function: the legal protection afforded by Article 6 ISD against the circumvention of an effective TPMs is a use "restricted by applicable legislation". Finally, a similar effect is also found in relation to illegal access to information systems in the cited Article 3 of Directive 2013/40/EU, which explicitly mentions the infringement of a security measure.

2. Contracts

28 Valid contractual provisions arguably follow a similar logic and may be classified as a second "hard" boundary to lawfulness. Contracts (or unilateral declarations with similar legal effects) appear particularly relevant in the case of Article 4 opt-outs in the CDSMD: once a right holder has validly reserved TDM uses for a work, access to it would become unlawful. This does not apply, of course, to situations where the law establishes otherwise, such as in the cases covered by Article3 (see Article 4(1) and 7(1) CDSMD) or Article 5(2) and (3) Software Directive. A similar relevance of contractual arrangements is observable in the cited cases of the

30 The underlying idea is that the right holder who has made available content online freely is presumed to have implicitly authorised access to the work for all Internet users. This shall be seen as an evolution of the concept of "lawfulness" that enriches its normative content with a more liberal approach which considers the reality of online access to content.

31 Article7 TFEU; European Parliament, the Council of the European Union and the European Commission, "Interinstitutional Agreement on Better Law-Making", Document 32016Q0512(01), of 13 April 2016.

32 "The first, third and fifth subparagraphs of Article 6(4) of Directive 2001/29/EC shall apply to Articles 3 to 6 of this Directive".

33 In these cases the rather complex and nationally fragmented procedure of Article6 ISD should be followed.

34 Case C-355/12 *Nintendo Co Ltd and Others v PC Box Srl and 9Net Srl* ECLI:EU:C:2014:25, para 25.

lawful acquirer of Article 5(1) Software Directive, in the legal source requirement found in *VOB*, and ultimately also in the CJEU stance expressed in the *ACI Adam* and *Copydan* rulings focusing on the initial consent of the right holder. In other words, the authorization by right holders is a clear indication of lawfulness. However, right holders' authorization (either in the form of a valid contractual provision or effective TPM) cannot form the only legal basis for lawful access.

3. Beyond Right Holders' Authorization

29 Recital 14 CDSMD is clear in this sense, where, next to contracts and open access policies, it lists "other lawful means" as well as "freely available online" as additional basis for lawful access. Furthermore, the fact that right holders' authorization cannot be the only legal basis for lawfulness is supported by a broader argument of general theory. By definition, an exception to an exclusive right identifies a situation where the authorization of right holders is not needed for the acts covered by the exception. If such an authorization was needed, there would not logically be space for an exception, representing that situation the default dynamic where an exclusive right is licensed (i.e., authorised) by right holders, usually via contracts. In other words, the very same definition of exception to an exclusive right is the absence of the need of authorization by the holder for the acts in question.

30 Therefore, creating a new condition of validity for an exception (lawful access/use/etc.) and pretending to define it as an act that needs to be authorised by its holder, is contradictory on a logical and terminological level as well as on a legal and doctrinal one. Consequently, to comply with the general principles of EU law, lawfulness can only mean *more* than authorization by right holders. This additional meaning must be defined, and a primary source is the text of the law itself. In the context of TDM, Recital 14 CDSMD is particularly relevant. It refers to access through "other lawful means" and to content that is "freely available online" as further examples of lawfulness.

4. Other Lawful Means

31 Regarding "other lawful means" it seems implicit in the texture of the law that these cases cover at least forms of authorization statutorily established, for instance exceptions and limitations to copyright, as well as other forms of free uses outside the scope of rights of economic exploitation (in this sense see Recital 9 CDSMD or Recital 33 ISD) or the exhaustion of some of these rights (e.g., Article 4(2) ISD).[35] This is systematically coherent. Within the field of intellectual property, and EU law more generally, it is not uncommon to identify "the law" or specific legal provisions as a valid basis to limit the scope, even entirely, of a right under EU law (see e.g., Article 3(2) Trade Secrets Directive; Article 6(1)(c) GDPR; or Article 15(b)(i) Data Act).

5. Freely Available Online

32 Recital 14 CDSMD includes also "freely available online" as a basis for lawfulness. However, establishing how far this novel notion can stretch the concept of "lawful access" is a matter of debate. Different interpretations are available. One plausible reading is that, outside of the clearly unlawful situations just described (effective TPMs and valid contracts), all other forms of access to content "freely available online" must be lawful. In other words, as long as the behaviour of an agent (e.g., a researcher) does not intentionally circumvent an effective TPM or breach a valid contractual provision, access is lawful, regardless of whether the work so accessed was made available with the consent of the author, thus also including situations that under the *ACI Adam* doctrine would qualify as unlawful sources. The reasons for this reconstruction, which is particularly attractive for cases of scientific research covered by Article 3 CDSMD, can be located in a literal, systematic, teleological and fundamental rights-based reading of the lawful access requirement in the CDSMD.[36] The reasons for treating scientific research as a special case are discussed in further detail below.

33 A second reading is similar in nature but affords a stronger relevance to the role of the knowledge element as developed by the CJEU in cases of communication to the public (e.g., *Svensson*[37], *GS Media*[38], *Filmspeler*, etc.). Following this line of argumentation, beneficiaries of Article 3 CDSMD will most likely also benefit from the presumption of not knowing that the accessed work was made available without the consent of the author and thus are covered by such a presumption,[39] at least until

35 Rec 9 CDSMD and 33 ISD offer examples of acts that fall outside the scope of the rights reproduction, respectively in the case of TDM and temporary reproductions.

36 Margoni (n 25).

37 Case C-466/12 Svensson and Others v Retriever Sverige AB ECLI:EU:C:2014:76.

38 Case C-160/15 GS Media BV v Sanoma Media Netherlands BV and Others ECLI:EU:C:2016:644.

39 The requirements are formally different, but they will coincide in most situations. Article3 is available to research organisation or cultural heritage institutions acting for research purposes while the presumption of "not

the time when they acquire direct knowledge for instance via a notification. Accordingly, attributing a role to the subjective status of the agent seems in line with a discrete judicial approach of the CJEU as well as with general principles of law opposing the idea of knowingly benefiting from an illicit act. In the case of Article 4 CDSMD (which will often but not always cover commercial TDM), such presumption will likely not operate and thus a higher standard of care (such as verifying the origin of the datasets when operating commercially) applies.[40] This argument is inter alia able to address effectively the different spirit of TDM performed for scientific research on the one hand and TDM performed by commercial actors on the other.

34 One possible objection to embracing the knowledge element in this context is that knowledge can be acquired in different ways, even after the TDM process ended, for instance, when receiving a notification informing the agent that a certain work had not been made available with the consent of the author. At this point, the agent is under an obligation to remove that work. However, while removing a picture from a website (e.g., *Renckhoff*[41]) or a link from a blog (*GS Media*) can be easily done, removing potentially hundreds of thousands of works from a training dataset or from a trained model would be very costly, perhaps even impossible from a technical point of view in the case of a trained AI model. As such, the net effect of this dynamic would be to push this sector in the direction of a pure licensing market and render the TDM exceptions ineffective, an outcome incompatible with EU law, as clearly established for instance in *VOB*[42] of *FAPL*.[43]

35 Additionally, whether the element of knowledge – developed by the CJEU in relation to specific cases of communication to the public – should apply to the right of reproduction may raise systematic objections.[44] To be clear, the fact that most instances of lawful use or access refer to cases of reproduction is not a decisive argument against the adoption of a standard developed for the different right communication to the public. But it does represent an element of discontinuity and may therefore justify a different treatment. Accepting knowledge as a defining component of lawfulness represents an important step whose consequences should be carefully assessed particularly in relation to "special" cases such as scientific research (further discussed below), which might therefore require different standards of knowledge.

36 Another aspect may play an important role in this analysis: is the meaning of "lawful access" in Articles 3 and 4 CDSMD the same? Whereas the logic of attributing an identical meaning to a concept that receives an almost identical formulation ("lawful access" in Article 3; "lawfully accessible" in Article 4) in two consecutive provisions of a same Directive is apparent, a deeper look at the history, scope, preamble and protected fundamental rights may suggest a different conclusion. Accordingly, defining the precise contours of lawfulness, even within two apparently close provisions, should likewise be object of a detailed analysis. The argument that scientific research as defined in Article 3 should again deserve a "special" treatment is further developed in the next section.

37 In conclusion, the requirement of lawfulness, and the activities attached to it, such as access, use, or source, appear to perform a similar regulatory function across different legal contexts. Yet they are also shaped by contingent circumstances. It is arguable that EU legislators and courts have used lawfulness to regulate information flows and creative markets. But the relevant subject matter differs significantly. The legal framework applies to different goods, such as software, databases, and e-books. It also applies to different activities, including private copying, temporary copying, and TDM. These rules were adopted in different historical periods, with 28 years separating the Software Directive from the CDSMD. They also respond to changing technological and socio-economic conditions, from the distribution of copies of software programs to the artificial generation of code.

38 For these reasons, any definition of lawfulness must remain within the "hard" boundaries identified above, namely effective TPMs, valid contracts, and other lawful means. At the same time, its application in a specific case must take account of the evolution and diversity of the subject matter and activities that the legislature sought to regulate. Within this

knowledge" is applicable in cases of absence of for-profit purpose (see e.g. *GS Media* (n 37) paras 49–51).

40 See *GS Media* (n 37) para 51.

41 Case C-161/17 Land Nordrhein-Westfalen v Renckhoff ECLI:EU:C:2018:634 (Renckhoff).

42 *VOB* (n 17) para 50.

43 *FAPL* (n 7) para 163.

44 TDM, temporary copies and private copies are all instances of reproduction, not communication to the public; the *VOB* case refers to public lending; the lawful acquirer of Article5 Software Dir. does not need authorization for acts of reproduction and adaptation, etc.

broader variety, some special cases require separate discussion.

III. The Specialty of Scientific Research

39 The interpretation of "lawful access" in relation to scientific research must begin with the specific structure and purpose of Article 3 CDSMD. That provision reflects a deliberate legislative choice to privilege scientific research as a matter of public interest. Recital 11 states that legal uncertainty surrounding TDM for scientific research should be addressed by a mandatory exception. That objective is implemented inter alia through a narrowly circumscribed beneficiary concept. Article 2(1) CDSMD defines "research organisation" cumulatively and restrictively: the entity must pursue scientific research as its primary objective; it must operate on a non-profit basis or reinvest profits in research, or act pursuant to a recognised public-interest mission; and it must be structured so that no undertaking exercising decisive influence enjoys preferential access to its results. These conditions ensure that the exception serves genuinely public-oriented research rather than commercial exploitation by proxy.

40 Article 3 is further limited by a strict purpose requirement and by safeguards concerning storage, retention and security of copies. Taken together, these elements demonstrate that the Union legislature struck a specific balance: research organisations are granted broader freedom to carry out TDM, but only within a framework that confines the activity to scientific purposes and prevents structural capture by commercial actors.

41 That balance must inform the interpretation of "lawful access." The wording of Article 3 refers to works "to which they have lawful access." The focus is thus placed on the relationship between the beneficiary and the act of access. As seen above, Recital 14 clarifies that lawful access includes access based on open access policies, contractual arrangements such as subscriptions, other lawful means, and content "freely available online." The introduction of this latter category would be deprived of autonomous meaning if it were reduced to a requirement that the original online posting be authorised in every instance. In the digital environment, the legal status of an upload is frequently indeterminate. Requiring research organisations to verify ex ante the lawfulness of each source would, in many settings, render Article 3 ineffective and reintroduce precisely the legal uncertainty the legislature sought to remove.

42 The Court's case law on unlawful sources must be read in its specific context. In *ACI Adam* and *VOB*, the Court held that certain exceptions under the ISD cannot extend to copies derived from unlawful sources, in order to preserve the fair balance between right holders and users. In parallel, in its case-law on online communication to the public, including *Svensson*, *GS Media*, *Filmspeler*, *Renckhoff* and *Stichting Brein*, the Court has developed a context-sensitive approach that pays close attention to the role, knowledge and profit-making intent of the actor concerned. That jurisprudence demonstrates that the concept of lawfulness is not abstract and uniform across the *acquis*, but instrument-specific and closely linked to the objectives and structure of the provision at issue. It cannot therefore be transposed mechanically to Article 3 CDSMD, which was adopted precisely to facilitate scientific research and to reduce legal barriers in that domain.

43 The special status of research under Union law reinforces that conclusion.[45] Academic freedom and the freedom of scientific research are protected under Article 13 of the Charter of Fundamental Rights. The CDSMD itself recognises the societal value of research and the limited harm to right holders resulting from the Article 3 exception. Within that constitutional and legislative framework, restrictions on research activities must be interpreted in a manner that preserves the effectiveness of the exception and respects the Union's commitment to knowledge creation, innovation and democratic participation.

IV. The Specialty of General-Purpose AI (GPAI)

44 Developments under the AI Act confirm the need for a conduct-based understanding of lawful access. Article 53(1)(c) of the AI Act requires providers of general-purpose AI (GPAI) models to put in place a policy to comply with Union copyright law, in particular to identify and respect rights reservations under Article 4(3) CDSMD. The Copyright Chapter of the GPAI Code of Practice,[46] notably Measure

45 European Copyright Society (2025), An EU Copyright Framework for Research: Opinion of the European Copyright Society, available at: https://europeancopyrightsociety.org/wp-content/uploads/2025/07/ecs_copyright-framework-for-research.pdf.

46 Code of Practice for General-Purpose AI Models, Copyright Chapter (Working Group Alexander Peukert/Céline Castets-Renard), 2025, https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. See also European Commission, Commission Opinion on the assessment of the General Purpose AI Code of Practice within the meaning of Article 56 of Regulation (EU) 2024/1689, C(2025) 5261 final; AI Board, Conclusion of the Artificial Intelligence Board on the Assessment of the General-Purpose AI Code of Practice pursuant to Article 56 of Regulation 2024/1689, https://ec.europa.eu/newsroom/dae/redirection/

1.2, operationalises that obligation in the context of web crawling. It requires providers not to circumvent effective TPMs, including paywalls and subscription systems, and to exclude websites that have been recognised by courts or public authorities in the Union as persistently infringing copyright on a commercial scale. Significantly, the measure does not impose a general obligation to verify the authorisation status of every individual work encountered online. It translates "lawfully accessible" into a standard of lawful conduct and due diligence.

45 Additionally, it is worth noting that if those conduct-based requirements are considered sufficient to ensure compliance with Article 4 CDSMD for commercial GPAI model providers, it would be incoherent and disproportionate to subject research organisations under Article 3 to stricter lawful-source verification obligations. Such an approach would invert the differentiated structure of the Directive by imposing the heaviest compliance burden on the very actors whom the legislature intended to privilege in the public interest. The differentiated regulatory design of Articles 3 and 4 presupposes that commercial actors are subject to broader exploitation rights combined with an opt-out mechanism, whereas research organisations benefit from a mandatory exception within a confined institutional and purposive framework.

V. The Specialty of Open-Source Software and AI Models

46 Whether AI coding assistants, such as Github Copilot, ChatGPT4, Google Gemini or Claude can be trained on free and open-source software (FOSS) is another interesting case deserving a dedicated analysis. The question is of considerable relevance given that AI coding assistants largely rely on FOSS as training data, proprietary closed source code not being available for such a use by third parties.

47 Whether the reproductions necessary for the training of these models are permitted on the basis of statutory limitations or only on the basis of FOSS licences – and if so, under what conditions – has not yet been decided by the courts.[47] In Europe, the TDM exception in Article 4 CDSMD is the focus of most debate.[48] Recital 14 CDSMD explicitly mentions works published on the base of "an open access policy". This may well suggest that the TDM exception should apply to open-source software, unless rights holders have opted-out. However, from the point of view of the open-source community such an interpretation could lead to an undesirable outcome: their software could be used for AI training without having to comply with the conditions of FOSS licences, including the keeping of the notices referring to the applicable licence, the mentioning of the authors and the application of the same licence to derivative works. A possible solution may be to combine the usual copyright notice accompanying FOSS licences with a reservation of rights in accordance with Article 4(3) CDSMD. Other projects grounded in similar "open" principles and licences, including Wikipeida, are likewise considering the impact of AI training on their open datasets and how elements such as attribution or other licensing restrictions may legally survive training activities.[49]

VI. The Specialty of Lawfulness in the Three-Step Test

48 Lawfulness has been considered in a few CJEU decisions regarding the three-step test. In *Infopaq II,* discussing information retrieval services, the Court clarified that as long as the five conditions of Article5(1) including "lawful use" are met, the three-step test of Article5(5) ISD is satisfied.[50] As discussed in more detail above, the Court clarified the importance of the exception for temporary acts of reproduction (and thus of its constituent elements including lawful use) as an essential element to "ensure the development and operation of new technologies" and "safeguard a fair balance between the rights … of right holders … and of users … who wish to avail themselves of those new technologies".[51]

49 On the other hand, in *ACI Adam*, when examining the private copy exception, the Court argued that a national provision that does not distinguish between *lawful and unlawful sources* in its implementation of Article 5(2)(b) ISD would adversely affect the normal exploitation of the work (step 2) and unreasonably prejudice the interest of right holders (step 3).[52]

50 Regarding TDM, absent any case law from the CJEU, scholarship has argued that the nature of "lawful access" seems closer to cases of temporary copies

document/118687.

47 In the United States, Doe v GitHub No 22-cv-06823-JST (ND Cal) addressed whether code snippets in the output of GitHub violated the copyright provisions of the DMCA; possible reproductions in the training process were not part of the complaint.

48 Software is not listed in Article's 3 scope.

49 See e.g,. the entry "Artificial intelligence in Wikimedia projects",

https://en.wikipedia.org/wiki/Artificial_intelligence_in_Wikimedia_projects (accessed 10 June 2026).

50 *Infopaq II* (n 6) para 57.

51 *FAPL* (n 7) para 164; *NLA* (n 9) para 24.

52 *ACI Adam* (n 13) paras 39 and 40; *Senftleben* (n 25) sec 3.3.

in technological processes than to cases of private copy. This difference is justified by the fact that private copies "made from an unlawful source would encourage the circulation of counterfeited or pirated works, thus inevitably reducing the volume of sales or of other lawful transactions relating to the protected work" and thus encroaching on the normal exploitation and legitimate interest parameters.[53] Conversely, in cases of temporary copies, and arguably of TDM, there are no copies that may operate a substitution effect and therefore conflict with the normal exploitation of the work.[54] Temporary copies and copies made in TDM processes are often provisional in nature and serve only the purposes of allowing a technological process. In the case of TDM, they enable the analysis of "text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations".[55] TDM copies are thus not suitable to "encourage the circulation of counterfeited or pirated works", nor they reduce "the volume of sales or of other lawful transactions relating to the protected work" as described in the different case of the private copies in *ACI Adam* requiring a lawful source.

51 Consequently, in light of the specific wording of Recital 14 CDSMD, in particular its reference to content that is "freely available online", and of the specific nature and limited scope of Article 3 CDSMD focusing on scientific research, the concept of lawful access broadly interpreted does not seem to conflict openly with the three-step test.[56] In particular, as evidenced above, the initial authorization of the rights holder cannot be used as the exclusive legal basis for the validity of an exception. This logic necessarily reflects in the way in which the test, and in particular its second step, should be interpreted.[57]

52 This interpretation is confined to acts of TDM, that is, acts involving the use of protected works as input for the training of AI models. As the literature makes clear, cases in which the output of a generative AI system, whose underlying model was trained through TDM activities, generates copies of protected works raise a distinct legal issue. Where, for example, the output of a generative AI application reproduces a protected work, this is better understood as a separate category of use and, arguably, as a case of *prima facie* infringement, unless justified by another legal basis.[58]

D. Conclusions

53 Lawfulness, in its various manifestations, regulates cultural production and information markets by distinguishing legitimate from illegitimate behaviour. It provides a flexible interpretative tool for assessing how EU copyright law applies in specific contexts, including software, databases, private copying, temporary copying, e-lending, TDM, and GPAI. This assessment cannot be made in the abstract. It must reflect the relevant technological and economic context. At the same time, legal certainty must be safeguarded and thus the flexibility of lawfulness is guided, and sometimes constrained, by a broader legal framework that provides both "hard" boundaries and "soft" guiding principles. Reconstructing this framework is not straightforward. It requires attention to historical, doctrinal, and theoretical layers. Yet it is necessary to ensure the legal certainty and consistency required by EU law in order to favour economic and social development as well as legal compliance. The following recommendations may be drawn from the analysis developed in the previous sections.

1. Lawfulness cannot be reduced to right holders' authorization. This applies in particular to exceptions and limitations conditioned on lawful access, lawful use or lawful source. The text, structure and function of EU copyright law support this conclusion. Since an exception to an exclusive right presupposes that authorization is not required, making authorization the sole condition of lawfulness would create circularity and would risk rendering exceptions ineffective.

2. Beyond authorization by right holders, there are other clear cases in which lawfulness is undisputed. These include situations where the law authorizes the relevant activity, including through exceptions, limitations or other free uses under copyright law or other legal sources; situations that fall outside the

53 M. Senftleben (2026) GenAI and the Copyright Three-Step Test – Do TDM Exceptions for AI Training Conflict with a Work's Normal Exploitation?, GRUR International, 75(1), https://doi.org/10.1093/grurint/ikaf113.

54 In addition, see Panel Report, United States – Section 110(5) of the US Copyright Act (WT/DS160/R, 15 June 2000) https://doi.org/10.30875/40287e1f-en.

55 Article 2(2) CDSMD.

56 For a detailed analysis of the compatibility of the CDSMD TDM exceptions with the European and international three-step test, see M. Senftleben (2022) Compliance of National TDM Rules with International Copyright Law – An Overrated Nonissue?, available at https://dx.doi.org/10.2139/ssrn.4134651.

57 Arguably, right holders will have incentives to see any unauthorized use as conflicting with the normal exploitation of their works, but for the reasons explained above, this interpretation is logically unsustainable and in contrast with international and EU interpretations of the 3ST.

58 Senftleben (n 25) 89-90, and the doctrine cited therein, especially fn 153.

scope of an exclusive right; and situations where the relevant exclusive right has been exhausted.

3. Situations not expressly regulated by law require a context-specific assessment. That assessment must respect the hard boundaries of lawfulness, including effective technological protection measures, valid and enforceable contractual provisions or rights reservations, and other legally binding restrictions. It must also be guided by softer interpretative criteria, including whether content is freely available online, the user's knowledge, the nature and purpose of the activity, the relevant fundamental rights framework and, where applicable, the three-step test. The interpretation adopted must preserve the policy objectives and effectiveness of the rule at issue.

4. Scientific research deserves special treatment within this framework. Article 3 CDSMD reflects a deliberate legislative choice to privilege scientific research through a mandatory exception limited to specific beneficiaries and purposes. In light of the public-interest function of research and its connection to academic freedom under EU law, lawful access should be interpreted in a manner that preserves the effectiveness of that exception and avoids disproportionate source-verification obligations for research organisations and cultural heritage institutions.

5. Other special situations have been identified and further ones may emerge. These include general-purpose AI, open-source software and AI models, and other contexts in which access, use and lawfulness interact with public-interest objectives, open licensing practices or technological innovation. Such situations should be assessed under the framework proposed in this Opinion. As a general principle, where a new situation warrants treatment comparable to that granted to scientific research, the conditions of lawfulness should be interpreted in a manner that favours the relevant activity, subject to the applicable legal safeguards.

Drafting committee: *Thomas Margoni; Martin Kretschmer; Axel Metzger; João Pedro Quintais; Tatiana Synodinou*

Signatories

Prof. Lionel Bently, Professor of Intellectual Property Law, University of Cambridge, United Kingdom

Prof. Christophe Geiger, Professor of Intellectual Property Law, Luiss Guido Carli University, Rome, Italy

Prof. Jonathan Griffiths, Professor of Intellectual Property Law, Queen Mary, University of London, United Kingdom

Prof. Martin Kretschmer, Professor of Intellectual Property Law, University of Glasgow and Director, CREATe, United Kingdom

Prof. Thomas Margoni, Research Professor of Intellectual Property Law, Centre for IT & IP Law (CiTiP), KU Leuven, Belgium

Prof. Axel Metzger, Professor of Civil and Intellectual Property Law, Humboldt-Universität, Berlin, Germany

Prof. Péter Mezei, Professor, University of Szeged, Hungary; Adjunct Professor (dosentti), University of Turku, Finland

Prof. Alexander Peukert, Professor of Civil, Commercial and Information Law, Goethe-University Frankfurt am Main, Germany

Prof. João Pedro Quintais, Associate Professor, University of Amsterdam, Institute for Information Law (IViR), Netherlands

Prof. Ole-Andreas Rognstad, Professor of Law, Department of Private Law/Centre for European Law, University of Oslo, Norway

Prof. Martin Senftleben, Professor of Intellectual Property Law and Director of the Institute for Information Law (IViR), University of Amsterdam, Netherlands

Prof. Caterina Sganga, Professor of Comparative Private Law, Scuola Superiore Sant'Anna (Pisa), Italy

Prof. Tatiana Eleni Synodinou, Professor of Private and Commercial Law, University of Cyprus, Cyprus

Prof. Mireille van Eechoud, Professor of Information Law, Institute for Information Law (IViR), University of Amsterdam, Netherlands

jipitec

Journal of
Intellectual Property,
Information Technology,
and Electronic Commerce
Law

www.jipitec.eu

31 jipitec 216