The Obligations of Providers of General-Purpose AI Models

Michael Veale[1,2] and João Pedro Quintais [2]

[1]Faculty of Laws, University College London
[2]Institute for Information Law, University of Amsterdam

To appear in Zanfir Fortuna, Malgieri, González Fuster and Mantelero (eds.) *The Artificial Intelligence Act — A Thematic Commentary* (Hart forthcoming 2026). This version: November 2025.

During the legislative process, the EU Artificial Intelligence (AI) Act was amended to include provisions related to general-purpose AI (GPAI) models. These broadly relate to transparency towards downstream users and relevant regulators, in addition to obligations connected to intellectual property. In this paper, we provide detailed analysis of these new provisions in the context of current technological applications and emerging trajectories, connecting them to computing literature and practice, and the broader context of connected and adjacent legal regimes, in particular copyright and relevant emerging case law. We find that there are a significant number of inclarities, tensions and contradictions both within the text, between the text and other legal regimes, and between the text and guideline documents, such as the Code of Practice on General-Purpose AI and recent guidelines by the European Commission. We identify a range of issues with the scoping of the provisions which may undermine its policy goals and create loopholes for regulatory avoidance, such as those relating to non-commercial models, open-source models, and model finetuning along the value chain. We find that the Code of Practice contains significant omissions and misstatements, some of which may present a compliance risk for an entity choosing to rely on the Code. We do not consider the provisions on GPAI models which present a systemic risk, which are dealt with elsewhere in the volume which this work will form a part of.

A Legislative Background 2

B Definitions 5

C Information and Documentation 5
I Overview 5
II Nature of the documentation . . . 6
(a) Level of detail 6
(b) Documentation updates . 8
(c) Retention 8
III Tasks and integration potential . . 8
IV Data 9
(a) Provenance 9
(b) Number of data points . 10
(c) Other measures and methodologies 11
(d) Interaction with other regimes 11
V Availability and confidentiality . . 11

D Copyright and Related Rights 13
I General issues in copyright and AI 13
(a) Locations of challenges . 13
(b) Distinct legal backdrops . 14
II Text and Data Mining (TDM) . . 15
(a) Basic regime 15
(b) Copyright–related obligations in the AI Act 16
(c) Value chains and model integration 16
(d) Distinct regimes for non-compliance 17
(e) Wide scope 17
III Policies to respect copyright . . . 18
(a) Copyright and model outputs 19
(b) Lawful access 20
(c) Reservation of rights . . 22
(d) Appropriate reservation method 23
(e) Granularity and temporality 25
(f) Applicable law 27
IV Training data summary 30

E Demonstrating Compliance 34
I Modalities 34
(a) Harmonised standards . . 34
(b) Common specifications . 35
(c) Codes of practice 35
(d) Common rules 38
(e) Adequate alternative means 38
II Authorised representatives 38
(a) Function 39
(b) Liability of authorised representatives 39

F Exclusions 40
I Free and Open-Source 40
II Non-Commercial Activities . . . 42
III Scientific Research and Development 43
(a) Models for research purposes 43
(b) Models during development 44
IV Value Chains and Fine-tuning . . 44

G Concluding remarks 47

A Legislative Background

The Commission draft of the AI Act published in April 2021 did not contain provisions around 'general-purpose AI' (GPAI) models or systems. The primary impact of that version on such models or systems would have arisen from prohibited practices, transparency obligations for specific uses like synthetic content generation, regulation under the New Legislative Framework approach for high-risk application areas, and the invalidation of Member State rules that imposed additional requirements on placing onto the market, putting into service or use of AI systems, such as GPAI systems, beyond those outlined in the (then) draft regulation.[1]

[1] Commission, 'Proposal for a Regulation of the 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; Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22 Computer Law Review

Obligations for GPAI providers were initially quite strict. In November 2021, the Slovenian Presidency of the Council circulated a draft which clarified that GPAI systems themselves should be considered as having no intended purpose and therefore should not, without specific intention, fall into the category of high-risk systems.[2] In May 2022, the French Presidency circulated additions to the text. They removed the Slovenian Presidency's exclusion and added new provisions stating that GPAI systems which have a possibility of use rather than an intention of use as high-risk systems or components thereof, should comply with a subset of the obligations that apply to high risk systems, as well as have a duty of co-operation with downstream providers, unless they explicitly and in good faith exclude high-risk uses in their information or instructions of use.[3] The Council's general position circulated in November 2022 changed little from this approach, other than delegating significant powers to the Commission to shape the details of how GPAI systems would fit into this regime through implementing acts.[4]

Later in that month, on 30 November 2022, OpenAI launched ChatGPT. This exposed legislators and the public to a version of the GPT 3.5 model released earlier in the year, this time through an attractive user interface.[5] ChatGPT reignited debate around GPAI throughout the next one-and-a-half years of legislative discussion on the file.[6] The Parliament's negotiating position introduced a variety of obligations on 'foundation model' providers, including obligations unique to these providers relating to copyright and energy usage. [7] These topics were, until that moment, not part of the AI Act draft and as such not subject to any prior impact assessment. The Council, led by France and Germany, in turn lobbied by GPAI model providers Mistral and Aleph Alpha, removed binding obligations for such providers beyond transparency provisions.[8] Those companies emphasised publicly that the final Act, which they initially opposed, was now 'fine' and 'entirely manageable' due to the firms undertaking

International 97.

2 Council, '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 – Presidency compromise text' (No 14278/21, 29 November 2021) recital 70(a), arts 3(12), 52(a).

3 Council, 'Proposition de Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union - Text de compromis de la présidence - Article 3, paragraphe 1 ter, Articles 4 bis à 4 quater, Annexe VI (3) et (4), considérant 12 bis bis' (No 9029/22, 13 May 2022) arts 4a, 4b, 4c.

4 Council, '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 - General approach' (No 14954/22, 25 November 2022) art 4b(1).

5 OpenAI, 'Introducing ChatGPT' (*OpenAI*, 30 November 2022) <https://openai.com/index/chatgpt/> accessed 7 January 2025.

6 Gian Volpicelli, 'ChatGPT Broke the EU Plan to Regulate AI' (*POLITICO*, 3 March 2023) <https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/> accessed 7 January 2025.

7 Parliament, 'Amendments adopted by the European Parliament on 14 June 2023 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 Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))' (P9_TA(2023)0236, 14 June 2023) art 28b.

8 Felix Duffy, 'AI Act: Von Der KI-Industrie in Die Zange Genommen' (*LobbyControl*, 12 March 2024) <https://www.lobbycontrol.de/macht-der-digitalkonzerne/ai-act-von-der-ki-industrie-in-die-zange-genommen-114508/> accessed 19 January 2025; Olivier Petitjean, 'AI Act : le troublant lobbying des « champions » européens, Mistral AI et Aleph Alpha' (*Observatoire des multinationales*, 11 March 2024) <https://multinationales.org/fr/actualites/mistralai-alephalpha-gafam-ai-europe> accessed 19 January 2025.

'a lot of work resulting in significant improvements'.[9]

The GPAI Code of Practice was published on 10 July 2025. It is a voluntary code, envisaged as a transitional instrument in the absence of standards created by private standards organisations, that helps demonstrate (but does not assure) compliance with the GPAI provisions. It was developed through a multi-stakeholder process consisting of largely self-nominated experts organised into working groups. The Code is accompanied by European Commission guidelines clarifying key concepts related to GPAI models.[10]

The Code of Practice is structured into three chapters: (i) transparency;[11] (ii) copyright;[12] and (iii) safety and security.[13] The chapters on transparency and copyright provide all GPAI model providers with a framework to demonstrate compliance with their obligations under Article 53. The Code's chapter on safety and security applies only to the obligations of a limited subset of providers of the most advanced models, namely those subject to the additional obligations governing providers of GPAI models with systemic risk under Article 55. In this chapter, we discuss only the obligations that apply across GPAI models — systemic risk obligations are discussed elsewhere in this volume.[14] We refer to the Code as appropriate throughout this chapter, while acknowledging that as the Code is only a voluntary text, the most significant contestation will likely arise instead from statutory interpretation.

We proceed as follows. First, we briefly consider the **definition** of GPAI models. Second, we introduce and analyse the obligations relating to **information and documentation**. Third, we consider the obligations relating to **copyright and related rights**, including the interaction with text and data mining. Fourth, we look at the (many) approaches possible to **demonstrate compliance** with this part of the regime. Fifth, we consider the wide array of **exclusions** present, reading this chapter in the context of the rest of the AI Act. We end with some overarching, concluding remarks.

---

9 Alexandre Piquard, 'Arthur Mensch, de Mistral AI : « En intelligence artificielle, nous pouvons rivaliser avec Google ou OpenAI »' (*Le Monde*, 26 February 2024) <https://www.lemonde.fr/economie/article/2024/02/26/mistral-ai-en-ia-nous-pouvons-rivaliser-avec-google-ou-openai_6218645_3234.html> accessed 19 January 2025. Interview with Jonas Andrulis, 'Big Tech Wants to Dominate AI Market: Aleph Alpha CEO' (*Bloomberg*, 24 January 2024) <https://www.bloomberg.com/news/videos/2024-01-24/aleph-alpha-plans-to-raise-more-capital-video> accessed 19 January 2025; Jonas Andrulis [@JonasAndrulis], Unnamed Post (*X*, 27 January 2024) <https://x.com/JonasAndrulis/status/1751234195937018208> accessed 19 January 2025.

10 European Commission, *Annex to the Communication to the Commission: Approval of the content of the draft Communication from the Commission – Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act)*, C(2025) 5045 final, Brussels, 18 July 2025.

11 Nuria Oliver and others, Code of Practice for General-Purpose AI Models, Copyright Chapter (European Commission 2025) (hereafter *GPAI Code of Practice, Transparency Chapter*) <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai> accessed 8 November 2025. There is additionally an attached "Model Documentation Form", found on the same website under this chapter but stored as a Word file, which we will follow the code and refer to as being part of this chapter.

12 Alexander Peukert and others, Code of Practice for General-Purpose AI Models, Copyright Chapter (European Commission 2025) (hereafter *GPAI Code of Practice, Copyright Chapter*) <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai> accessed 8 November 2025.

13 Matthias Samwald and others, Code of Practice for General-Purpose AI Models, Safety and Security Chapter (European Commission 2025) (hereafter *GPAI Code of Practice, Safety and Security Chapter*) <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai> accessed 8 November 2025.

14 Chapter X [GPAI systemic risk], this volume.

B Definitions

When stripped of the superfluous aspects, the GPAI model definition is a functional one which brings into scope (i) AI models, which (ii) display significant generality, are (iii) capable of competently performing a wide range of distinct tasks, and which are (iv) made available on the market.[15] The recitals advise that a presumption of generality should be applied for models with over a billion parameters trained on a large amount of data,[16] and give the example of a 'large generative AI model' able to generate content, as such content generation 'can readily accommodate a large range of distinctive tasks'.[17] The one-billion parameter example given encompasses, at the time of writing in 2025, many models which run comfortably on personal consumer devices as well as those which require specialised cloud hardware.[18] The European Commission state that they intend to use a proxy for the parameters and data presumption of training compute floating point operations (FLOPs), with models greater than 10^{23} FLOPs which are able to generate language receiving a presumption of general-purpose AI model.[19] The placing on the market of GPAI systems, effectively software which integrates such a model, triggers obligations for the model if they have not been triggered already.[20] As mentioned above, there is an additional rule-set for GPAI models determined to have systemic risk — these are covered in a separate chapter of this commentary.[21] There are furthermore a variety of exclusions that may be sought and claimed — we discuss these towards the end of this chapter.[22]

C Information and Documentation

I Overview

The obligations imposed on providers of GPAI models to provide information and documentation relating to these models are set forth in Article 53 AI Act. These requirements are split into obligations to produce 'technical documentation' to be given, upon request, to regulators (the AI Office and national competent authorities), and 'documentation' made for and made available to actors integrating GPAI models in the value chain (downstream providers).[23]

Annexes XI and XII of the AI Act specify the information required to be provided to regulators and providers respectively. The Code of Practice provides a modest elaboration of

15 AI Act, art 3(63).
16 AI Act, recital 98.
17 AI Act, recital 99.
18 For example, Stable Diffusion XL, a 2023 image generation model which can easily run on modern devices, has approximately 3.5bn parameters, while Google's Gemma 2 language model, optimised for phones and similar devices, starts at 2.6bn parameters.
19 European Commission (n 10) para 17.
20 AI Act, recital 97.
21 AI Act, arts 3(65), 55. See generally Chapter X [GPAI systemic risk], this volume.
22 See *infra*, 'Exclusions'.
23 AI Act, recital 101, 53(1)(a–b).

these provisions as part of its Model Documentation Form template.[24] The Act notes that the purpose of these provisions is to provide information for compliance with both the AI Act as well as 'other regulations'.[25] The Commission has the power to amend these annexes in light of new technological developments for a five-year period — a power renewed thereafter as long as the Parliament or Council do not object.[26] The Commission is also specifically empowered to use delegated acts to specify methodologies for computation and energy consumption of models.[27]

Table 1 shows the difference in scope at the time of passing of the AI Act based on the annexes.[28] Important to note is that while we can assume that for widely sold systems, information given to providers is *de facto* public (and thus can easily influence other governance mechanisms such as private law), the information only provided to regulators is not obliged to be public. None of this data is systematically required to be indexed in a public database, unlike the information provided to deployers in the high-risk regime, which must be.[29]

In the sections that follow, we will not go line-by-line through the requirements in Table 1. Some are self-evident, as indicated by the Code of Practice which in its template Model Documentation Form can do little more than repeat them verbatim or add fairly obvious details. In places, the Model Documentation Form proposes a more limited interpretation of the statute. This may have a practical effect on regulatory enforcement or political will, but like the rest of the Code of Practice, it has no legal effect.[30]

Below, we focus on the areas raised by these documentation requirements which have legally complex interpretations or interactions with the rest of the Act.

II Nature of the documentation

(a) Level of detail

The value chain obligations in the AI Act specify a floor for the level of detail in the documentation to providers. These obligations are designed to facilitate co-operation between any provider of components to high-risk AI systems by written agreement, but the quality conditions that apply to such information can be applied to GPAI model documentation, which the legislator has indicated is so important as a form of this value chain as to be provided *ex ante* without written prompting. Such documentation must 'specify the necessary information, capabilities, technical access and other assistance based on the generally acknowledged state of the art, in order to enable the provider of the high-risk AI system to fully comply with the obligations set out in this Regulation'.[31] This requires GPAI model providers to have

[24] GPAI Code of Practice, Transparency Chapter, annex (hereafter *Model Documentation Form*).
[25] AI Act, recital 101.
[26] AI Act, arts 53(5), 97(2).
[27] AI Act, art 53(5).
[28] Annex XI supports the obligation in AI Act, art 53(1)(a); Annex XII supports the obligation in AI Act, art 53(1)(b)(ii).
[29] AI Act, art 71. Chapter X [database], this volume.
[30] See *infra*, 'Demonstrating Compliance'.
[31] AI Act, art 25(4).

Provision		Scope Regulators (R) Providers (P)
the tasks that the model is intended to perform and the type and nature of AI systems in which it can be integrated		R, P
the acceptable use policies applicable		R, P
the date of release and methods of distribution		R, P
how the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, where applicable		P
the versions of relevant software related to the use of the GPAI model, where applicable		P
the architecture and number of parameters		R, P
the modality (eg text, image) and format of inputs and outputs		R, P
the licence for the model		R, P
the technical means (eg instructions for use, infrastructure, tools) required for the GPAI model to be integrated into AI systems		R, P
the design specifications of the model and training process, including training methodologies and techniques, the key design choices including the rationale and assumptions made; what the model is designed to optimise for and the relevance of the different parameters, as applicable		R
information on the data used for training, testing and validation, where applicable, including	the type and provenance of data and curation methodologies (eg cleaning, filtering, etc.),	R, P
	the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases, where applicable	R
the computational resources used to train the model (eg number of floating point operations), training time, and other relevant details related to the training		R
known or estimated energy consumption of the model (where the energy consumption of the model is unknown, the energy consumption may be based on information about computational resources used.)		R

Table 1: Comparison of information requirement provisions under Article 53(1)(a-b), including Annexes XI and XII.

reference to all the high-risk AI requirements when developing this documentation, and to envisage how they could be complied with in light of the information GPAI model providers provide, unless all high-risk applications are entirely foreclosed by the GPAI model provider. Downstream providers may still make use of the value chain obligation to write to the GPAI model providers where the documentation remains insufficient for their compliance needs.[32]

[32] See generally Chapter X [value chain obligations], this volume.

(b) Documentation updates

Providers must keep the documentation 'up-to-date'.[33] The Code of Practice fails to provide guidance on what this obligation means, other than stating there must be a 'relevant' change.[34] When a new model is released, when new acceptable use policies change, when distribution methods change, or when the delegated act requirements on, for example, energy measurement, change, documentation must be updated.[35] Documentation to providers should also be changed when external features that affect the validity of the documentation change, such as when the 'generally acknowledged state of the art' changes, for example where new information on the interaction of a model with other components becomes known.[36]

(c) Retention

Where held by authorised representatives of GPAI model providers, such documentation must be retained for ten years.[37] No explicit time limit is laid out where authorised representatives are not used (eg for Union–established GPAI model providers), which appears to be a drafting omission as a ten-year time limit applies to providers, importers, and authorised representatives of providers of high-risk AI systems.[38] The Code of Practice provides for ten years for all signatories.[39]

III Tasks and integration potential

To what extent can a GPAI model be intended for a high-risk system, but wrapped thinly into a system which manages to shed this intent? In other words, what role does the required documentation of the GPAI model's play in the determination of that system as high-risk or not? The majority of Annex III high-risk AI systems only fall into these categories where they are *intended* for use in this area by the provider.[40] However, note that many GPAI models can be turned into systems with only a very thin and simplistic computational wrapper to animate and query the model, which even can be done on users' own devices.[41]

This relates more broadly to the extent to which (and with what measures) an AI system provider should be *excluding* (for example, contractually) specific high-risk use of a system. The AI Act foresees regulatory action at this area, providing that where market surveillance authorities believe that a GPAI system (including a model and a system released in parallel) 'can be used' by deployers for a high-risk purpose, it may require investigation under both the

33 AI Act, art 53(1–2).
34 GPAI Code of Practice, Transparency Chapter, 5.
35 It is possible that a delegated act would attempt to explicitly avoid retroactive updating of model documentation; we cannot comment on the validity of such a situation without seeing a specific instrument.
36 AI Act, art 25(4).
37 AI Act, art 54(3)(b).
38 AI Act, art 18, 22(3)(b), 23(5).
39 GPAI Code of Practice, Transparency Chapter, 5.
40 AI Act, annex III; see generally Chapter X [high risk classification], this volume.
41 See eg the *Ollama* system.

GPAI regime and the high-risk AI regime and thus requires regulatory cooperation.[42] Tasks described in the documentation are likely to play a role in these considerations. Relatedly, when assessing whether a purpose is or is not intended, the AI Act instructs regulators to consider the 'context and conditions of use'.[43] This will also include the task list described in the GPAI model documentation.[44] The Code of Practice contains no consideration of this issue.

IV Data

Information on utilised data has become an important matter of secrecy in GPAI model training at the time of writing, for entwined reasons of competitive advantage from using specific data sources or methods and risk of liability from revealing that data used is protected by copyright or subject to other claims arising from regimes such as data protection or *sui generis* database protection.

(a) Provenance

Obligations around the provenance of data are two-tiered, both related to transparency. Downstream actors should be provided information concerning the provenance of data used, while regulators should be additionally provided information on 'how the data was obtained and selected'. Given that the GPAI model provider is already obliged to produce 'a sufficiently detailed summary about the content used for training' in all situations – according to a predefined template[45] – the provenance obligation should be seen to both complement this, as well as draw our attention to the specificity of the term *data provenance* in this context (used in Annexes XI and XII[46]), which has a distinct meaning from a 'detailed summary about the content'.

Data provenance is a term of art in computer science, referring to data's lineage: the relationships amongst all the elements, including sources, processing steps, contextual information and dependencies, that contributed to the existence of a piece of that data.[47] It refers to operational steps to understand the lineage of each data point, and therefore this duty cannot be discharged simply by stating generically that the data was, for example, scraped from the web. Despite being designed in part by computer scientists who have, in other situations, been very sensitive to whether legal definitions connect to and reflect computer science definitions, and the obligation on providers to interpret the terms in light of the technical state-of-the-art, the Code of Practice incorrectly characterises data provenance

42 AI Act, art 75(2).

43 AI Act, art 3(12).

44 For clarification, this issue is separate from the potential classification of a regular GPAI model as a GPAI model with systemic risks, which follows from the rules of AI Act, art 51 and annex XIII. See generally Chapter X [GPAI systemic risk], this volume.

45 European Commission, *Explanatory Notice and Template for the Public Summary of Training Content for General-Purpose AI Models required by Article 53(1)(d) of Regulation (EU) 2024/1689 (AI Act)*, C(2025) 5235 final, Brussels, 24 July 2025.

46 AI Act, annex XI, para 2(c); annex XII, para 2(c).

47 Lucian Carata and others, 'A Primer on Provenance' (2014) 57 Communications of the ACM 52, 52.

as being a tickbox choice between 'web crawling', 'private non-publicly available datasets obtained from third parties', 'user data', 'publicly available datasets', 'synthetic data that is not publicly accessible (when created directly by or on behalf of the provider)' and 'data collected through any other means [with a small text box]'.[48] This is not to say that provenance would have to be analysed per data point — the obligations are designed to be 'proportionate',[49] which indicates data with similar provenance should be grouped and have the characteristics of their provenance described. This would include, for example, the dates it was obtained (including originally by any third parties, to emphasise lineage), as well as the broad methods, such as by Web scraper, synthetic generation or model distillation, or similar.

As a result of this, it appears that the Code of Practice will be insufficient for GPAI providers to discharge the data provenance obligation in the AI Act.

Furthermore, additional information must be provided to regulators upon request relating to 'how the data was obtained and selected'. For a web-scraper, for example, this would include, *inter alia*, identification details of the scraper(s) used, such as for Web scrapers, their IP addresses and HTTP headers, which identify them to servers and allow servers to customise their response, and any logic of the systems scraping the data which might change how the data is 'selected', such as its approach towards different opt-out protocols such as *robots.txt* and those relevant for purposes of compliance with the copyright related obligation under Article 53(1)(c). The Code of Practice clarifies that this should include annotation information, synthetic data generation, and contextual information on data acquisition from third parties, but unhelpfully provides no guidance or analysis concerning how to discharge the obligation in the case of many other extremely common formats, including the types they have listed under provenance, including 'web crawling' and 'user data'.[50]

(b) Number of data points

A transparency obligation on the 'number of data points' will not always be so meaningful for language models, as such systems convert the human-interpretable input of text into data and then analyse them in relation to each other. This will need to be interpreted appropriately, such as by representing a breakdown of tokens, sentences and documents, to provide a meaningful and comparable description of the data used. Similar arguments can be made for other models, such as video or audio generation models. Scope and main characteristics will vary depending on the data sources, but using language as an example, obvious candidates would be the topics of the text and proportions of languages in the training dataset. The Code of Practice authors escape providing a recommendation by allowing providers to choose however they would like to interpret the law in this situation, so long as they name the unit they use.[51] This means that this data is not comparable, and can be selectively presented by regulated entities.

---

48 Model Documentation Form, 1.
49 AI Act, recital 101.
50 Model Documentation Form, 2.
51 Model Documentation Form, 2.

(c) Other measures and methodologies

Providers in their documentation should be provided with 'curation methodologies'[52] for the data. Examples of this would be cleaning and filtering. The use of methodology implies information about the process itself. For example, was it automated or manual? Using which systems, or which sources of labour? The Code of Practice however instructs providers only to provide a 'general description of the data processing' in this case, which is not a definition of 'methodology'.[53] They recommend only 300 words, where most GPAI models will use many methodologies at different stages and for different subsets of the data. It is unclear whether this should meet the bar for compliance.

Even more specific is the inclusion of 'all other' measures used to detect either data source unsuitability or identifiable biases in the information for regulators. This refers to an *exhaustive* list of measures taken and would include both technical measures as well as social or organisational measures. The term 'detection' is distinct from 'curation'. 'Curation' implies some action; detection measures need not have resulted in the data being removed or its impact being mitigated, simply being diagnosed and, for the purposes of compliance, documented. The Code of Practice removes the word 'all', which is directly *contra legum,* and leaves it almost certain that signatories will be non-compliant by following the Code alone.[54]

(d) Interaction with other regimes

The extent to which information can be hidden by GPAI model providers from downstream providers and only made available to regulators (or not at all) should be considered in light of the 'other regulations' supposed to be supported by the AI Act's transparency measures.[55] For example, much data used by GPAI model providers is likely to be personal data under EU data protection law,[56] and where this data does not come from data subjects directly, such as through web-scraping, information must either be given to data subjects concerning the source of this data,[57] or communicated publicly.[58] Where such obligations on public communication have to be fulfilled, it makes it difficult to argue that documentation requirements have to be limited for reasons of, for example, trade secrecy.

V Availability and confidentiality

Information prepared only for regulators may not be available to the public, but its existence may nonetheless have impact. This information will likely become an easy target of international civil disclosure or discovery processes.

---

52 AI Act, annex XI, para 2(c); annex XII, para 2(c).
53 Model Documentation Form, 2.
54 Model Documentation Form, 2.
55 AI Act, recital 101.
56 cf, in relation to web-scraping, Case C-101/01 *Lindqvist* ECLI:EU:C:2003:596 [25]; Case C-131/12 *Google Spain* ECLI:EU:C:2014:317 [26].
57 GDPR, art 14(2)(f).
58 GDPR, art 14(5).

Information to providers must be available to those 'providers of AI systems who intend to integrate' the GPAI model into such systems. For the large, platformised, AI-as-a-service providers — such as the hyperscalers[59] — this information will likely be *de facto* publicly available, as anyone can purchase and intend to integrate such services. However, where GPAI model providers are serving smaller and more specific AI system markets, such as in law enforcement, it may be the case that these documents may only be provided directly and may remain confidential.

It is additionally foreseeable that such documents may be subject to non-disclosure agreements. Such agreements would however have to ensure that any integrating high-risk system deployers could reuse information required for compliance with their own downstream obligations under the Act, which may include making parts of the relevant information public.[60]

While Article 53(7) AI Act instructs that information obtained pursuant to Article 53 shall be treated with the confidentiality indicated in Article 78, that provision in turn does not obviously apply to integrators of GPAI models, as it is not clear that as regulated entities they are 'any other natural or legal person involved in the application of this Regulation'. Indeed, Article 78(2) summarises the addressees of this provision as 'authorities'. Following this reading — and unlike authorities — providers are bound only by private law confidentiality obligations, rather than any further regime placed upon them by this Act. The Code of Practice incorrectly claims that Article 78 applies to providers, using the term 'recipients' (which does not appear in the text) expanding the scope of confidentiality in a manner that is not binding.[61] This point is further emphasised by the entire absence of any specified penalty (or route to judicial remedy) for breach of Article 78 in the Act, which would be expected if it placed an obligation upon providers.[62]

A key component of the private law approach will be the reliance on trade secrets protection under the national implementations of the Trade Secrets Directive,[63] which establishes a set of rules to facilitate access to comparable civil action across the EU in the case of trade secret misappropriation. These rules are applicable when the appropriation, use or disclosure of trade secrets occurs without the consent of the trade secret holder, and through the use of dishonest means, breach of law or breach of contract. Such an approach would not protect documentation as a whole — it would only protect aspects of the documentation which could legitimately be described as a trade secret.

---

59 Hyperscaler AI providers are large cloud service companies that offer massive-scale computing power, AI infrastructure, and machine learning services. Examples include Amazon Web Services, Microsoft Azure, Google Cloud Platform and IBM Cloud.

60 AI Act, art 25(4).

61 GPAI Code of Practice, Transparency Chapter, 2, 6.

62 AI Act, art 99.

63 Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure OJ L 157/1.

D Copyright and Related Rights

I General issues in copyright and AI

(a) Locations of challenges

Copyright issues arise at various stages of the AI lifecycle or value chain. This includes the input or training data stage, the AI model itself, and the outputs generated by or with the assistance of an AI model or system.[64] We provide a brief summary of the field below to orient readers.

From the **input perspective**, training and developing AI models involves activities such as web scraping, pre-training, and training, which may involve copyright-relevant reproductions and fall under the broad legal definition of text and data mining (TDM) in the Copyright in the Digital Single Market Directive (CDSM Directive).[65] Most copyright–relevant provisions of the AI Act relate to this input or training stage, as do the relevant measures in the copyright chapter in the GPAI Model Code of Practice and copyright references in the template for the public summary of training content for GPAI models.

From the **model perspective**, key questions centre on the legal status of model weights as protected databases,[66] and whether sufficient memorisation of copyrighted content within a model constitutes unauthorised reproduction.[67]

From the **output perspective**, critical questions include whether AI-generated outputs qualify for copyright protection, constitute derivative works, or infringe upon third-party works used in training.[68] In the EU, a notable and evolving issue is the extent to which

---

64 Andrés Guadamuz, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73 GRUR International 111; EUIPO, *Development of Generative Artificial Intelligence from a Copyright Perspective* (EUIPO 2025) <https://doi.org/10.2814/3893780> accessed 21 May 2025; Nicola Lucchi, 'Generative AI and Copyright - Training, Creation, Regulation' (European Parliament (JURI) 2025) <https://www.europarl.europa.eu/thinktank/en/document/IUST_STU(2025)774095> accessed 29 September 2025; João Pedro Quintais, 'Generative AI, Copyright and the AI Act' (2025) 56 Computer Law & Security Review 106107.

65 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSMD) art 2(2).

66 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases OJ L 77/20.

67 A Feder Cooper and others, 'Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models' (arXiv, 18 May 2025) <http://arxiv.org/abs/2505.12546> accessed 21 May 2025; A Feder Cooper and James Grimmelmann, 'The Files Are in the Computer: On Copyright, Memorization, and Generative AI' (arXiv, 18 July 2024) <http://arxiv.org/abs/2404.12590> accessed 31 July 2024; EUIPO (n 49); Ivo Emanuilov and Thomas Margoni, 'Forget Me Not: Memorisation in Generative Sequence Models Trained on Open Source Licensed Code' (Zenodo, February 2024) <https://zenodo.org/records/10635479/preview/GenAI_Memorisation_Open_Source_IEmanuilov_TMargoni-Preprint_Zenodo.pdf> accessed 31 July 2024; Nuno Sousa e Silva, 'Are AI Models' Weights Protected Databases?' (*Kluwer Copyright Blog*, 18 January 2024) <https://copyrightblog.kluweriplaw.com/2024/01/18/are-ai-models-weights-protected-databases/> accessed 31 July 2024. At time of writing there were already two divergent lower level cases in European courts on this matter. Contrast *Getty Images v Stability AI* [2025] EWHC 2863 (Ch) (Bus and Prop Ct, Intellectual Property List (ChD), 4 November 2025), where the court rejects that the AI model in question stores or reproduce any copyright works, with *Gema v Open AI*, Regional Court of Munich I, 11 November 2025, Case No 42 O 14139/24, where the court ruled that, for copyright purposes, the AI model in question had memorised song lyrics, and that these lyrics were reproducibly fixed and embodied in the model's parameters, amounting to an unauthorised reproduction.

68 The latter issue, often linked to memorization, should instead be viewed from the perspective of regurgitation, extraction, and reconstruction during output generation, as highlighted by Cooper and Grimmelmann (n 67); Matthew Sag, 'Fairness

copyright exceptions apply to AI-generated outputs, including exceptions for freedom of expression, transformative use, quotation, criticism, review, caricature, parody, and pastiche.[69]

Cross-cutting concerns are also relevant and bring additional complexities. One relates to the interplay between copyright law and the private ordering of AI outputs through service providers' terms and conditions, which can affect authorship, ownership, and enforcement.[70] Another relates to how the territoriality principle influences the legal assessment of different copyright law issues.[71] Finally, a pressing policy concern is how creators should be remunerated for the use of their works in generative AI models.[72]

(b) Distinct legal backdrops

The AI Act and copyright law operate on fundamentally different principles, making their interaction complex. The AI Act is primarily a public law instrument, designed through a product safety lens, imposing systemic compliance obligations on certain providers. In contrast, copyright law is a private law framework that grants rights holders exclusive (or remuneration) rights. Copyright is further considered a fundamental right under the Charter.[73] These distinctions lead to differing approaches to enforcement and remedies, and introducing copyright into this frame has created a 'fusion' of two distinct legal approaches.[74]

Moreover, the relationship between the AI Act and EU copyright law is unclear. The AI Act acknowledges that it is wading into a significant tension, noting that generative AI models 'present unique innovation opportunities but also challenges to artists, authors, and other creators' in how their content is 'created, distributed, used, and consumed'.[75] Its recitals state that compliance is 'without prejudice to Union copyright law' and that the regulation 'does not affect the enforcement of [EU] copyright rules',[76] yet such an exclusion is missing from the Act's binding articles, unlike similar exclusions of intermediary liability law.[77] This ambiguity is further complicated by the interplay between the AI Act and the TDM exceptions in the CDSM Directive. Since the Directive allows for varying levels of implementation across EU member states, national differences in copyright law add another

and Fair Use in Generative AI' (2024) 92 Fordham Law Review 1887.

69 NB that the upcoming Court of Justice of the European Union (CJEU) judgment in Pelham II may further define 'pastiche' as an autonomous EU legal concept, potentially impacting the treatment of AI outputs. See Case C-590/23 *CG and YN v Pelham GmbH, SD and UP* ECLI:EU:C:2025:452 (Opinion of AG Emiliou).

70 See eg Oleksandr Bulayenko and others, 'AI Music Outputs: Challenges to the Copyright Legal Framework' (reCreating Europe Reports 4072806 2022) <https://papers.ssrn.com/abstract=4072806> accessed 16 October 2024; Lilian Edwards and others, 'Private Ordering and Generative AI What Can We Learn from Model Terms and Conditions?' in Mimi Zou and others (eds), The Cambridge Handbook of Generative AI and the Law (Cambridge University Press 2025).

71 João Pedro Quintais, 'Copyright, the AI Act and Extraterritoriality' (The Lisbon Council 2025) Policy Brief <https://papers.ssrn.com/abstract=5316132> accessed 23 October 2025.

72 Christophe Geiger and Vincenzo Iaia, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' (2024) 52 Computer Law & Security Review 105925; Martin Senftleben, 'Generative AI and Author Remuneration' (2023) 54 IIC - International Review of Intellectual Property and Competition Law 1535; João Pedro Quintais, 'Generative AI, Copyright and the AI Act' (2025) 56 Computer Law & Security Review 106107.

73 Charter, art 17(2).

74 Alexander Peukert, 'Copyright in the Artificial Intelligence Act – A Primer' (2024) 73 GRUR International 497.

75 AI Act, recital 105.

76 AI Act, recitals 108–9.

77 AI Act, art 2(5).

layer of complexity to their alignment with the AI Act, particularly due to the maximum harmonisation of obligations on GPAI model providers intended by the instrument.[78]

II Text and Data Mining (TDM)

(a) Basic regime

TDM is defined in the CDSM Directive as 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends, and correlations'.[79] The TDM regime comprises of two mandatory copyright exceptions established in the CDSM Directive. The first permits TDM conducted for scientific research by 'research organisations' and 'cultural heritage institutions', provided they have lawful access to the works or subject matter, and meet some additional conditions.[80] The second provides an exception for reproductions and extractions of lawfully accessed works/subject matter for the purposes of TDM not limited to a research purpose or organisational type; however this exception is subject to 'reservation' by rights holders.[81] This reservation, commonly referred to as an opt-out requirement, includes through 'machine-readable means in the case of content made publicly available online', for instance through the use of metadata and terms and conditions of a website or a service. Opting-out does not affect the ability of a relevant research or heritage organisations to claim the first TDM exception.[82]

The AI Act views the development of GPAI models as inherently involving acts of TDM that, when relating to copyright–protected content (as is generally the case), must comply with the conditions in the CDSM Directive and where appropriate, the rights reservation mechanism therein.[83] Not all scholars agree with this view. Some argue that all copyright–relevant reproductions and extractions during AI training qualify as TDM,[84] while others exclude certain acts of reproduction taking place during the training stage.[85] Some suggest AI training activities fall or should fall entirely outside the TDM definition altogether.[86]

---

78 AI Act, art 1(2)(e), recital 1.

79 CDSM Directive, art 2(2).

80 CDSM Directive, art 3; Thomas Margoni and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71 GRUR International 685.

81 CDSM Directive, art 4.

82 Rossana Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 European Intellectual Property Review 322; Péter Mezei, 'A Saviour or a Dead End? Reservation of Rights in the Age of Generative AI' (2024) 46 European Intellectual Property Review <https://papers.ssrn.com/abstract=4695119> accessed 25 June 2024; EUIPO (n 64).

83 ibid.

84 Giuseppe B Abbamonte, 'The Application of the Copyright TDM Exceptions and Transparency Requirements in the AI Act to the Training of Generative AI' (2024) 46 European Intellectual Property Review 479.

85 Eleonora Rosati, 'Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law' [2024] European Journal of Risk Regulation <https://doi.org/10.1017/err.2024.72>. See also, with a similar conclusion for other research exceptions in the copyright acquis and national law, Eleonora Rosati, 'Copyright Exceptions and Fair Use Defences for AI Training Done for "Research" and "Learning," or the Inescapable Licensing Horizon' [2025] European Journal of Risk Regulation 1.

86 Tim W Dornis, 'The Training of Generative AI Is Not Text and Data Mining' [2025] European Intellectual Property Review <https://papers.ssrn.com/abstract=4993782> accessed 13 January 2025. This view is however difficult to reconcile with EU

The role of TDM in the interaction of AI and copyright does, however, have clear limits. TDM exceptions do not cover subsequent acts of communication to the public or making available of the results. The scope of these exceptions covers only acts of reproduction and extraction, and the Directive places clear boundaries on the subsequent use of protected content made pursuant to the TDM exceptions.[87] The exceptions do not cover the making available to the public of datasets resulting from TDM that include copies of protected content (eg a training dataset), or even making any copies during the TDM process that fail to meet the requirements of the exception, which concern security and necessity, and differ by exception type.[88]

(b) Copyright–related obligations in the AI Act

Against this backdrop, the AI Act establishes two key copyright-related obligations for all GPAI model providers. Firstly, providers must implement a 'policy' to ensure compliance with EU copyright law.[89] In particular, they must identify and respect, including through state-of-the-art technologies, reservations of rights (opt-outs) expressed under Article 4(3) of the CDSM Directive. Secondly, providers must draft and publicly disclose a sufficiently detailed summary of the content used to train their GPAI models. This summary must follow a template provided by the AI Office.[90]

Before analysing each obligation in detail, it is important to highlight some common features.

(c) Value chains and model integration

Firstly, because these obligations apply specifically to GPAI *model* providers, it is not always straightforward to understand when they might extend to other players in the AI value chain. Upstream, we find actors responsible for web scraping, crawling, training data supply or brokerage. Downstream, we find actors integrating GPAI models into AI systems, potentially scraping or crawling further information as a part of processing queries (eg through retrieval–augmented generation).

Vertically integrated providers cannot escape the obligations for models by only placing them on the market within an AI system. The obligations will apply to the embedded model in that context.[91] However, it is much less clear where the boundaries of these obligations end when models are very deeply integrated into a system. The lines between these upstream or downstream activities and the models entangled with them, and the systems

legal provisions. See also Lucchi (n 64). (not clearly committing to this view but advancing a purposive interpretation to this effect).

87 CDSM Directive, arts 3(2), 4(2).

88 ibid.

89 AI Act, art 53(1)(c).

90 AI Act, art 53(1)(d).

91 AI Act, recital 97. See also Peukert (n 74); João Pedro Quintais, 'Generative AI, Copyright and the AI Act' (2025) 56 Computer Law & Security Review 106107.

or upstream scraping may not always be technologically clear, and may become more unclear as technological development continues. For example, what if an AI system which queries the web at runtime additionally retains this data for future, routine model training and finetuning? What if future architectures emerge which blur the concepts of training and querying? Where the model and the system are not cleanly separable — when training and deployment does not proceed in discrete, linear phases — to remain effective the AI Act would have to apply model requirements to the entire system. AI system providers that do not also provide the internal models, or retrain or finetune existing models as part of system functionality, in contrast, would not be subject to GPAI model requirements commensurate with the changes applied.[92]

Upstream providers that engage in TDM but do not qualify as GPAI model providers may still avail themselves of the TDM opportunities in the CDSM Directive — as long as they meet their conditions, which in relation to retention and security, they may find difficult as part of a supply chain. They are, however, generally excluded from the obligations set forth in the AI Act.

(d) Distinct regimes for non-compliance

Secondly, as the AI Act operates independently from the CDSM Directive, GPAI model providers' failure to comply with AI Act obligations does not automatically constitute copyright infringement. Non-compliance may result in administrative fines under the Act. Nevertheless, under certain conditions, such non-compliance could play a decisive role in determining whether a provider fails to meet the requirements of the TDM exceptions under the CDSM Directive, potentially leading to a finding of copyright infringement. This interaction is revisited below.[93]

(e) Wide scope

Thirdly, the copyright obligations under the AI Act have broad applicability, overriding some of the Regulation's exclusions. These exclusions are discussed further below.[94] Specifically, they apply even to GPAI models released under free and open-source licenses and cover all model types and sizes.[95] Compliance with these obligations should be commensurate with the nature of the model provider and allow simplified compliance pathways for SMEs, including startups. For instance, in cases of model modification or fine-tuning, the relevant obligations are limited to those modifications or adjustments, such as by supplementing or updating existing documentation.[96]

---

92 AI Act, recital 109.
93 See *infra*, 'Policies to respect'.
94 See *infra*, 'Exclusions'.
95 AI Act recital 104, art 53(2). Note the exceptions for paragraphs (c) and (d).
96 AI Act, recital 109.

III Policies to respect copyright

The AI Act mandates that GPAI model providers establish a policy to respect EU copyright law. However, it is unclear exactly what this obligation constitutes. Is it a strict requirement to ensure compliance? A lower requirement for providers to adopt reasonable efforts towards compliance? Only a minimum formal documentation requirement to have a copyright *policy* in place? While the precise interpretation remains uncertain, the recitals appear to rule out a superficial, documentation–focussed reading of the obligation, noting the copyright standards that the Act requires that providers should be 'applying', and emphasising compliance with the reservation of rights in particular.[97]

This also seems to be the reading advanced by the copyright chapter in the GPAI Code of Practice. This document states that its signatories must draft, regularly update, and *implement* a single, consolidated copyright policy that describes how they comply with relevant legal obligations. They must also assign internal responsibility for overseeing and *executing* this policy. Additionally, signatories are encouraged to make a summary of their copyright policy publicly available and to keep it up to date.[98]

Importantly, the Code's copyright chapter focuses exclusively on this policy obligation, stating that its aim is to 'contribute to the proper application of the obligation laid down in Article 53(1)(c) of the AI Act'. In this respect, a number of additional remarks are justified.

First, the Code's copyright chapter sets out a number of measures for signatories to implement in order to demonstrate compliance with the policies obligation, while noting that adherence to the Code does not *per se* constitute compliance with Union law on copyright and related rights.[99]

Second, this chapter clarifies that it does not affect the interpretation or enforcement of EU copyright and related rights law, which remains under the jurisdiction of national courts and ultimately the CJEU.[100]

Third, the Code's signatories acknowledge that EU copyright law grants preventive, consent-based exclusive rights, subject to exceptions. One such exception is the TDM exception under Article 4(1) of the CDSM Directive, which applies only when users have lawful access and rights have not been expressly reserved by rightsholders under Article 4(3).[101]

Fourth, 'commitments' requiring proportionate measures must be scaled appropriately to the size of providers, with special consideration for the interests of SMEs and startups.[102]

Fifth, the chapter does not interfere with existing agreements between signatories and rightsholders that authorize the use of protected works.[103]

---

97 AI Act, recital 106.
98 GPAI Code of Practice, Copyright Chapter, measure 1.1.
99 GPAI Code of Practice, Copyright Chapter, recital (a).
100 GPAI Code of Practice, Copyright Chapter, recital (b).
101 GPAI Code of Practice, Copyright Chapter, recital (c).
102 GPAI Code of Practice, Copyright Chapter, recital (d).
103 GPAI Code of Practice, Copyright Chapter, recital (e).

Sixth, the commitments to show compliance with the policies obligation in Article 53(1)(c) of the AI Act are considered to complement providers' separate obligation under Article 53(1)(d) to publish detailed summaries of the training data used for GPAI models, following an AI Office template.[104] We return to this point on the transparency template below.[105]

(a) Copyright and model outputs

This obligation broadly pertains to compliance with EU copyright law, meaning it is not necessarily confined to the input or training stage of AI models. The policy obligation could extend to the moderation of model outputs.[106] Microsoft, for example, deploys an explicitly 'layered approach' to AI content moderation, *including* mitigations at the model layer alongside mitigations within the systems that surround them.[107] The AI Act is silent on the distribution of responsibilities for mitigation at the output layer between model providers and system providers, but it would appear that, insofar as model outputs are relevant to EU copyright law, the AI Act includes an obligation to include consideration of them in the policy. This has raised concerns of a risk of overzealous blocking of outputs in the model output, disregarding legitimate exemptions in the copyright regime such as parody, education or critique,[108] in ways which both mirror the concerns around so-called 'upload filters' in the CDSM Directive,[109] as well as much older debates around how technical protection mechanisms override the careful societal balances struck in intellectual property law.[110] Such an evolution of third-party and proprietary tools mirrors in the AI context developments in content filtering for online platforms—such as YouTube's *Content ID* and independent tools like *Pex*. Notably, like for content filtering tools by platforms, these AI opt-out tools could eventually be extended beyond training-stage compliance to downstream content moderation, helping enforce copyright in AI-generated outputs.[111] This development, along with its potential chilling effects upon freedom of expression, is visible in the GPAI Code of Practice's copyright chapter, which contains a specific measure to mitigate the risk of copyright-infringing outputs by downstream AI systems that integrate GPAI models.[112]

In essence, the Code's signatories commit to implementing two types of measures. The

---

104 GPAI Code of Practice, Copyright Chapter, recital (f).

105 See *infra*, 'Training data summary'.

106 Peukert (n 74); Quintais (n 91). This is suggested for example at GPAI Code of Practice, Copyright Chapter, measure 1.4 ('Mitigate the risk of copyright-infringing outputs').

107 Azure AI Foundry, 'Responsible AI for Azure AI Foundry' (*Microsoft*, 20 February 2025) <https://learn.microsoft.com/en-us/azure/ai-studio/responsible-use-of-ai-overview> accessed 23 February 2025.

108 Caroline De Cock, 'From Upload to Output Filters? How the AI Act's Code of Practice Could Threaten Freedom of Expression and Thought' (*Tech Policy Press*, 17 January 2025) <https://techpolicy.press/from-upload-to-output-filters-how-the-ai-acts-code-of-practice-could-threaten-freedom-of-expression-and-thought> accessed 23 February 2025.

109 CDSM Directive, art 17. Note however that to the extent that AI system providers or deployers do not qualify as "online content-sharing services providers" they are not subject to these requirements.

110 Dan L Burk and Julie E Cohen, 'Fair Use Infrastructure for Rights Management Systems' (2001) 15 Harvard Journal of Law & Technology 41.

111 Arvind Narayanan and Sayash Kapoor, 'Generative AI's End-Run around Copyright Won't Be Resolved by the Courts' (*AI Snake Oil*, 20 March 2023) <https://www.aisnakeoil.com/p/generative-ais-end-run-around-copyright> accessed 31 July 2024.

112 GPAI Code of Practice, Copyright Chapter, measure 1.4.

first is to implement 'appropriate and proportionate technical safeguards' that prevent models from generating infringing reproductions of copyrighted training content. The second is to prohibit copyright-infringing uses of their models through their acceptable use policies (or equivalent terms of use or documentation) or, for open-source releases of their GPAI models, clearly warn users against such uses in the accompanying documentation.[113] Crucially, the Code clarifies that this commitment applies whether the Signatory integrates the model into its own AI systems or provides it to another entity under a contractual arrangement.[114]

The blurry distinction between a 'model' and a 'system' discussed above returns with a vengeance when considering output filtering obligations. Many GPAI providers do not release downloadable models with open weights and architectures, but release new models via APIs. However, these differ from models released in an open-source or weight-available manner, as it is rare that it is just the *model* involved in an API query. Instead, APIs that claim to query a model typically query an entire software stack, including systems that classify inputs and outputs against certain content moderation policies to prevent misuse.[115] AI *models* themselves are notoriously difficult to robustly prevent from outputting certain types of content, however when embedded in a system, common techniques of automated content moderation can be added, as flawed as they may be.[116] GPAI model providers may be tempted to try and argue that the concept of model should be interpreted narrowly, and seek to only apply the copyright policy there, considering the model in its content moderation wrapper as a 'system' to attempt to bring it outside of scope and avoid having to apply a copyright policy to the outputs.[117] Such debates will also likely be influenced by the interpretation of the scope of the AI Act's labelling requirements for synthetic content under Article 50(2).[118]

(b) Lawful access

A shared and common requirement of Articles 3 and 4 of the CDSM Directive is that of lawful access. The TDM exceptions apply only to the extent that the materials subject to the restricted acts have been lawfully accessed by the beneficiaries of the exceptions. Crucially, the focus of the legal provisions lies on the lawfulness of the access rather than on the lawfulness of the accessed copy or source.[119]

The concept of lawful access is not defined in the Directive, but some interpretative

[113] GPAI Code of Practice, Copyright Chapter, measure 1.4.

[114] GPAI Code of Practice, Copyright Chapter, measure 1.4, para 2.

[115] See eg Amazon, 'Amazon Bedrock Abuse Detection' (*AWS Amazon Bedrock*) <https://docs.aws.amazon.com/bedrock/latest/userguide/abuse-detection.html> accessed 8 July 2024.

[116] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945.

[117] Note however that the transparency obligation for labelling synthetic content in AI Act, art 50(2) does apply to AI systems rather than to models, which could be read as suggesting that output moderation measures are meant to apply at the system level.

[118] See Chapter X [watermarking and synthetic content, art 50 provisions], this volume. On this topic, see eg Kateryna Militsyna, 'Can Copyright Law Benefit from the Marking Requirement of the AI Act?' [2025] IIC - International Review of Intellectual Property and Competition Law <https://doi.org/10.1007/s40319-025-01624-2> accessed 28 October 2025.

[119] Thomas Margoni, 'Saving Research: Lawful Access to Unlawful Sources under Art. 3 CDSM Directive?' (*Kluwer Copyright Blog*, 22 December 2023) <https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/> accessed 31 July 2024.

guidance can be derived from the recitals. The notion is likewise not clarified elsewhere in the copyright *acquis*, although there are multiple references to it, and to adjacent concepts, in other legal instruments and in the case law of the CJEU.[120] Relevant judgments include those concerning the right of reproduction and its exceptions (*ACI Adam* and *Copydan*),[121] as well as the scope of the right of communication to the public (*Svensson*, *Stichting Brein*, *GS Media*, *Renckhoff*)[122]. Such related notions include 'lawful use', 'lawful user', 'lawful source' and 'legal access', among others.[123] As noted by Margoni, CJEU case law on this matter ranges from a general prohibition of reproductions from unlawful sources (as in *ACI Adam*) to the establishment of a rebuttable presumption of the unlawfulness of the source accessed by for-profit providers (as in *GS Media*).[124]

The preamble of the CDSM Directive identifies four categories of lawful access in relation to Article 3.[125] First, access based on contractual arrangements between rightsholders and research organisations. This includes subscriptions,[126] with the further clarification that persons affiliated with a beneficiary, such as a research organisation or cultural heritage institution, and covered by those subscriptions, should also be deemed to have lawful access (for example, employee researchers or librarians).[127] Second, access based on open access policies. Third, a residual, catch-all category comprising access 'through other lawful means'. Fourth, access to content that is 'freely available online'.

Although the CDSM Directive recitals explicitly link these categories to TDM under Article 3, it is arguable that they should apply *mutatis mutandis* to Article 4 as well. Article 4, for its part, is explicitly supported only by a single recital, which states that the exception should apply only where the work or other subject matter is 'accessed lawfully by the beneficiary, including when it has been made available to the public online'.[128]

There is a significant amount of scholarly discussion surrounding the concept of lawful access.[129] Among that, one of the key points of discussion concerns the fourth category, namely the meaning of lawful access to content that is 'freely available online', a formulation

---

120 Tatiana-Eleni Synodinou, 'Who Is a Lawful User in European Copyright Law? From a Variable Geometry to a Taxonomy of Lawful Use' in Tatiana-Eleni Synodinou and others (eds), *EU Internet Law in the Digital Era: Regulation and Enforcement* (Springer International Publishing 2020) <https://doi.org/10.1007/978-3-030-25579-4_2> accessed 29 October 2025; Tatiana Eleni Synodinou, 'Lawfulness for Users in European Copyright Law: Acquis and Perspectives' (2019) 10 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 20.

121 Case C-435/12 *ACI Adam BV and Others v Stichting de Thuiskopie and Stichting Onderhandelingen Thuiskopie vergoeding* ECLI:EU:C:2014:254; Case C-463/12 *Copydan Båndkopi v Nokia Danmark A/S* ECLI:EU:C:2015:144.

122 Case C-466/12 *Nils Svensson and Others v Retriever Sverige AB* ECLI:EU:C:2014:76; Case C-610/15 *Stichting Brein v Ziggo BV and XS4All Internet BV* ECLI:EU:C:2017:456; Case C-160/15 *GS Media BV v Sanoma Media Netherlands BV and Others* ECLI:EU:C:2016:221; Case C-161/17 *Land Nordrhein-Westfalen v Dirk Renckhoff* ECLI:EU:C:2018:634.

123 Synodinou (n 120); Kacper Szkalej, 'The Paradox of Lawful Text and Data Mining? Some Experiences from the Research Sector and Where We (Should) Go from Here' (2025) 74 GRUR International 307.

124 Margoni (n 119).

125 CDSM Directive, recital 14.

126 CDSM Directive, recitals 10, 14.

127 On the potential challenges this category poses to research organizations, including in the context of AI development, see Szkalej (n 123).

128 CDSM Directive, recital 18.

129 Giorgos Vrakas, 'A Literature Review of "Lawful" Text and Data Mining [Version 1; Peer Review: 3 Approved with Reservations]' (2024) 4:153 Open Research Europe <https://open-research-europe.ec.europa.eu/articles/4-153> accessed 29 October 2025.

that has previously appeared in CJEU case law.[130] As a matter of interpretation, it seems plausible that the scope of this category is broader than that of the preceding ones. However, it remains unclear whether 'freely available online' (in the Article 3 recitals) and 'made available to the public online' (in the Article 4 recital) have identical or distinct meanings.[131]

Still, given the focus of the exceptions on the lawfulness of access rather than on the source itself, the differing types of beneficiaries and purposes of the exceptions in Articles 3 and 4, and the fact that CJEU case law indicates different treatment of commercial and non-commercial users in assessing whether content is 'freely available online', there are strong arguments in favour of adopting a broader interpretation of this category in relation to TDM under Article 3 of the CDSM Directive.[132]

The GPAI Code of Practice's copyright chapter addresses this legal quandary through Measure 1.2, which seeks to ensure that signatories reproduce and extract only lawfully accessible materials in their TDM and AI training web-crawling activities. In this context, signatories make two commitments designed to delineate lawful from unlawful access in relation to web crawling. First, they commit not to circumvent effective technological measures, as defined in the InfoSoc Directive, that prevent or restrict unauthorised acts concerning protected materials, including by respecting technological access controls such as subscription models or paywalls.[133] Second, they commit to exclude from their web crawling any websites that make content publicly available and have been recognised, at the time of crawling, by courts or public authorities in the EU or EEA as repeatedly infringing copyright on a commercial scale. For compliance purposes, a dynamic list linking to official lists of such websites, issued by the relevant EU and EEA bodies, will be made publicly available on an EU website.[134]

(c) Reservation of rights

The explicit emphasis of the policy obligation in Article 53(1)(c) is, however, to ensure the identification and respect of rights reservations (opt-outs) under Article 4(3) of the CDSM Directive using state-of-the-art technologies when building GPAI models.[135] This is also the topic on which the majority of the Measures in the copyright chapter of the Code of Practice focus on.

Several legal ambiguities remain regarding how this applies to GPAI model training, which propagate forwards to what would constitute an adequate policy for the purposes of the AI Act.

---

130 Margoni (n 119); Ula Furgal and others, 'Report on the European Copyright Society's Annual Conference 2025' [2025] GRUR International ikaf119.

131 cf CDSM Directive, recitals 14, 18.

132 Margoni (n 119); Furgal and others (n 130).

133 GPAI Code of Practice, Copyright Chapter, measure 1.2, para 1(a); 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, OJ L 167/10 ('InfoSoc Directive'), art 6(3).

134 GPAI Code of Practice, Copyright Chapter, measure 1.2, para 1(b).

135 AI Act, art 53(1)(c), recital 106.

(d) Appropriate reservation method

A reservation of rights must always be carried out in an 'appropriate manner'.[136] For publicly available online material, the articles of the CDSM Directive indicate that 'machine-readable means' are such an appropriate manner,[137] although the recitals go further to stipulate that such means are the 'only' appropriate method in that situation.[138] Writing in relation to the CDSM Directive rather than the AI Act, Hamman argues that a reservation 'must be explicit (specific for a given content and use) and automatable (employing a well-defined technical protocol)', considering, based in part on definitions of machine-readability in other statutes, that a reservation needs to be structured for a machine to parse it.[139] A range of relevant standards already exist.[140]

The AI Act complicates this issue due to the combination of its inclusion, beyond the explicit text of the CDSM Directive, of an obligation to use 'state-of-the-art technologies' to identify and comply with a reservation of rights, and the example in the recitals of the CDSM Directive of 'terms and conditions' as an example of a machine-readable reservation. Terms and conditions, written in natural language, have not historically been able to explicitly signal anything unambiguously to computers, however typical GPAI model provider will be proficient with some of the most state-of-the-art technologies for machine reading available, in the form of large language models for natural language processing. National courts have been divided as to what extent this means that terms and conditions, typically written for humans, might constitute an effective reservation, albeit in cases not related to GPAI model providers. The Hamburg District Court indicated (in *obiter dicta*) in the *LAION* case that it believes clear entries in terms and conditions could meet the machine readability requirement. It went so far to specifically and explicitly disagree with varied points made by Hamman, and highlighted the importance of having reference to technological developments affecting what machine readability currently means.[141] In contrast, in *HowardsHome*, the Amsterdam District Court appears to endorse the requirement that opt-outs must be at least machine

[136] CDSM Directive, art 4(3).

[137] ibid.

[138] CDSM Directive, recital 18.

[139] Hanjo Hamann, 'Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and Their (In)Compatibility with Article 4(3) of the Copyright DSM Directive' (2024) 15 JIPITEC 102. Hamann draws on definitions of machine readability else in EU law, in particular looking at the Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, OJ L172/56, art 2(13) ("a file format structured so that software applications can easily identify, recognise and extract specific data, including individual statements of fact, and their internal structure").

[140] See generally Hamann (n 139); Paul Keller and Zuzanna Warso, 'Defining Best Practices for Opting out of ML Training' (Open Future Foundation 2023) <https://openfuture.eu/publication/defining-best-practices-for-opting-out-of-ml-training> accessed 31 July 2024.

[141] Landgericht Hamburg, 27 September 2024 (310 O 227/23) ECLI:DE:LGHH:2024:0927.310O227.23.00, paras 100 *et seq*. See generally Paul Goldstein, Christianne Stuetzle and Susan Bischoff, 'Kneschke vs. LAION - Landmark Ruling on TDM Exceptions for AI Training Data – Parts 1 and 2' (*Kluwer Copyright Blog*, 13 November 2024) <https://copyrightblog.kluweriplaw.com/2024/11/13/kneschke-vs-laion-landmark-ruling-on-tdm-exceptions-for-ai-training-data-part-1/> accessed 13 January 2025; Jonathan Pukas and Jan Bernd Nordemann, 'German Regional Court (Landgericht) of Hamburg Paves the Way to Treat the Reproduction of Works as AI Training Data under the EU Text and Data Mining Exceptions' (*Kluwer Copyright Blog*, 25 October 2024) <https://copyrightblog.kluweriplaw.com/2024/10/25/german-regional-court-landgericht-of-hamburg-paves-the-way-to-treat-the-reproduction-of-works-as-ai-training-data-under-the-eu-text-and-data-mining-exceptions/> accessed 25 October 2024.

readable and explicit or express (in the sense of properly targeted). It refused to extrapolate that because the website owner used *robots.txt* to block some specific AI scrapers that they meant to block similar scrapers as well.[142] At least in relation to the AI Act rather than the CDSM Directive, this issue may be somewhat resolved before a CJEU ruling by the adoption of one of many of the compliance demonstration mechanisms in the AI Act specifying the appropriate standard for machine readability, the former of which provides a loose mechanism of demonstrating compliance, and the latter a firm presumption of conformity.[143]

On this topic, the copyright chapter of the Code of Practice sets out a multi-pronged approach in its most detailed provision, Measure 1.3, aimed at identifying and complying with rights reservations when crawling the web.

Firstly, it is important to make an observation on the limited scope of this measure. The commitment in the Code applies *only* to 'machine-readable reservations of rights expressed pursuant to Article 4(3)' of the CDSM Directive.[144] That is to say, the Code's commitments in this respect *do not apply to other non-machine-readable types of reservations*, such as the aforementioned natural language reservations, despite other 'appropriate' forms of reservation being foreseen by the CDSM Directive at least for material other than publicly available online material. Such non-machine-readable reservations might still be considered valid for the purposes of TDM exception and EU copyright law,[145] and no such scope limitation is present in the text of the AI Act. It may be the case that, given this omission and depending on the type of material used in training, relevant providers will need to demonstrate they go beyond the Code in order to fully comply with the Act.

Against this background, signatories make two main commitments. First, they commit to using web crawlers that respect instructions set out under the Robots Exclusion Protocol (robots.txt), as defined by the Internet Engineering Task Force (IETF) and any future technically feasible versions.[146] Second, they commit to identify and comply with other suitable machine-readable protocols that express rights reservations under Article 4(3) CDSM Directive. These may include metadata-based methods (asset- or location-based) that are recognised by international or European standardisation bodies, or are otherwise considered state-of-the-art, technically viable, and widely adopted by rightsholders. Such approaches should reflect sectoral diversity and be developed through inclusive EU-level discussions among rightsholders, AI providers, and other relevant stakeholders, as a practical interim step toward formal standardisation.[147] Signatories are encouraged, but not obliged, to support

---

142 Rechtbank Amsterdam, 30 October 2024 (C/13/737170 / HA ZA 23-690) ECLI:NL:RBAMS:2024:6563, para 4.33. See generally Etienne Valk and Iris Toepoel, 'DPG Media et al vs. HowardsHome – A National Ruling on DSM's Press Publishers' Rights and TDM Exceptions' (*Kluwer Copyright Blog*, 16 January 2025) <https://copyrightblog.kluweriplaw.com/2025/01/16/dpg-media-et-al-vs-howardshome-a-national-ruling-on-dsms-press-publishers-rights-and-tdm-exceptions/> accessed 30 January 2025.

143 See *infra*, 'Demonstrating Compliance'.

144 GPAI Code of Practice, Copyright Chapter, measure 1.3, para 1.

145 GPAI Code of Practice, Copyright Chapter, measure 1.3, para 2.

146 GPAI Code of Practice, Copyright Chapter, measure 1.3, para 2(a). Note that *robots.txt* was an informal standard from the mid-nineties, and only in 2022 was it codified by the IETF, see generally Martijn Koster and others, *Robots Exclusion Protocol* (IETF RFC 9309, Request for Comments 2022) <https://doi.org/10.17487/RFC9309>.https://doi.org/10.17487/RFC9309>.

147 GPAI Code of Practice, Copyright Chapter, measure 1.3, para 1(b).

such efforts.[148]

Furthermore, this measure applies to web crawling both carried out by signatories or third parties acting 'on their behalf'. The AI Act is silent on the interaction between the AI value chain and the scope of the copyright policy, and the Code provides no further instruction in this regard. We can expect this topic to become complicated. For example, many GPAI providers, as well as downstream cleaned databases, rely on versions of the Common Crawl, a large and regularly collected dump of the web originally created for researchers. Common Crawl neither respects *robots.txt* nor scans content after collection, notably even including known child sexual abuse imagery in its dataset.[149] It is unclear to what extent or how far either the AI Act obligations, or the Code's interpretation of them, would spread upstream in the AI value chain.

Signatories also commit to take 'appropriate measures' to enable affected rightsholders to access information about the web crawlers they use, including their robots.txt features and other methods for identifying and complying with rights reservations.[150] Such information may prove important for more than just rights reservation, as model providers have often rotated their scrapers' IPs and obfuscated their client header information in order to avoid increased efforts to block them.[151] Private governance efforts in this area by content distribution networks, particularly Cloudflare, have already focussed on requiring AI scrapers to reveal their identities, backed by a threat of widespread blocking by these internet intermediaries.[152]

(e) Granularity and temporality

Other uncertainties relate to the granularity of the opt-out. Does a reservation applied to the work itself apply universally, or is a separate reservation needed for each digital copy? Should reservations be applied at the level of websites or services, or more specifically in a case-by-case manner to individual pieces of media? This is particularly important given that the contemporary web is characterised by platforms where rightsholders are often individual users, rather than the overarching service operator, yet *robots.txt* is defined at the level of website structure.

A 2025 EUIPO study distinguishes between two types of opt-out measures, namely: 'legally-driven' opt-out measures (eg unilateral declarations by rights holders, licensing

[148] GPAI Code of Practice, Copyright Chapter, measure 1.3, para 3.

[149] David Thiel, 'Identifying and Eliminating CSAM in Generative ML Training Data and Models' (*Stanford Internet Observatory, Cyber Policy Center*, 2023) <https://doi.org/10.25740/kh752sm9123> accessed 12 January 2024 (describing LAION, a machine learning image dataset derived from the Common Crawl).

[150] GPAI Code of Practice, Copyright Chapter, measure 1.3, para 4 (adding that this information should be made publicly available, with an option for automatic notification of updates — eg via a web feed — without prejudice to the right of information under Directive 2004/48/EC, art 8.

[151] See eg Jess Weatherbed, 'Anthropic's Crawler Is Ignoring Websites' Anti-AI Scraping Policies' (*The Verge*, 25 July 2024) <https://www.theverge.com/2024/7/25/24205943/anthropic-ai-web-crawler-claudebot-ifixit-scraping-training-data> accessed 8 October 2024; Elena Yovcheva, 'Why We Blocked Bytespider and ClaudeBot' (*Servebolt*, 27 May 2024) <https://servebolt.com/articles/servebolts-decision-to-block-bytespider-and-claudebot/> accessed 8 October 2024.

[152] Alex Bocharov and others, 'Declare Your AIndependence: Block AI Bots, Scrapers and Crawlers with a Single Click' (*The Cloudflare Blog*, 3 July 2024) <https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click> accessed 8 October 2024.

constraints, and website-level terms and conditions), and 'technically-driven' opt-out measures, including the robots exclusion protocol, TDM reservation protocols, tools adapted for content provenance and authenticity, and broader copyright management solutions still in development.[153]

Our view is that legally-driven measures would hardly meet the requirements of Article 4(3) for TDM of 'publicly available online content'. Technically-driven measures, for their part, are more clearly aligned with the requirements of Article 4(3) for TDM of such content. These can be categorised as either location-based — linked to the specific place where the content is hosted or found online — or asset-based, meaning they are tied directly to the content itself, regardless of its online location.[154] At time of writing, examples of technical measures already in use that serve an opt-out function include opt-out tools by AI model providers like Google (Google-Extended),[155] the TDM Reservation Protocol,[156] and third-party solutions like Spawning AI's tools (eg the Do Not Train Tool Suite).[157] Despite these measures, 'no single opt-out mechanism has emerged as the sole standard used by rights holders'[158] and legal uncertainties remain regarding how the opt-out requirement applies to GPAI model training.[159]

Among those uncertainties is the issue of temporality: it is unclear whether an opt-out can be exercised at any stage during the pre-training and training process (*ex ante* and *ex post* TDM) or only when content is publicly available online and before it is mined (*ex ante*), such as during web scraping.[160] When assessing a policy under the AI Act, the AI Office will be faced with the question of how it deals with each of these challenges, likely without clear guidance from the Court of Justice.[161]

---

153 EUIPO (n 67)). Examples of the latter include the Liccium Trust Engine Infrastructure or Valuenode's Open Rights Data Exchange platform.

154 EUIPO (n 67)).

155 See these and other examples provided in Paul Keller and Zuzanna Warso, Defining Best Practices for Opting out of ML Training (Open Future Foundation, 2023); Quintais (n 72).

156 W3C Community Group, *TDM Reservation Protocol (TDMRep)* (W3C Community Group, 2024).

157 Other third-party solutions, such as Cloudflare's crawl blocker — which allows content owners to block AI crawlers and prevent their works from being mined — perform a similar function but may not be considered opt-out tools within the meaning of CDSM Directive, art 4(3). See EUIPO (n 67See EUIPO (n 67); Quintais (n 72). The EUIPO study considers that these measures fail to meet 'the legal criteria for right reservation'.

158 EUIPO (n 67).

159 Addressing these different questions, see eg Giuseppe B. *Computer Law & Security Review*, vol. 56, 2025, p. 106107.

160 Addressing these different questions, see eg Quintais (n 91); Hamann (n 139); Mezei (n 82); Abbamonte (n 84).

161 nb that at time of writing there is a relevant preliminary reference to the CJEU on issues of TDM and AI. See preliminary reference to the CJEU in case C-250/25, *Like Company v Google Ireland Limited* from the Budapest Környéki Törvényszék (Hungary). For early comment, see Andres Guadamuz, 'First Case on AI and Copyright Referred to the CJEU' (*TechnoLlama*, 27 May 2025) <https://www.technollama.co.uk/first-case-on-ai-and-copyright-referred-to-the-cjeu> accessed 5 June 2025; Paul Keller, 'Do AI Models Dream of Dolphins in Lake Balaton?' (*Kluwer Copyright Blog*, 28 May 2025) <https://copyrightblog.kluweriplaw.com/2025/05/28/do-ai-models-dream-of-dolphins-in-lake-balaton/> accessed 5 June 2025; Tim W Dornis and Nicola Lucchi, 'Generative AI and the Scope of EU Copyright Law: A Doctrinal Analysis in Light of C-250/25' (Social Science Research Network, 14 August 2025) <https://papers.ssrn.com/abstract=5391439> accessed 28 October 2025.

(f) Applicable law

A specific feature of the 'policies to respect copyright' is their potential extraterritorial effect.[162] Recital 106 of the AI Act states that the obligation should apply even if the TDM activities in question take place outside the EU. The reason for this is to prevent regulatory arbitrage by 'ensuring a level playing field' among GPAI model providers, such that 'no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union.'

This stands in stark contrast to the operation of copyright law. It is universally accepted that copyright protection and its attached exclusive rights are granted on a territory-by-territory basis (territoriality principle). Furthermore, in the EU, *lex loci protectionis* (the law of the country for which protection is claimed) is the choice of law rule that applies to copyright infringement, covering both the requirements for protection and that protection's scope.[163] This means that if relevant TDM acts, namely reproductions and extractions, for (pre-)training a GPAI model occur outside the EU, the law applicable is the law of the location of these acts, rather than the location where the model is later made available. If such acts take place in such a third country and there is no relevant point of attachment to an EU member state, then the EU TDM regime (through its national transpositions) is not relevant, regardless of any later placing on the market of the model in the EU.[164]

Some authors qualify this AI Act rule as a 'market entry requirement', derived from the general rule in Article 2(1)(a) of the AI Act. According to this provision, the Regulation applies, *inter alia*, to providers placing GPAI models on the market in the Union, regardless of whether those providers are established or located within the Union or in a third country.[165] This approach follows the Act's product safety logic, as well as the logic of EU law in general in that domain — if a product (model) is placed on the EU market, it must comply with EU law.

Some go even further, arguing that, as a result of this rule, EU copyright directives 'acquire direct effect vis-à-vis private legal entities outside the EU'.[166] However, this interpretation is difficult to defend.

Firstly, recitals have an essentially explanatory function.[167] They have no binding legal force and therefore no ability to create rights or duties. While they are an interpretative aid, they cannot be used to derogate from provisions or interpret them in a manner clearly contrary to their wording.[168] A non-binding recital of the AI Act cannot alter the globally-

---

[162] For detailed analysis on this topic, see Quintais (n 71).

[163] Regulation (EC) No 864/2007 of the European Parliament and of the Council of 11 July 2007 on the law applicable to non-contractual obligations (Rome II) OJ L 199/40, art 8.

[164] Senftleben (n 72) 1548–9.

[165] Abbamonte (n 84).

[166] Malte Stieper and Michael Denga, The international reach of EU copyright through the AI Act (Institut für Wirtschaftsrecht 2024) <https://opendata.uni-halle.de//handle/1981185920/118909> accessed 22 October 2024.

[167] Maarten den Heijer, Teun van Os van den Abeelen and Antanina Maslyka, 'On the Use and Misuse of Recitals in European Union Law' (*Amsterdam Law School Research Paper 2019-31*, 30 August 2019) <https://papers.ssrn.com/abstract=3445372> accessed 23 February 2025.

[168] Case C-136/04, *Deutsches Milch-Kontor* ECLI:EU:C:2005:716 [32]; Case C-134/08, *Tyson Parketthandel* ECLI:EU:C:2009:229

understood, substantive rules of an entirely different regime. At most, it can clarify that the copyright policy obligation in Article 53(1)(c) should apply in the same way that EU law has to be applied by manufacturers outside the Union seeking to access the Union market. Through this lens, the recital could be seen as support for a consistent scope within the product safety paradigm. If the recital succeeds in its expansive efforts, it provides a wider scope of substantive protection to rightsholders, albeit only under the public law, regulatory regime of the AI Act rather than the private law copyright framework.

However, it could also be seen as requiring an interpretation in a manner clearly contrary to the Article's wording. The provision itself only requires compliance with Union law on copyright and related rights. Considering the binding articles alone (ie the enacting terms), the obvious assumption would be that this would be after conflict of laws rules are applied, and within the normal operation of copyright territoriality.

Some scholars have proposed intermediate solutions, such as scoping the law based on whether or not content was hosted on EU servers.[169] This appears to be an attempt to establish a localisation fiction, in which relevant TDM activities are deemed to take place in the EU, thereby aligning with the *lex loci protectionis* principle. But as van Eechoud notes, such localisation fictions typically fail when internet communication is involved.[170] Where they concern the servers being scraped themselves, such fictions fail particularly hard because to increase speed and lower bandwidth burden on the global internet, content is often cached through content delivery networks (commonly called CDNs) — servers close to the user, often in their jurisdiction.[171] The 'uncertain location' of servers and similar intermediary services has left the CJEU unwilling to use their location or their provider's place of establishment as a basis for jurisdictional rules.[172] Insofar as this argument might relate to the servers undertaking the scraping themselves, these may be more fixed and observable. However in the absence of heavily geo-blocked content, this would provide little of the protection that the recital appears to hope for.

Others have suggested we might find a defence for such an effect in international and EU copyright law, in particular considering CJEU case law on infringement localisation. If TDM–related extraction and reproduction are 'functionally' essential to AI model training, and those models are made available in the EU, perhaps applying EU law is justified, as these acts are part of a broader process linked to the EU.[173] The practical problem with this argument is the factual and legal distinction between (1) TDM activities used to train and build a model and (2) making that model available on the EU market.[174] The different actors involved in such practices in both loosely and tightly integrated capacities across the

[16]; Case C-162/97, *Nilsson* ECLI:EU:C:1998:554 [54].

169 Alexander Peukert, 'Copyright in the Artificial Intelligence Act – A Primer' (2024) 73 GRUR International 497.

170 Mireille MM van Eechoud, 'Territoriality and the Quest for a Unitary Copyright Title' (2024) 55 IIC - International Review of Intellectual Property and Competition Law 66, 74.

171 See eg in relation to Netflix as an indicative empirical study, Timm Böttger and others, 'Open Connect Everywhere: A Glimpse at the Internet Ecosystem through the Lens of the Netflix CDN' (2018) 48 SIGCOMM Comput. Commun. Rev. 28.

172 Case C-523/10 *Wintersteiger* ECLI:EU:C:2012:220 [36].

173 Rosati (n 85).

174 See eg Stieper and Denga (n 166). treating training, design and output as 'separate regulatory subjects'.

value chain, and the difficult-to-reason-about distinction between data and models trained from it, makes such a leap hard to make, let alone on the basis of a recital alone. Value-chain obligations in law are complex, and require careful regulatory design, and it may be difficult for a court to establish these on its own with such little legislative steer.

One possible solution to address the extraterritoriality issue could be to tackle it in the GPAI Code of Practice. Previous versions of the Code either mentioned Recital 106 or more clearly suggested a type of voluntary extraterritoriality.[175] The current version of Code's copyright chapter eliminated these suggestions and does not mandate extraterritoriality or, notably, even mention Recital 106.[176] This shift raises the question of whether some of the Code's measures can or should be understood as promoting a voluntary and mild (even implicit) approach to extending the copyright policies obligation in Article 53(1)(c) beyond the territorial borders of European law. The answer to that question is unclear and depends on how the Code's copyright chapter is interpreted. On the one hand, it is possible to read Measures 1.2 and 1.3 as indicating that signatories voluntarily commit to comply with the AI Act's copyright policy obligation — including lawful access and rights reservation requirements — as they pertain to web crawling activities carried out by them or on their behalf in the model training phase, regardless of whether these activities occur inside or outside the EU. On the other hand, it is also plausible to argue that, since nothing in these measures explicitly references extraterritorial application, they should be interpreted in accordance with EU law. That is to say, based on the arguments above, these measures should not apply to TDM activities taking place outside the EU, whether carried out by the GPAI model provider or related third parties.[177]

In our view, underlying the Code of Practice is a form of a copyright value chain approach, whereby signatories to this voluntary code are asked to take reasonable efforts to establish both upstream and downstream copyright compliance

Such a value chain approach appears to be aligned with EU copyright law more than a blunt form of extraterritoriality. In the AI Act, it is currently enabled by meta-regulation as well as strong policy and market incentives.[178] An effective approach is however likely more sophisticated than the current legal text can support. Further standardisation of this area may solidify this direction, but for the richest providers, codes of conduct and harmonised standards will remain voluntary, as alternative means of compliance will be affordable. Much will likely hinge on whether the CJEU confidently expands these aspects of the AI Act,

175 Quintais (n 72); Sévérine Dusollier and others, 'Copyright and Generative AI' (2025) 16 JIPITEC 1.

176 The only reference to territoriality is found on the section entitled and restating the 'Legal text: Article 53(1)(c) AI Act', to the effect that Signatories remain responsible for ensuring that the Measures in their copyright policy comply with each Member State's implementation of EU copyright and related rights law, particularly CDSM Directive, art 4(3), before undertaking any copyright-relevant act in the territory of the relevant Member State, as failure to do so may result in liability under Union law. See GPAI Code of Practice, Copyright Chapter, 4.

177 Quintais (n 71).

178 Lee A Bygrave and Rebecca Schmidt, 'Regulating Non-High-Risk AI Systems under the EU's Artificial Intelligence Act, with Special Focus on the Role of Soft Law' (Social Science Research Network, 24 October 2024) <https://papers.ssrn.com/abstract=4997886> accessed 24 October 2024; Alexander Peukert, 'Copyright and the Meta-Regulation of Intermediary Services and Artificial Intelligence' (*Kluwer Copyright Blog*, 13 June 2024) <https://copyrightblog.kluweriplaw.com/2024/06/13/copyright-and-the-meta-regulation-of-intermediary-services-and-artificial-intelligence/> accessed 5 February 2025.

including in interplay with the standards and codes being produced, as well on any parallel developments in EU copyright law which will feed into the AI Act's administrative regime. As the enforcer of this section, whether many of these questions make it to the CJEU at all will also depend heavily on the capacity and political will of the AI Office to bring these cases at all.

IV Training data summary

The second copyright-related obligation in the AI Act is found in Article 53(1)(d). According to this provision, GPAI model providers must prepare and make publicly available a sufficiently detailed summary of the data used in pre-training and training in GPAI models, following a template provided by the AI Office. Article 3(29) AI Act defines 'training data' as data used for training an AI system through fitting its learnable parameters.

This obligation extends beyond copyright. The data in question and the legitimate interests considered are not limited to copyright and copyright holders but also include, for example, privacy and data protection, research, non-discrimination, respect for diversity, and fair competition. Thus, copyright interests should not be the sole factor in defining the scope of this obligation. [179]

This obligation does not require GPAI providers to itemise all copyrighted material included in their training datasets with clear identification of rights ownership claims across jurisdictions. Such a requirement would make compliance virtually impossible due to factors such as the low threshold of originality, the territorial fragmentation of copyright and its ownership, the absence of a copyright registration requirement, and the poor state of copyright metadata.[180] This is clear from the AI Act's recitals, which elaborate that 'sufficiently detailed' means 'generally comprehensive in its scope instead of technically detailed', giving by way of example the listing of the main datasets and a 'narrative' explanation (rather than itemised list) of other data sources used.[181] As is common in EU technology law relating to informational disclosure, trade secrets and confidentiality are mentioned as a consideration but a balance is supposed to be struck. The use of the clear term 'comprehensive' indicates that the balance is to be struck in relation to the detail, rather than in relation to the scope. The use of the term 'data sources' should be seen to encompass other inputs to model training, which could be less conventional data sources such as other models, through techniques such as synthetic data production, model distillation and 'LLM-as-a-teacher'.[182]

---

[179] Zuzanna Warso, Maximilian Gahntz and Paul Keller, 'Sufficiently Detailed? Towards Robust Training Data Transparency' (Open Future Foundation; Mozilla 2024) <https://openfuture.eu/publication/towards-robust-training-data-transparency> accessed 25 June 2024; Zuzanna Warso, Maximilian Gahntz and Paul Keller, 'Blueprint of the Template for the Summary of Content Used to Train General-Purpose AI Models (Article 53(1)d AIA) – v.2.0' (Open Future Foundation; Mozilla 2024) Policy Brief <https://openfuture.eu/blog/sufficiently-detailed-summary-v2-0-of-the-blueprint-for-gpai-training-data/>.

[180] See eg Martin Senftleben and others, 'Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies and the Opportunity Arising from Article 17 of the CDSM Directive' (2022) 13 JIPITEC 1; Geiger and Iaia (n 72).

[181] AI Act, recital 107.

[182] Michal Gal and Orla Lynskey, 'Synthetic Data: Legal Implications of the Data-Generation Revolution' (2023) 109 Iowa Law Review 1087; Jianping Gou and others, 'Knowledge Distillation: A Survey' (2021) 129 International Journal of Computer Vision 1789.

The AI Office is responsible for providing the template that meets these requirements and for monitoring compliance with these obligations. Importantly, there should be no need for the Office to conduct a work-by-work assessment of the training data in tewil, rms of copyright compliance.[183] It is also crucial to emphasise that the obligation of transparency and disclosure associated with the template does not affect the availability of other remedies at the disposal of rights holders, such as the right to information in the IP Enforcement Directive.[184]

In July 2025, the Commission released an *Explanatory Notice and Template for the Public Summary of Training Content* to help GPAI model providers summarise the data used to train their models.[185]

The first section, on general information, requires details identifying the provider and model, including versioned model name(s), model dependencies, and the date the model was placed on the Union market. It also covers modalities (eg text, image, audio, video, other), the overall size of the training data (for each modality and within ranges of tokens, images, etc.), and other characteristics such as types of copyrighted content and formats. Providers must also give general information about the training data after pre-processing and before model training.

The second section of the template concerns data sources. It requires disclosure of the main datasets[186] used to train the model, including major private or public databases, a detailed description of data scraped online by or for the provider (with a summary of key domain names), and descriptions of all other data sources (eg user or synthetic data) to ensure a complete summary of the training content. It includes six types of sources. The first four are particularly relevant for copyright purposes.

The first source is that publicly available datasets, meaning those compiled by third parties, freely available, and downloadable in full or in parts. Examples include datasets or collections hosted on public repositories, online platforms, specialised websites, or snapshots of Common Crawl. The template notes that free public availability does not necessarily mean the data are free of rights, as content may still be subject to licensing terms or conditions of use, including free or open licences that prohibit certain uses such as model training. A dataset is deemed 'large' if its data for any modality exceeds 3% of all publicly available training data for that modality. Size should be measured after pre-processing and without splitting to avoid reporting circumvention.

The second data source are private, non-public datasets obtained from third parties. These cover: (1) licensed datasets obtained through agreements with rightsholders, their represent-

---

183 AI Act, recital 108.

184 Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights OJ L 157/45, art 8. See eg European Commission, *Explanatory Notice and Template for the Public Summary of Training Content for General-Purpose AI Models required by Article 53(1)(d) of Regulation (EU) 2024/1689 (AI Act)*, C(2025) 5235 final, Brussels, 24 July 2025, 6.

185 European Commission (n 184).

186 The template clarifies that a 'dataset' should be understood as a single, pre-packaged collection of data. It adds that the filtering and pre-processing of data collected from the same pre-packaged collection should not be considered a new dataset to be disclosed separately. See European Commission (n 184), Template, section 2.

atives, collective management organisations, or content aggregators; and (2) private datasets used without such agreements.

The third data source covers data crawled or scraped directly by the GPAI provider or on their behalf, excluding publicly available third-party datasets (eg those on platforms like Common Crawl) already listed under the first category. This requires inter alia a disclosure of a summary of the list of top domain names crawled and scraped from online sources in a summarised narrative form, including a list the internet domain names representing the top 10% of scraped content by volume, across modalities where applicable. SMEs, including start-ups, must list the top 5% or up to 1,000 domains to ensure proportionality, in line with Recital 109 of the AI Act.[187]The fourth data source covers user data collected across the provider's services and products—such as email, social media, content platforms, or interactions with the provider's AI systems (eg user inputs and prompts). It excludes data licensed through commercial agreements or customer data used to fine-tune models for specific purposes.

The fifth source is synthetic data, meaning data generated by or for the provider to train a model on another AI model's outputs, such as through distillation or alignment (eg reinforcement learning with AI feedback). It excludes using AI to clean or enrich data.[188] In theory at least, synthetic data does not meet the requirements of human authorship and creativity to merit copyright protection under EU or US law,[189] meaning that this data source is less relevant for copyright purposes.

The sixth and final source is a residual category — 'other sources of data' — covering data outside the previous categories, such as offline collections, self-digitised media, human-labelled datasets commissioned by the provider, or human-generated data from reinforcement learning.

The third and final section of the Template concerns 'data processing aspects' and partly interacts with the copyright policy obligations in Article 53(1)(c) of the AI Act and the GPAI Code of Practice. The first part addresses measures taken by GPAI providers to respect reservations of rights under Article 4(3) of the CDSM Directive, as required by Article 53(1)(c) of the AI Act. It asks whether the provider is a signatory to the GPAI Code of Practice and to describe measures implemented before and during data collection to honour TDM opt-outs, including any protocols or solutions used by the provider or relevant third parties.[190] The next part concerns measures to prevent or remove illegal content under EU law, including unauthorised use of copyrighted material, from training data (eg through blacklists, keywords, or model-based classifiers), without disclosing internal business practices

[187] European Commission (n 184), Template, section 2.3.

[188] European Commission (n 184), Template, section 2. Note that this category may overlap with others. Publicly available datasets (Section 2.1) should be reported there, while synthetic datasets created by third parties for the provider should be reported in this section, not in Section 2.2.2.

[189] P Bernt Hugenholtz and João Pedro Quintais, 'Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?' [2021] IIC - International Review of Intellectual Property and Competition Law <https://doi.org/10.1007/s40319-021-01115-0>; PB Hugenholtz, 'Copyright and the Expression Engine: Idea and Expression in AI-Assisted Creations' (2025) 100 Chicago-Kent Law Review 251; US Copyright Office, 'Copyright and Artificial Intelligence, Part 2: Copyrightability' (USCO 2025) <https://www.copyright.gov/ai/> accessed 1 November 2025.

[190] European Commission (n 184) 6–7 (on the template's approach to balancing disclosure with trade secrets and confidential business information).

or trade secrets.[191]

The Summary must be publicly available no later than when the model is placed on the Union market. It should be published prominently on the provider's official website, clearly indicating the covered model(s) and version(s), as specified in point (30) above. The Summary must also accompany the model across all public distribution channels (eg online platforms).[192] The obligation to make Summaries publicly available applies from 2 August 2025. Providers of models placed on the market before that date must publish the corresponding Summary by 2 August 2027. If, despite best efforts, a provider cannot supply certain required information because it is unavailable or its retrieval would impose a disproportionate burden, the Summary should clearly identify and justify these information gaps. Supervision and enforcement by the AI Office will begin on 2 August 2026.[193]

From a copyright perspective, the transparency obligation in Article 53(1)(d) AI Act, including the publication by model providers of the associated template can help clarify three key aspects essential for determining compliance with the TDM exception under Article 4 of the CDSM Directive.[194] Firstly, noting the uncertainties about this developing regime in the discussion above,[195] it may help establish whether the relevant TDM activities took place within, or have sufficient connections to, the EU territory. This is crucial for determining whether the exclusive rights of reproduction and extraction apply and, consequently, whether the TDM exception is triggered.[196] Secondly, assuming EU law applies, this obligation can assist in assessing whether the GPAI model provider complied with the lawful access requirement of the TDM exceptions in the CDSM Directive, particularly in cases involving web scraping.[197] Finally, by providing information on the training sources and potentially the time period of data collection, this obligation may help determine whether the TDM activities respected any opt-outs that were in place at the relevant source and time. Since rights holders have the ability to exclude their content from TDM under Article 4 CDSM Directive, having access to this information could be instrumental in verifying whether such exclusions were properly observed. [198] By addressing these three aspects, the obligation and associated transparency template have the potential to enhance legal clarity and facilitate compliance with the TDM exception within the EU copyright framework as well as in relation to the AI Office's assessment of the copyright policy.

---

191 European Commission (n 184), Template, section 3.2.
192 European Commission (n 184) 7.
193 European Commission (n 184) 7.
194 European Commission (n 184), recital 8 (making the broader point of compliance with EU copyright law and its enforcement).
195 See eg *supra*, 'Granularity and temporality'.
196 See Quintais (n 71).
197 See generally Margoni (n 119).
198 Abbamonte (n 84) 484.

E Demonstrating Compliance

The AI Act is mostly modelled on an extended version of the New Legislative Framework (NLF), the EU's flagship product safety regime.[199] In such laws, compliance choices are typically a choice between applying what are typically called the *essential requirements* of the legislation directly, applying an approved *harmonised standard* from a European Standardisation Organisation (ESO),[200] or, as an 'exceptional fall back', applying a *common specification* — effectively a Commission-led implementing act subject to the comitology procedure.[201] GPAI model provider requirements both draw on and further adapt this legal approach.[202] We describe these adapted and extended modalities of compliance demonstration below.

I Modalities

We describe the different ways the compliance regime can function below. Poor statutory drafting — likely the result of the messy legislative process — means that while all the possible compliance mechanisms exist in the statute with specific reference to GPAI model providers, not all are explicitly listed in Article 53(4).

(a) Harmonised standards

Approved harmonised standards grant a presumption of conformity with the Act to the extent that they cover the obligations in question,[203] and shifts the burden of proving compliance from the regulatee to the regulator.[204] The Commission is obliged by the Act to issue a harmonised standardisation request for GPAI models 'without undue delay'.[205] The downsides of harmonised standards in this area are well documented. They form part of EU law but are proprietary documents, leading to significant tensions around openness and cost.[206] This tension is heightened due to the interplay between international standards and

[199] See Chapter X [high risk systems and standardisation], this volume.

[200] These are CEN, CENELEC and ETSI. See Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council OJ L 316/12, annex I.

[201] AI Act, art 41(1). This is exceptionally rare — very few active common specifications exist. See Commission, *Non-Paper: Horizontal Approach on Common Specifications* SI(2022) 259, 3 (made public pursuant to a request and appeal by the authors under Regulation (EU) 1049/2001 on June 5 2025 at https://perma.cc/MQ37-488K). This is reflected in their description as an 'exceptional fall back' in AI Act, recital 121. See generally on common specifications in the AI Act, Hans-W Micklitz, 'The Role of Standards in Future EU Digital Policy Legislation' (*ANEC/BEUC* July 2023) <https://perma.cc/3WVY-J2AD>.

[202] Differences remain. For example, there is no clear notion of *conformity assessment*, no interaction with the notified bodies laid out in the AI Act, and enforcement is envisaged at European level, which is not generally the case for an NLF instrument. cf See Chapter X [high risk systems and standardisation], this volume.

[203] AI Act, arts 53(4), 40(1).

[204] Commission Notice, The 'Blue Guide' on the Implementation of EU Product Rules 2022 [2022] OJ C 247/1 (Blue Guide) 8.

[205] AI Act, art 40(2).

[206] Case C-613/14 *James Elliot* ECLI:EU:C:2016:821; Case C-588/21 P *Public.Resource.Org* ECLI:EU:C:2024:201. See generally Olia Kanevskaia, 'Is It Really All about the Money? The Future of European Standardization after PublicResourceOrg' (2025) 16 European Journal of Risk Regulation 344 <https://doi.org/10.1017/err.2024.64>.

harmonised standards. International standards organisations are challenging the free release of proprietary standards now mandated under EU law as damaging to their wider interests,[207] which only becomes more complex where EU standards are composed of licensed parts (eg common definitions) from international standards.[208] Such bodies and processes also have limited legitimacy in a fundamental rights context.[209]

(b) Common specifications

Common specifications for GPAI models are not mentioned in Article 53 but are possible to create for GPAI models, either when a harmonised standard request or production has failed or the Commission does not expect to publish a reference to a standard within a reasonable period.[210] As ESOs have little experience with fundamental rights concerns, and the Commission admits that their ability and competence to deal with them is unclear, the AI Act goes beyond prior NLF regimes and allows the Commission to draft common specifications when proposed standards insufficiently address such concerns.[211] Common specifications impart a similar presumption of conformity to harmonised standards, and must be repealed when a suitable harmonised standard is adopted.[212] Scrutiny of common specifications through the comitology procedure is typically via Member State nominated experts, which brings accountability challenges and raises questions around whether this set-up is sufficiently representative of the broad stakeholder base the governance GPAI models appears to call for.[213] Unlike all scrutiny of implementing legislation, passing a common specification for the AI Act will require a qualified majority rather than simply a lack of an objection.[214] Furthermore, a single member state can trigger a review of a specification by the Commission.[215]

(c) Codes of practice

In the AI Act, codes of practice for GPAI models are a co-regulatory mechanism designed to provide additional clarity before and until the creation and approval of harmonised stand-

[207] Case T-631/24 *International Electrotechnical Commission and ISO v Commission* (pending).

[208] See generally Justus Baron and others, 'Making the Rules: The Governance of Standard Development Organizations and Their Policies on Intellectual Property Rights' (*Joint Research Centre, European Commission*, 2019) <https://doi.org/10.2760/48536>.

[209] Micklitz (n 201201); Veale and Zuiderveen Borgesius (n 11); on the difficulty of civil society participation, see generally Jean-Christophe Graz and Christophe Hauert, 'Translating Technical Diplomacy: The Participation of Civil Society Organisations in International Standardisation' (2019) 33 Global Society 163, 168; Alexandre Mallard, 'L'écriture des normes' (2000) 18 Réseaux 37, 51.

[210] AI Act art 40(1)(a).

[211] AI Act art 40(1)(a)(iii). In this context, the Commission (n 201201) motivates such a provision, stating that the 'full ability and competence of [ESOs]' in the area of fundamental rights 'remains to be fully assessed'.

[212] AI Act, arts 40(3–4).

[213] Gijs Jan Brandsma, *Controlling Comitology* (Palgrave Macmillan 2013).

[214] AI Act, art 98(2). Because such acts will concern the "protection of the health or safety of humans" given the objectives of the AI Act, they will need qualified majority approval rather than just no objection. Regulation (EU) No 182/2011 on rules and general principles concerning mechanisms for control by Member States of the Commission's exercise of implementing powers OJ L 55/13 ('Comitology Regulation'), art 5(4).

[215] AI Act, art 41(6).

ards.[216] The GPAI Code of Practice was, in part, a result of tensions around timing. To understand the necessary runway for the harmonised standards described above, we can look at the high-risk AI regulatory timeline. The substantive obligations in Chapter III only apply from August 2027,[217] yet the Commission already released a mandate for draft standards in relation to these areas in late 2022, almost two years before the AI Act passed,[218] stemming from the Commission's expressed anxiety around imposing obligations without such standards.[219] As the GPAI provisions were added during trialogue and shifted in form up until the approval of the AI Act, such a move was not possible.

For all GPAI model providers, codes produced under this must at least cover the information and documentation obligations and the training data summary — the requirement omitted the copyright policy.[220] Nonetheless, the Code published in July did still contain a chapter on copyright (as has been discussed above). The AI Office must also aim to ensure that the Code contains, as appropriate, key performance indicators (KPI), and that these are reported upon by code participants — at the time of writing, there are no apparent KPIs listed in the Code, and so compliance with this provision is uncertain.

The Act characterises a code of practice as something providers 'participate in', 'adhere to', or make 'commitments' in relation to.[221] While standards require application, codes of practice operate more like a club. Being part of this club brings limited legal effect — the potential to 'rely on' the code to 'demonstrate compliance'[222] — notably short of the presumption of conformity given by a harmonised standard or common specification — and an obligation for the Commission to consider commitments as a mitigating factor in any enforcement action.[223] The Commission has further stated that its enforcement strategy will consist, for signatories, of checking their adherence to the Code, rather than assessing them more holistically.[224] As being part of this club entails reporting obligations against KPIs, and that no independent obligation to report against KPIs exists, treating the code as a standard alone is a substantively different activity than operating it as a club. In such a case, it would seem appropriate that the AI Office did not apply even the limited benefits described above during an enforcement process.

The AI Office is subject to several best-efforts obligations in the AI Act in this regard. They must aim to ensure wide stakeholder needs, including affected persons, are taken into account in a code. They did this through a call for interest in July 2024, which led to an extremely

---

216 AI Act, art 53(4).

217 AI Act, art 113(c).

218 Commission, 'Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence' (5 December 2022) <https://ec.europa.eu/docsroom/documents/52376> accessed 17 January 2025.

219 See eg the impact assessment, where the Commission's economic calculations assume that harmonised standards are available before the AI Act provisions apply in the Union, in Commission, 'Impact Assessment accompanying the 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, 69.

220 A list of the relevant parts this should include in AI Act, art 56(2) interestingly omits the copyright policy obligation, although not very much should be read into that as the provision is clear that it remains within the required scope of the provision.

221 AI Act, arts 53(4), 56(5–6), 101(1).

222 AI Act, art 53(4).

223 AI Act, art 101(1).

224 European Commission (n 10), para 94.

large number of people self-nominating as experts for this process. The Commission ended up with around 1,000 participants at its first (online) meeting, in a process which only opened up after loud public criticism, and was described as 'unwieldy', with an 'extremely compressed timeline', and which 'left little room for thoughtful deliberation or balanced input'.[225]

The first Code of Practice had statutory obligations as to its timeframe, which were missed, with the Code published on 10 July 2025 rather than 7 May 2025. This did however fall before a second, firmer deadline of 2 August 2025, at which point the Commission could have exercised its ability to make common rules (see below). Now this deadline has passed, in order to do this, the AI Office would have to deem an existing code inadequate.[226]

Out of the wide array of EU uses of regulatory instruments called codes, the AI Act's codes of practice most resemble codes of conduct in the GDPR.[227] Such codes are also envisaged, facilitated and monitored by regulators or EU bodies, can be used to demonstrate compliance but grant no presumption of conformity, may be considered as a liability reduction tool,[228] and may be given Union–wide effect through an implementing act.[229] GDPR codes of conduct do not, however, have a highly visible development process, nor are they so easily superseded as codes of practice in the AI Act are. There may be a lesson in this — given a very similar incentive structure, few GDPR codes of conduct have been approved at national level, none at European level at the time of writing, and their impact has been correspondingly negligible.

Codes of practice appear to be short-lived instruments. The Commission describes them as a 'temporary tool'.[230] The Act states that providers of GPAI models may rely on Codes of Practice to demonstrate compliance only 'until a harmonised standard is published'.[231] The Commission is furthermore legally obliged to work towards this moment 'without undue delay'.[232]

Despite this, the AI Act envisages a process where the AI Office and the European Artificial Intelligence Board review the code of practice and produce a published assessment of its adequacy, and that the AI Office should 'encourage and facilitate the review and adaptation of the codes'.[233] The process of who would be involved in such a review is not outlined in the law.

We may expect a tension between the short-lived nature of the Code of Practice due to the shadow of harmonised standards, and the mechanisms of review, which imply a longer life

---

225 Cynthia Kroet, 'Drafting of AI Code of Practice Faces at Least One Month Delay' (*Euronews*, 20 February 2025) <https://www.euronews.com/next/2025/02/20/drafting-of-ai-code-of-practice-faces-at-least-one-month-delay> accessed 10 March 2025; Martin Ebers, 'When Guidance Becomes Overreach' (*Verfassungsblog*, 8 April 2025) <https://verfassungsblog.de/when-guidance-becomes-overreach-gpai-codeofpractice-aiact/> accessed 12 November 2025.

226 AI Act, art 53(9)

227 Comparing these different forms of codes, see generally Carl Vander Maelen, 'Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legalisation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof' (2022) 47 European Law Review 752.

228 AI Act, art 101(1).

229 GDPR, arts 40–41.

230 European Commission (n 10), para 100.

231 AI Act, art 53(4).

232 AI Act, art 40(2).

233 AI Act, art 56(6), 56(8).

for this regulatory mechanism. Given that both forums are significantly led and influenced by private actors, this may create practical challenges, if industrial actors believe that one path is likely to be more commercially advantageous than the other — whether they want to have 'another go' at reshaping the rules, or whether they wish to avoid the potential ratcheting-up of standards that a fresh process might bring.

(d) Common rules

Common rules are an AI Act novelty that only relate to GPAI models.[149] They are to failed codes of practice what common specifications are to failed harmonised standards — a Commission–led, comitology–scrutinised fall back. They grant no presumption of conformity, but are otherwise similar to a common specification other than the explicit ability for a Member State to trigger a review, or the required annulment of such rules upon the approval of a code of practice.

(e) Adequate alternative means

Providers always have the option to demonstrate compliance without regard to the modalities above. Such a demonstration would involve both elaborating on the requirements in relation to the state-of-the-art, and then fulfilling these obligations. The burden of proof is on the provider to prove compliance, which will require them to convincingly evidence and document their interpretative choices and the approach they use to comply with it. The Commission will assess this evidence.[234] Where a common specification exists but providers choose not to use it, providers must 'duly justify that they have adopted technical solutions that meet the requirements [...] to a level at least equivalent thereto' of the statute.[235] This is not a requirement present for harmonised standards as an inelegant solution to avoid providers being forced by law to purchase proprietary documents. No similar obligation is explicitly present for codes of practice or common rules, but their similarity to common specifications combined with their openness means that an obligation to refer to them can be read horizontally. It is worth noting that the wealth of many general-purpose AI firms at the time of writing might render this a significantly more feasible and preferable option for some than it usually is in an NLF instrument.

II Authorised representatives

Some GPAI model providers require authorised representatives in order to demonstrate compliance with the AI Act.

---

234 AI Act, art 53(4) ('Providers of general-purpose AI models who do not adhere to an approved code of practice or do not comply with a European harmonised standard shall demonstrate alternative adequate means of compliance for assessment by the Commission.')

235 AI Act, art 41(5).

(a) Function

An authorised representative is a natural or legal person established within the Union who has received and accepted a written mandate to perform and carry out, on behalf of a GPAI model provider, the obligations and procedures established in the AI Act.[236] All GPAI model providers in scope of the GPAI model obligations located outside the Union must mandate an authorised representative,[237] and such a mandate must be accepted before a model is made available in the Union.[238] They verify and retain documentation of the provider in relation to obligations of the AI Act, as well as to provide information to regulators and to co-operate in action related to larger AI value chains.[239] This obligation echoes authorised representative obligations for providers of high-risk AI models.[240] Authorised representatives in the AI Act have no explicit requirements of skill or competence, unlike other regimes such as the Medical Devices Regulation,[241] nor a registration obligation.[242] Regulators from that domain using this registration information to audit authorised representatives have found them to often be post boxes with few, if any local staff, which does not bode well for accountability within the AI Act.[243]

(b) Liability of authorised representatives

The AI Act does not explicitly state that authorised representatives of extraterritorial GPAI model providers are liable for infringements of their own obligations in the regulation. This appears to be another drafting error, as authorised representatives of high-risk AI system providers are explicitly written into the penalties regime.[244] It seems sensible to read parallel liability across from that. In any case, the authorised representative regime in the AI Act is restrained in terms of liability for the failures of the GPAI model provider. To illustrate, in other instruments, such as the Medical Devices Regulation, authorised representatives can explicitly be held jointly and severally liable with manufacturers for defective devices,[245] while in the AI Act, they are not even mentioned in the section on penalties for the GPAI model obligations.[246]

---

[236] AI Act, art 3(5).

[237] AI Act, arts 3(5), 54.

[238] AI Act, recital 82 (in relation to high-risk systems; no similar recital exists corresponding to art 54).

[239] AI Act, art 54.

[240] AI Act, art 22.

[241] Regulation (EU) 2017/745 of the European Parliament and of the Council 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 OJ L 117/1 (Medical Devices Regulation) OJ L 117/1, art 15(6) (requiring specific degree qualifications and one year's experience in medical device regulation, or four years' experience in such regulation).

[242] Medical Devices Regulation, art 31.

[243] Dutch Health and Youth Care Inspectorate, *Authorised representatives of medical devices do not always meet the basic requirements* (Ministry of Health, Welfare and Sport, the Netherlands May 2024) <https://perma.cc/MW8R-J3E9> accessed 12 November 2025.

[244] AI Act, art 99(4)(b).

[245] Medical Devices Regulation, art 11(5).

[246] AI Act, art 101.

F Exclusions

A multi-layered range of exclusions applies to the obligations of GPAI model providers in the Act: free and open-source licenses, non-commercial activities, scientific research and development, and finetuning. These have to be understood in the specific technical and political economic context of the contemporary GPAI industry, and are crucial to understanding the operation and interpretation of Article 53.

I Free and Open-Source

In the AI Act, obligations placed on providers or their representatives are significantly dialled down if a model is released under a 'free and open-source license', as such a license 'should be considered to ensure high levels of transparency and openness'.[247] It is assumed that access to the code of the model removes the need for regulators or users (and potential high-risk AI system providers) to have access to documentation pursuant to Article 53(1)(a) and (b), for regulators and potential deployers, respectively. The only additional requirement for models released under a free and open source license is that the models' parameters (including the weights), the information on the model architecture, and the information on model usage, are made publicly available. This, however, falls significantly short of the type of disclosure/transparency required in Article 53(1)(a) and (b), including the minimum elements set out in Annex XII.

On inspection of the specifically exempted provisions, it becomes evident that models released under the Act's definition of 'free and open source' are unlikely automatically allow individuals to, through analysis or examination, obtain the missing information that non-open source models would have to document for regulators and providers. Under the AI Act, a 'licence should be considered to be free and open-source also when it allows users to run, copy, distribute, study, change and improve software and data, including models under the condition that the original provider of the model is credited, the identical or comparable terms of distribution are respected'.[248] This definition does not give model users or regulators access or rights to access any of what the regulation literature describes as *credence* attributes about a machine learning model — properties which are inherently difficult or impossible to understand from its examination or use alone.[249] From a model alone, it is not possible to reliably glean 'the type and provenance of data and curation methodologies (eg cleaning, filtering, etc.), the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases',[250] because such data has been subsumed into the weights of the model, and while models can leak and

[247] AI Act, arts 53(2), 54(6), recital 102. Note that while there is a general scope exclusion in Article 2(12) for open-source systems, this applies only to AI systems, not AI models.

[248] AI Act, recital 102.

[249] See generally on credence attributes of machine learning models, Kira JM Matus and Michael Veale, 'Certification Systems for Machine Learning: Lessons from Sustainability' (2022) 16 Regulation & Governance 177.

[250] AI Act, annex XI, section 1, para 2(c); this obligation is also present in relation to Article 53(1)(b) in a weaker form in Annex XII, para 2(c).

memorise parts of their training data,[251] this characteristic does not extend to providing such full and detailed information about training. The recitals even acknowledge this, noting that 'the release of general-purpose AI models under free and open-source licence does not necessarily reveal substantial information on the data set used for the training or fine-tuning of the model' — stating that as a justification for providing a training data summary for copyright purposes but decidedly ignoring its implications for data aspects of the information requirements. Computational training resources or energy use can also not be reliably estimated from a model itself,[252] as the visible characteristics in a final model such as the structure or number of parameters, say nothing about how many times they were adjusted over many computationally-intensive rounds ('epochs') of training. This is because the model structure, including the total number of parameters, is typically set at the point the training begins. It is also unlikely that examination of the model would clearly reveal the 'rationale and assumptions' of the designer.[253]

The European Commission has attempted to further specify what is and is not open source, in guidance which is not legally binding but will inform its enforcement. While it generally says that usage restrictions will disqualify a model from this categorisation, calling out common tactics used by large 'open source' AI releases such as scale thresholds triggering a new license requirement (used by eg Meta) or limitations on commercial use (as used by eg Cohere), they permit proportionate behavioural use licensing, which is the use of a software license to try and preclude societally harmful uses of a model (as seen in eg OpenRAIL, the Open Responsible AI License).[254]

This challenge is entwined with current controversies and debates over what makes a machine learned model 'open-source'. The Open Source Initiative has indicated that they do not consider models 'open-source' unless, additionally to the EU definition, the source code for training and a complete description of training data, pre-processing instructions, and information on where to locate it either publicly or from third parties, is released.[255] The assumption that practical scrutability and accountability can automatically flow from such openness has been subject to serious challenge in the scholarship, particularly because transparency tells little about how models perform in specific contexts; the deployment that is a prerequisite to analysis is often restricted to those with significant compute power; and

---

[251] European Data Protection Board, 'Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models' (*EDPB*, 17 December 2024) <https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf> accessed 18 December 2024.

[252] AI Act, annex XI, section 1, para 2(d–e).

[253] AI Act, annex XI, section 1, para 2(b).

[254] European Commission (n 10), paras 83–8; Cohere, 'CC-BY-NC 4.0 License with Acceptable Use Addendum' (2025) <https://cohere.com/cohere-labs-cc-by-nc-license> accessed 11 November 2025; Meta, 'Meta LLaMA 3 Community License Agreement (18 April 2024) <https://www.llama.com/llama3/license/> accessed 11 November 2025; on behavioural use licensing, see generally Danish Contractor and others, 'Behavioral Use Licensing for Responsible AI', *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) <https://dl.acm.org/doi/10.1145/3531146.3533143>.

[255] Open Source Initiative, 'The Open Source AI Definition – 1.0' (*Open Source Initiative*, 2024) <https://opensource.org/ai/open-source-ai-definition> accessed 7 January 2025. This tension has also become important because some of the most currently salient, contemporary general-purpose AI models that claim to be licensed as open-source, such as Meta's LLaMA series of models, have significant restrictions on how the model can be used commercially, such as restricting users from using it to train additional models, or to power services delivered to more than 700m monthly users; see generally Steven J Vaughan-Nichols, 'Meta's Llama 2 Is Not Open Source' (*The Register*, 21 July 2023)., <https://www.theregister.com/2023/07/21/llama_is_not_open_source/> accessed 7 January 2025.

models may have been fine-tuned and extended, and as such the layer of transparency any one provider can provide is inherently limited.[256]

II Non-Commercial Activities

Only models that are made 'available on the market' are within scope of Article 53, which only includes any GPAI model supplied for distribution or use in the Union market 'in the course of a commercial activity, for payment or free of charge'.[257] This opens up the possibility for a non-commercial entity to provide such a model and not be subject to the regulations. While many of the top models today are developed by companies with commercial interests, several are also from organisations more difficult to classify in this way. The Falcon series of models, for example, is produced by the Technology Innovation Institute, a research institute funded by the government of Abu Dhabi.[258] Many models are produced by EleutherAI, which is a 'non-profit research institute'.[259] The Netherlands is also training a 'public interest' Dutch–English language model, 'GPT-NL'.[260] Guidance from the Commission on the scope of the GPAI provisions surprisingly fails to engage with this question.[261]

Such an exemption will often not be straightforward to apply. Simply because a model is availably freely and open-source does not mean it also benefits from this non-commercial exemption. This is well-illustrated by the new Product Liability Directive, which only applies its own open-source exemption to 'free and open-source software *that is developed or supplied outside the course of a commercial activity*' (emphasis added), clearly establishing the expectation that both commercial and non-commercial open-source software exist.[262] 'Openness' has long been used for commercial gain, and companies contributing to both traditional open-source software and contemporary open-source AI often do so for strategic, commercial reasons, such as to encourage others to improve the software, or to disrupt proprietary offerings from and market consolidation by competitors.[263]

Relatedly, just because an entity is established as a non-profit, does not mean it will be acting non-commercially in relation to a particular GPAI model. The CJEU has specifically highlighted, in the context of eligibility for trademark protection, that 'the fact that a

[256] David Gray Widder, Meredith Whittaker and Sarah Myers West, 'Why "Open" AI Systems Are Actually Closed, and Why This Matters' (2024) 635 Nature 827.

[257] AI Act, art 3(10).

[258] Ebtesam Almazrouei and others, 'The Falcon Series of Open Language Models' (*arXiv*, 2023) <https://doi.org/10.48550/arXiv.2311.16867> accessed 22 January 2025.

[259] Stella Biderman and others, 'The View from 30,000 Feet: Preface to the Second EleutherAI Retrospective' (*EleutherAI Blog*, 2 March 2023) <https://blog.eleuther.ai/year-two-preface/> accessed 22 January 2025.

[260] Tom Barbereau and Lieke Dom, 'GPT-NL: Towards a Public Interest Large Language Model', *2nd Workshop on 'Public Interest AI' co-located with the 47th German Conference on AI (KI 2024)* (2024) <https://resolver.tno.nl/uuid:0c3607e0-14fc-4ed6-abc2-731e07debb2b> accessed 19 January 2025.

[261] European Commission (n 10), section 3.

[262] Interestingly, the Product Liability Directive's exemption around open-source software excludes only 'free and open-source software that is developed or supplied outside the course of a commercial activity', a significantly narrower scope which, if replicated in the AI Act, would have taken many AI models out of this exclusion. See 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 OJ L, art 2(2).

[263] Widder, Whittaker and Myers West (n 256).

charitable association does not seek to make profit does not mean that its objective cannot be to create and, later, to preserve an outlet for its goods or services', and acknowledged that funding and incentives for such entities might come through a variety of means, such as subsidies.[264] When considering related issues, such as whether an activity is 'economic', or whether a service is 'supplied in return for remuneration', the CJEU has repeatedly emphasised that all the circumstances in which it is supplied have to be examined, in a case-by-case assessment, due to the difficulty of providing a precise definition.[265] As a result, it may be difficult for many types of entity, particularly any that carry out commercial activities in any related area, to confidently claim this exemption.

The future may see further difficulty in classifying models due to decentralised training. GPAI models are often finetuned by individuals, arguably in a non-commercial, private manner, and shared socially on model marketplaces. For example, the *GPT-4chan* language model, trained on the toxic posting on the 4chan imageboard's */pol/* forum, was finetuned and released by a private, individual hobbyist, before being voluntarily removed from its hosting intermediary, Hugging Face, for its ability to be misused.[266] Model marketplaces like Civitai have emerged as hybrid social network platforms designed to make machine learning finetuning — often of abusive or arguably illegal models — accessible to private individuals.[267] Technical peer-to-peer finetuning and querying protocols are also emerging, in the style of BitTorrent, to make computation for such finetuning accessible to private individuals.[268] There is no unambiguous exemption for private providers in the AI Act as there is for private (natural persons) deployers,[269] meaning that commerciality will need to be assessed in these situations, which may prove difficult. Production and consumption are always blurred in content-heavy social networks, and model finetuning is proving no different.

III Scientific Research and Development

(a) Models for research purposes

Actors may benefit from a related layer of exclusion. If an AI model is 'specifically developed and put into service for the sole purpose of scientific research and development', it is further excluded from any obligations in the regulation, including those in Article 53.[270] This is again, surprisingly, omitted from the Commission's guidance on the GPAI provisions' scope.[271]

---

[264] Case C-442/07 *Radetzky* ECLI:EU:C:2008:696 [17]–[18].

[265] Case C-612/21 *Gmina O.* ECLI:EU:C:2023:279 [35]; Case C-520/14 *Gemeente Borsele* ECLI:EU:C:2016:334 [29]; Case C-230/94 *Enkler* ECLI:EU:C:1996:352 [27].

[266] Robert Gorwa and Michael Veale, 'Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries' (2024) 16 Law, Innovation and Technology 341 <https://doi.org/nq2g>.

[267] Ibid.

[268] Alexander Borzunov and others, 'Petals: Collaborative Inference and Fine-Tuning of Large Models' (arXiv, 2 March 2023) <https://doi.org/n299>; Jue Wang and others, 'CocktailSGD: Fine-Tuning Foundation Models over 500Mbps Networks', *Proceedings of the 40th International Conference on Machine Learning* (PMLR 2023).

[269] AI Act, art 2(10).

[270] AI Act, art 2(6).

[271] European Commission (n 10).

This includes models developed in the course of both commercial and non-commercial activities. The term 'AI model' is undefined in the law, but presumably includes at least GPAI models (which are defined) as a subset of these. Accessibility and reusability by others are key criteria for data-driven scientific research,[272] and so unlike with many research exemptions, it cannot be assumed that models and systems developed for research alone are not going to be, in practice, generally available to the public for open-ended reuse.

(b) Models during development

AI Act obligations do not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service.[273] However, the Commission acknowledges that this is somewhat at tension with several regulatory requirements. GPAI model providers are supposed to notify the Commission around the training compute threshold in order to flag a model that may have systemic risk, from which follow obligations that must be carried out prior to a model being placed on the market.[274] GPAI model providers are also practically obliged to document their research and development process else they will not be able to meet the essential requirements around, for example, training data description.[275] The tension between these obligations and the law's stated scope in this regard appears to be another drafting omission. The Commission again ignores this provision entirely in its scope guidance.[276]

IV Value Chains and Fine-tuning

Algorithmic systems are structured in supply chains, and GPAI models no less so.[277] Models build on previous models in many ways. Some new models use existing models as datasets —or 'teachers' — to train or finetune them with a variety of 'knowledge distillation' methods.[278] Other models are finetuned versions of previous ones. Models can be finetuned in their entirety, for example where their parameters are allowed to adjust based on new data. An important range of 'adaptor' methods also exist, which finetune models to specific tasks without changing all their parameters,[279] typically allowing an add-on to be distributed

---

272 Mark D Wilkinson and others, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship' (2016) 3 Scientific Data 160018.

273 AI Act, art 2(8).

274 AI Act, recital 114. See generally European Commission, *General-Purpose AI Models in the AI Act – Questions & Answers* (European Union 14 March 2025) <https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers> accessed 13 March 2025. See generally Chapter X [systemic risk], this volume.

275 European Commission (n 274).

276 European Commission (n 10).

277 Jennifer Cobbe, Michael Veale and Jatinder Singh, 'Understanding Accountability in Algorithmic Supply Chains', *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2023) <https://doi.org/gsb98p>.

278 Xiaohan Xu and others, 'A Survey on Knowledge Distillation of Large Language Models' (arXiv, 21 October 2024) <https://doi.org/10.48550/arXiv.2402.13116>.

279 Zeyu Han and others, 'Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey' [2024] Transactions on Machine Learning Research <https://doi.org/n25b>; Clifton Poth and others, 'Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning' (arXiv, 18 November 2023) https://doi.org/10.48550/arXiv.2311.11077. Adaptor

that can be applied to an already obtained module to create the finetuned version. These methods require a copy of the original model as well as the adaptor to create the new model. Finetuning can radically change the behaviour of a model, from giving it new capabilities to breaking safeguards designed to protect against harmful or illegal queries.[280]

The AI Act outlines the ways finetuning interacts with the scope of the law. Recital 109 states that 'the obligations for providers of general-purpose AI models should be limited to that modification or fine-tuning, for example by complementing the already existing technical documentation with information on the modifications, including new training data sources'. This implies that the requirements under Article 53 are somewhat decomposable into multiple parts, which broadly does appear to be possible as they are primarily descriptive, unlike the requirements for models with systemic risk, which require publishing of testing and benchmarks which are necessarily a function of both the finetuned component and the original model combined.[281]

However, it is not clear all models that are finetuned will be based on models which have obligations under Article 53. Any models which are exempt for the reasons above — open-source models, models produced by non-commercial entities, or models produced solely for scientific research and development, will not require the information obliged by Article 53(1)(a–b). Non-commercial and scientific research and development models do not give rise to Article 53(1)(c–b) obligations either.

Furthermore, the Commission, in guidance, 'deems that it is not necessary for every modification of a general-purpose AI model to lead to the downstream modifier being considered the provider of the modified general-purpose AI model'.[282] They state that modification must lead to a 'significant' change in the model's 'generality, capabilities, or systemic risk', which they measure by an entirely arbitrary indicative criterion of whether the 'training compute used for the modification is greater than a third of the training compute of the original model'.[283] This has no clear basis in the statute, and creates enormous loopholes. This would mean that an entity slightly modifying a model and placing it on the market would not have to, for example, declare anything about the information they had used (copyrighted or not), nor themselves follow a copyright policy relating to either input or output with any finetuned data they had added. The figure of a 'third of the training compute of the original model' is entirely divorced from current AI practice, where on platforms such as Civitai, individuals use parameter-efficient finetuning methods such as LoRAs to adjust models with copyrighted information using far, far less than a third of the compute used to train common base models on that platform, such as Stable Diffusion or Flux, in the process giving them the abilities to produce material subject to intellectual property rights or

methods notably include LoRAs (low-rank adaptations), for which there exist significant distribution model marketplaces specialised in their distribution such as Civitai. See generally Edward J Hu and others, 'LoRA: Low-Rank Adaptation of Large Language Models', *International Conference on Learning Representations (ICLR)* (2021) https://doi.org/10.48550/arXiv.2106.09685.

280 See eg Andy Arditi and others, 'Refusal in Language Models Is Mediated by a Single Direction', *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)* (2024) <https://doi.org/10.48550/arXiv.2406.11717>.

281 AI Act, art 55. See generally Chapter X [systemic risk], this volume.

282 European Commission (n 10), para 60.

283 European Commission (n 10), paras 61, 63.

used for intimate image abuse.[284] Such methods are widely adopted within cloud firms too, which use LoRAs to allow commercial entities to customise and in turn resell models.[285] This position from the European Commission is particularly absurd given that, as described above, the statute itself already limits the obligations for a finetuning entity to the change that they made, meaning that the regulatory burden is minor. The Commission already acknowledges that its rule-of-thumb for what constitutes a change is itself largely unworkable because training data compute is not something that providers are obliged to communicate to users (only to regulators, see Table 1 above), and so they in turn replace it with a reference to a static set of numbers, which again are divorced from the literature on parameter-efficient finetuning.[286]

The Commission also provides no indication of what happens in general when a provider finetunes an open-source model and releases it as a closed source model. In this case as well as several other combinations of the above exemptions, a finetuner cannot simply build on previously provided documentation, because it does not exist. Examining the model itself, which will often be possible to an entity finetuning a model (particularly an open-source model), does not give visibility into the credence properties of a model, including energy use and computation, rationale and assumptions, and data choice, curation and issues around bias.[287] Nor will, in the above cases, such an entity have access to previously created documentation under Article 53(1)(a–b), and potentially also not under (d), as there may be no obligation to provide it, which they can 'complement' with their own additions.[288]

One interpretation might be that the exempt model is simply an input factor, and the documentation effectively has a legally-permitted hole in it. Another would be that the provider building on the exempt model has an obligation to make best efforts to fill in this hole; for example with information known about the exempted model, or information that could be gleaned from its openness. A third approach would be that the provider building on the model is the one making it available in an unexempted format and has to comply with Article 53 in its entirety. This would be, in effect, making a distinction between value chains of *regulated* models and value chains of hitherto *unregulated* models. Providers would be likely to protest at such an obligation, which may be very difficult to do given the lack of information about prior training and development. Yet such a situation may get very confusing with models such as those intended solely for scientific research by the provider, but which are open-source and available for download by another entity which decides to build on them for a commercial purpose. Such research-purposed models were exempt even from the copyright-motivated training data summary, leaving the obligations of the downstream provider unclear.

The Commission recognises this complexity, stating that the 'specific circumstances in which a downstream entity becomes a provider of a new model is a difficult question with

---

[284] Gorwa and Veale (n 266).

[285] See eg Microsoft, 'Customize a Model with Azure OpenAI in Azure AI Foundry Models' (*Azure OpenAI*, 2 July 2025) <https://learn.microsoft.com/en-us/azure/ai-foundry/openai/how-to/fine-tuning> accessed 9 August 2025.

[286] European Commission (n 10), para 64.

[287] See *supra*, 'Free and Open-Source'; see generally Matus and Veale (n 249).

[288] AI Act, recital 109.

potentially large economic implications, since many organisations and individuals fine-tune or otherwise modify general-purpose AI models developed by another entity'.[289] However, not further guidance is provided, except a clarification that independently of the qualification as a GPAI model provider of a downstream entity that incorporates such a model into their system, that entity 'that entity must comply with the relevant AI Act requirements and obligations for AI systems'.[290]

G Concluding remarks

The provisions regarding GPAI models in the AI Act are brief compared to those for the other aspects of this large and complex regulation. They are, however, deceptively complicated. This complexity has only been made worse by poor drafting, which itself is a result of the last-minute addition of this section and the many political disputes that accompanied it. This legal uncertainty has been exacerbated by significant omissions — whether accidental or deliberate — in guidance and the published Code of Practice.

The main difficulties of the GPAI model provisions arise from interactions and intersections. These include intersections with other legal regimes, such as copyright law and trade secrets, as well as interactions with the exemptions and limitations in the Act—particularly given the complex value chains in which GPAI models are involved. Additionally, there are interactions between the various mechanisms available to demonstrate compliance or obtain presumptions of conformity. With so many possible instruments vying to simplify compliance, it is in practice possible that compliance will become *more* fragmented, confusing, and — for some — easy to game. These Articles will need, in the AI Office, a regulator with clear vision and political will. Time will tell whether or not they have one.

[289] European Commission (n 274).
[290] European Commission (n 274).