Article 55

Obligations of providers of general-purpose AI models with systemic risk

Commentary by Hannes Bastians, Madalina Nicolai (joint first authors) | Submitted: May 2026

AI Act provision

Article 55

1. In addition to the obligations listed in Articles 53 and 54, providers of general-purpose AI models with systemic risk shall:

> (a) perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks;
>
> (b) assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk;
>
> (c) keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them;
>
> (d) ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model.

2. Providers of general-purpose AI models with systemic risk may rely on codes of practice within the meaning of Article 56 to demonstrate compliance with the obligations set out in paragraph 1 of this Article, until a harmonised standard is published. Compliance with European harmonised standards grants providers the presumption of conformity to the extent that those standards cover those obligations. Providers of general-purpose AI models with systemic risks 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.

3. Any information or documentation obtained pursuant to this Article, including trade secrets, shall be treated in accordance with the confidentiality obligations set out in Article 78.

Recitals

Recital 110

General-purpose AI models could pose systemic risks which include, but are not limited to, any actual or reasonably foreseeable negative effects in relation to major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content. Systemic risks should be understood to increase with model capabilities and model reach, can arise along the entire lifecycle of the model, and are influenced by conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails and other factors. In particular, international approaches have so far identified the need to pay attention to risks from potential intentional misuse or unintended issues of control relating to alignment with human intent; chemical, biological, radiological, and nuclear risks, such as the ways in which barriers to entry can be lowered, including for weapons development, design acquisition, or use; offensive cyber capabilities, such as the ways in vulnerability discovery, exploitation, or operational use can be enabled; the effects of interaction and tool use, including for example the capacity to control physical systems and interfere with critical infrastructure; risks from models of making copies of themselves or 'self-replicating' or training other models; the ways in which models can give rise to harmful bias and discrimination with risks to individuals, communities or societies; the facilitation of disinformation or harming privacy with threats to democratic values and human rights; risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community.

Recital 114

The providers of general-purpose AI models presenting systemic risks should be subject, in addition to the obligations provided for providers of general-purpose AI models, to obligations aimed at identifying and mitigating those risks and ensuring an adequate level of cybersecurity protection, regardless of whether it is provided as a standalone model or embedded in an AI system or a product. To achieve those objectives, this Regulation should require providers to perform the necessary model evaluations, in particular prior to its first placing on the market, including conducting and documenting adversarial testing of models, also, as appropriate, through internal or independent external testing. In addition, providers of general-purpose AI models with systemic risks should continuously assess and mitigate systemic risks, including for example by putting in place risk-management policies, such as accountability and governance processes, implementing post-market monitoring, taking appropriate measures along the entire model's lifecycle and cooperating with relevant actors along the AI value chain.

Recital 115

Providers of general-purpose AI models with systemic risks should assess and mitigate possible systemic risks. If, despite efforts to identify and prevent risks related to a general-purpose AI model that may present systemic risks, the development or use of the model causes a serious incident, the general-purpose AI model provider should without undue delay keep track of the incident and report any relevant information and possible corrective measures to the Commission and national competent authorities.

Furthermore, providers should ensure an adequate level of cybersecurity protection for the model and its physical infrastructure, if appropriate, along the entire model lifecycle. Cybersecurity protection related to systemic risks associated with malicious use or attacks should duly consider accidental model leakage, unauthorised releases, circumvention of safety measures, and defence against cyberattacks, unauthorised access or model theft. That protection could be facilitated by securing model weights, algorithms, servers, and data sets, such as through operational security measures for information security, specific cybersecurity policies, adequate technical and established solutions, and cyber and physical access controls, appropriate to the relevant circumstances and the risks involved.

Recital 117

The codes of practice should represent a central tool for the proper compliance with the obligations provided for under this Regulation for providers of general-purpose AI models. Providers should be able to rely on codes of practice to demonstrate compliance with the obligations. By means of implementing acts, the Commission may decide to approve a code of practice and give it a general validity within the Union, or, alternatively, to provide common rules for the implementation of the relevant obligations, if, by the time this Regulation becomes applicable, a code of practice cannot be finalised or is not deemed adequate by the AI Office. Once a harmonised standard is published and assessed as suitable to cover the relevant obligations by the AI Office, compliance with a European harmonised standard should grant providers the presumption of conformity. Providers of general-purpose AI models should furthermore be able to demonstrate compliance using alternative adequate means, if codes of practice or harmonised standards are not available, or they choose not to rely on those.

Recital 164

The AI Office should be able to take the necessary actions to monitor the effective implementation of and compliance with the obligations for providers of general-purpose AI models laid down in this Regulation. The AI Office should be able to investigate possible infringements in accordance with the powers provided for in this Regulation, including by requesting documentation and information, by conducting evaluations, as well as by requesting measures from providers of general-purpose AI models. When conducting evaluations, in order to make use of independent expertise, the AI Office should be able to involve independent experts to carry out the evaluations on its behalf. Compliance with the obligations should be enforceable, *inter alia*, through requests to take appropriate measures, including risk mitigation measures in the case of identified systemic risks as well as restricting the making available on the market, withdrawing or recalling the model. As a safeguard, where needed beyond the procedural rights provided for in this Regulation, providers of general-purpose AI models should have the procedural rights provided for in Article 18 of Regulation (EU) 2019/1020, which should apply mutatis mutandis, without prejudice to more specific procedural rights provided for by this Regulation.

Select bibliography

- Finck M, *The EU Artificial Intelligence Act: A Commentary* (OUP 2026).
- Beurskens M, 'Art. 55 Pflichten der Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko' in David Bomhard, Fritz-Ulli Pieper & Susanne

Wende (eds), *KI-VO: Verordnung über künstliche Intelligenz* (1st edn, Deutscher Fachverlag, 2025).

- Bernsteiner C and Schmitt T, 'Art. 55 Pflichten der Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko' in Mario Martini and Christiane Wendehorst (eds), *KI-VO: Verordnung über Künstliche Intelligenz* (2nd edn, C.H. Beck, 2026).
- Schneider A, 'Artikel 55 Pflichten der Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko' in J Schefzig and R Killan (eds), *Beck'scher Online-Kommentar KI-Recht* (5th edn, C.H. Beck, 2025).
- Joerges C and others, 'European Product Safety, Internal Market Policy and the New Approach to Technical Harmonisation and Standards' (EUI Working Papers LAW, nos. 91/10-14, Florence, 1991).
- Anderljung M and others, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (arXiv, 7 Nov 2023).
- Fraser H, and Villarino J-M, 'Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough.' (2024) 15 European Journal of Risk Regulation 431.
- Teichmann F, 'Risk, Reasonableness and Residual Harm under the EU AI Act: A Conceptual Framework for Proportional Ex-Ante Controls' (2026) European Journal of Risk Regulation.
- Wei K and Heim L, 'Designing Incident Reporting Systems for Harms from General-Purpose AI' in *Proceedings of the AAAI Conference on Artificial Intelligence* (2026) 38016.
- Chatzipanagiotis M, 'Incident Reporting and Investigation Under the AI Act: Some Insights from Aviation' (2026) 34 International Journal of Law and Information Technology eaaf019.
- Nolte H, Rateike M, and Finck M, 'Robustness and Cybersecurity in the EU Artificial Intelligence Act' in *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025) 283.
- Nevo S and others, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models* (RAND 2024).

The authors thank Maarten Herbosch for his invaluable editorial and substantive support with this chapter.

Commentary

1. General remarks

1.1. Introduction

1. Article 55 of the AI Act lays down the additional obligations applicable to providers of general-purpose AI ("GPAI") models presenting systemic risk. These obligations are in addition to those imposed on providers of GPAI models under Article 53 and apply specifically to models classified as presenting systemic risk pursuant to Article 51.[1] In particular, providers must conduct and document state-of-the-art model evaluations, including adversarial testing;[2] assess and mitigate systemic risks at the Union level and their possible sources;[3] keep track of and report relevant information concerning serious incidents;[4] and ensure an adequate level of cybersecurity protection in relation to the model and its physical infrastructure.[5] Article 55(2) then sets out the mechanisms through which providers may demonstrate compliance with those obligations. In particular, providers may rely on harmonised standards or on codes of practice adopted pursuant to Article

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 1689/1 ("AI Act"), art 55(1) and recital 114.
[2] AI Act, art 55(1)(a).
[3] AI Act, art 55(1)(b).
[4] AI Act, art 55(1)(c).
[5] AI Act, art 55(1)(d).

56.[6] Article 55(3) clarifies that any information or documentation obtained pursuant to Article 55, including trade secrets, must be treated in accordance with the confidentiality obligations laid down in Article 78.

2. The inclusion of Article 55 during the trilogue negotiations is also significant for understanding the provision's aims and interpretative difficulties.[7] Neither the Commission proposal nor the Council's general approach contained a regime dedicated to GPAI models with systemic risk. The provision instead emerged in response to rapid developments in GPAI capabilities and increasing policy concerns regarding *AI foundation models* during the later stages of the legislative process.[8] As a result, several concepts central to Article 55, such as *state-of-the-art* or *adequate* cybersecurity protection, remain relatively open-textured. Their interpretation has and will continue to be shaped by forthcoming standards and non-binding guidance, most notably the Safety and Security Chapter of the GPAI Code of Practice.[9]

3. The temporal application of Article 55 also warrants brief clarification. Pursuant to Article 113(3)(b), the provisions of Chapter V, including Article 55, apply from 2 August 2025,[10] while the AI Office will be empowered to assess compliance from 2 August 2026.[11] Providers of general-purpose AI models with systemic risk that were placed on the market before 2 August 2025 may nevertheless benefit from a longer transitional period and must take the necessary steps to comply with the obligations set out in the Regulation by 2 August 2027.[12]

1.2. Structure and overview

4. This chapter analyses Article 55 in the order of the provision, proceeding paragraph by paragraph. The chapter begins with Article 55(1), which introduces the substantive obligations applicable to providers of GPAI models presenting systemic risk. Before turning to the individual obligations contained in subparagraphs (a) through (d), the analysis first addresses the structure and logic of Article 55(1) as a whole. This section also addresses the interpretative value of the Safety and Security Chapter of the GPAI Code of Practice, thereby laying the groundwork for the references made to the Code throughout this chapter as an interpretative aid for resolving ambiguities concerning the operationalisation of Article 55(1).

5. The analysis of Article 55(1)(a) examines the obligation to conduct and document state-of-the-art model evaluations, including adversarial testing. Particular attention is devoted to the meaning of

[6] AI Act, art 55(2); see the commentary on Article 56 in this work.

[7] See Clara Hainsdorf and others, 'Dawn of the EU's AI Act: Political Agreement Reached on World's First Comprehensive Horizontal AI Regulation' (*White & Case*, 14 December 2023) <https://www.whitecase.com/insight-alert/dawn-eus-ai-act-political-agreement-reached-worlds-first-comprehensive-horizontal-ai> accessed 15 May 2026.

[8] Breffni Banks, 'AI Act Definitive Text Endorsed by EU Member States' (*IMRO*, 2 February 2024) <https://imro.ie/industry-news/ai-act-definitive-text-endorsed-by-eu-member-states/> accessed 15 May 2026.

[9] European Commission, 'Code of Practice for General-Purpose AI Models – Safety and Security Chapter' (2025) <https://ec.europa.eu/newsroom/dae/redirection/document/118119>; see the commentary on Article 56 in this work.

[10] AI Act, art 113(b); see the forthcoming commentary on Article 113 in this work.

[11] AI Act, art 113.

[12] AI Act, art 111(3); see the commentary on Article 111 in this work.

the 'state of the art' requirement, the relationship between internal and external evaluations, the role of adversarial testing, and the scope of the accompanying documentation obligations.

6. The discussion then turns to Article 55(1)(b), which requires providers to assess and mitigate possible systemic risks at the Union level. This section analyses the meaning of 'possible systemic risks', the threshold of foreseeability implied by the provision, and the relationship between systemic risk assessment and broader risk management methodologies. It further examines the distinction and interaction between AI safety and AI security risks, as well as the extent to which Article 55(1)(b) imposes ongoing monitoring and mitigation obligations throughout the model lifecycle.

7. The section on Article 55(1)(c) addresses the obligation to keep track of, document and report relevant information about serious incidents and possible corrective measures. It analyses the notion of 'serious incident' in the specific context of GPAI models with systemic risk and examines the relationship between Article 55(1)(c) and the incident reporting framework applicable elsewhere under the AI Act. Particular attention is also devoted to the reporting timelines and the information that needs to be included in reports, as well as to the practical and evidentiary challenges associated with identifying incidents linked to GPAI models with systemic risk.

8. The final subparagraph, Article 55(1)(d), requires providers to ensure an adequate level of cybersecurity protection for GPAI models presenting systemic risk and their physical infrastructure. The accompanying analysis considers the objective scope of protection, the meaning of an 'adequate' level of cybersecurity, and the relationship between cybersecurity obligations and broader systemic risk mitigation measures. It also evaluates how the GPAI Code of Practice operationalises cybersecurity obligations.

9. The discussion of Article 55(2) addresses some of the different compliance pathways available to providers. It analyses the legal role of harmonised standards, approved codes of practice, and alternative adequate means of compliance, while clarifying the legal effects and practical significance of each of these pathways. Particular attention is devoted to the distinction between codes of practice and harmonised standards, as well as to the implications of adherence and non-adherence to the GPAI Code of Practice.

10. Finally, this chapter examines Article 55(3), which governs the confidentiality treatment of information and documentation obtained pursuant to Article 55. This section analyses the relationship between confidentiality protections, trade secrets, cybersecurity concerns, and the supervisory powers of the AI Office and the Commission. It also considers some of the limits of confidentiality claims in the context of public enforcement and regulatory transparency.

2. Substance

2.1. Article 55(1): Additional obligations for providers of GPAI models with systemic risk

11. Article 55(1) introduces a set of substantive obligations for providers of GPAI models with systemic risk. These obligations apply in addition to those already imposed on providers of GPAI models

under Article 53[13] and, where applicable, Article 54 AI Act.[14] The inclusion of these supplementary obligations is justified on the basis that GPAI models with systemic risks may give rise to 'potential significantly negative effects' that exceed those associated with general-purpose AI models more broadly.[15] The severity of systemic risk – defined as a risk capable of propagating at scale across the AI value chain and producing significant impact on the Union market, including through society-wide harm – thus warrants the inclusion of commensurate obligations, in line with the AI Act's underlying risk-based logic.[16] In addition to the severity of systemic risk, Article 55(1) also responds to the inherent unpredictability of the most advanced GPAI models. Systemic risk is understood to increase with model capabilities,[17] which in turn may only emerge and become apparent after market placement, in ways that cannot be predicted or anticipated at the pre-deployment stage.[18] This creates what may be characterised as an epistemic problem as much as a severity problem: the risks are not only uniquely severe but are inherently difficult to anticipate and mitigate, even by those who develop the models.[19] GPAI models with systemic risk therefore present a fundamentally different type of regulatory challenge,[20] which in turn justifies the imposition of the additional obligations under Article 55(1).

[13] See the commentary on Article 53 in this work; AI Act, recital 114: 'The providers of general-purpose AI models presenting systemic risks should be subject, in addition to the obligations provided for providers of general-purpose AI models, to obligations aimed at identifying and mitigating those risks'.

[14] See the commentary on Article 54 in this work.

[15] AI Act, recitals 97 and 104; for instance, the presence of systemic risk precludes the application of exemptions available to other general-purpose AI models. GPAI models with systemic risk are precluded from exemptions as regards the transparency-related requirements imposed on general-purpose AI models.

[16] See AI act, art 3(65) and recital 110; also see on the nature of systemic risk the Code of Practice, Safety and Security Chapter (n 9); see the forthcoming commentary on Article 3(65) in this work.; see also AI Act, recital 26; European Commission, 'Communication from the Commission - Commission Guidelines on the Scope of the Obligations for Providers of General-Purpose AI Models Established by Regulation (EU) 2024/1689 (AI Act)' C(2025) 7719 final ("Commission Guidelines"), para 67; AI Act, recital 109, 'Compliance with the obligations applicable to the providers of general-purpose AI models should be commensurate and proportionate to the type of model provider'.

[17] AI Act, recital 110. Also see the forthcoming commentary on Article 3(65) in this work.

[18] Markus Anderljung and others, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (arXiv, 7 November 2023) <https://doi.org/10.48550/arXiv.2307.03718> accessed 15 May 2026, 10–11.

[19] See Giacomo Zanotti, Daniele Chiffi and Viola Schiaffonati, 'AI-Related Risk: An Epistemological Approach' (2024) 37 Philosophy & Technology 66; Yoshua Bengio and others, 'International AI Safety Report 2026' (DSIT 2026/001, 2026) <https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf> accessed 15 May 2026, 36.

[20] Adrian Schneider, 'Artikel 55 Pflichten der Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko' in Jens Schefzig and Robert Killan (eds), *Beck'scher-Onlinekommentar KI-Recht* (C.H. Beck, 5th edn, 2025) para 3 ['Dies trägt dem Grundgedanken Rechnung, dass bei KI-Modellen mit allgemeinem Verwendungszweck mit zunehmender Rechenleistung auch das Gefahrenpotenzial für Anwender steigt.']; See Thorsten Ammann und Jan Pohle 'KI-Verordnung – Was bisher geschah und jetzt zu tun ist' (2024) Compliance Berater 137 ['Der Grundgedanke ist hier, dass GPAI-Systeme mit zunehmender Rechenleistung tendenziell vermehrt unvorhersehbarere Ergebnisse hervorbringen und damit tendenziell ein höheres Gefahrenpotenzial für den Anwender darstellen.']; Michael Beurskens, 'Art. 55 Pflichten der Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko' in David Bomhard, Fritz-Ulli Pieper, and Susanne Wende (eds), *KI-VO Verordnung über künstliche Intelligenz* (1st edn, Deutscher Fachverlag 2025) para 3. Also see the forthcoming commentary on Article 3(65) in this work.

12. The nature of systemic risk also informs the interpretation of the obligations outlined in Article 55(1),[21] which are to be read in light of the objective to assess and mitigate systemic risk[22] with a degree of scrutiny and level of detail proportionate to the risks involved.[23] It should be underscored that the purpose of the obligations under Article 55(1) is not to prevent systemic risks from materialising altogether, but rather to achieve 'a comprehensive prevention in the sense of minimising the probability of occurrence as much as possible and preparing as effectively as possible should foreseen or unexpected systemic risks materialise.'[24] The obligations apply along the model's entire lifecycle and regardless of whether the GPAI model with systemic risk is provided as a standalone model or embedded in an AI system or a product.[25] This means providers must take appropriate measures along the entire model's lifecycle and cooperate with relevant actors along the AI value chain;[26] the fact that the downstream system complies with the corresponding regulatory requirements does not release the provider of the GPAI model with systemic risk from fulfilling its obligations under Article 55.[27]

13. The provider obligations are outlined across the four subparagraphs of Article 55(1). Under Article 55(1)(a), providers must conduct model evaluations using 'standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks'.[28] Article 55(1)(b) requires providers to 'assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use' of the GPAI model with systemic risk.[29] Pursuant to Article 55(1)(c), providers must 'keep track of, document, and report [...] relevant information about serious incidents and possible corrective measures to address them'.[30] Finally, Article 55(1)(d) obliges providers 'to ensure an adequate level of cybersecurity protection' for both the GPAI model with systemic risk and its physical infrastructure.[31]

14. Each of these obligations will be analysed in turn, following the order of the provision, although the relationship between the obligations themselves is not necessarily sequential. The obligations in

---

[21] See Section 2.1.3.7.; see Code of Practice, Safety and Security Chapter (n 9) recital (i) 'The Signatories recognise that all Commitments and Measures shall be interpreted in light of *the objective to assess and mitigate systemic risks*.' (emphasis added).

[22] See Code of Practice, Safety and Security Chapter (n 9) recital (i) 'The Signatories recognise that all Commitments and Measures shall be interpreted in light of *the objective to assess and mitigate systemic risks*.' (emphasis added).

[23] ibid recital (c); AI Act, art 56(2)(d).

[24] Clemens Bernsteiner and Thomas Rainer Schmitt, 'Art. 55 Pflichten der Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko' in Mario Martini and Christiane Wendehorst (eds), ***KI-VO: Verordnung über Künstliche Intelligenz*** (2nd edn, C.H. Beck, 2026) para 2 (quote translated from German) ['Ziel ist dabei freilich nicht die komplette Vermeidung der Realisierung dieser Risiken, sondern eine umfassende Prävention iSd bestmöglichen Verringerung der Eintrittswahrscheinlichkeit und der bestmöglichen Vorbereitung für den Fall, dass sich vorhergesehene oder unerwartete systemische Risiken doch manifestieren.'].

[25] AI Act, recital 114 ['The providers of general-purpose AI models presenting systemic risks should be subject [...] to obligations aimed at identifying and mitigating those risks [...] regardless of whether it is provided as a standalone model or embedded in an AI system or a product.']

[26] AI Act, recital 114; Code of Practice, Safety and Security Chapter (n 9) recital (a) ['The Signatories recognise that providers of general-purpose AI models with systemic risk should continuously assess and mitigate systemic risks, [...] cooperating with and taking into account relevant actors along the AI value chain']; For an analysis on the notion of 'lifecycle', see Section 2.2.1. in the forthcoming chapter on Modifications in this work.

[27] Beurskens (n 20) para 4.

[28] AI Act, art 55(1)(a).

[29] AI Act, art 55(1)(b).

[30] AI Act, art 55(1)(c).

[31] AI Act, art 55(1)(d).

Article 55(1) are *continuous*,[32] meaning that the risk assessment and mitigation process must be revisited along the model's lifecycle as circumstances change.[33] The obligations are also *iterative*, meaning that the steps of the risk management process are cyclical and repeat themselves, feeding back into one another, at least until all risks have been reduced to an acceptable level.[34] Although Article 55 does not explicitly describe the obligations as such – a notable omission given that the risk management process for high-risk AI systems under Article 9 is expressly characterised as a 'continuous iterative process'[35] – *iteration* is nonetheless a central feature of risk management.[36]

15. This understanding of the obligations as continuous and iterative therefore means that the four subparagraphs should be read as interacting with and reinforcing one another, rather than as discrete and self-contained requirements.[37] This is particularly evident in the relationship between Article 55(1)(a) and 55(1)(b). Article 55(1)(a) focuses on the use of state-of-the-art model evaluations with a view to identifying systemic risk. However, risk identification itself constitutes the first stage of the established risk management pipeline reflected in international standards and seemingly invoked by the EU legislature elsewhere in the AI Act, most notably in the context of high-risk AI systems.[38] Accordingly, the risk identification obligation under Article 55(1)(a) should be understood as forming part of the systemic risk assessment and mitigation obligations imposed under Article 55(1)(b).

16. Similarly, the obligations under Article 55(1)(c) concerning the tracking, documenting, and reporting of relevant information about serious incidents may contribute to identifying previously unrecognised failure modes or expose the inadequacy of existing mitigation measures.[39] Article 55(1)(d), in turn, introduces cybersecurity obligations aimed at protecting the GPAI model with systemic risk and its physical infrastructure against malicious interference or compromise.[40] Although such measures should be distinguished from safety-oriented mitigation measures directed

---

[32] AI Act, recital 114.

[33] ISO, 'Risk Management — Principles and Guidelines' (ISO 2009) ISO 31000:2009(E) <https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-1:v1:en> accessed 15 May 2026, s 3 (j) 'Risk management is dynamic, iterative and responsive to change', and in more detail '[r]isk management continually senses and responds to change. As external and internal events occur, context and knowledge change, monitoring and review of risks take place, new risks emerge, some change, and others disappear.'.

[34] Also see Code of Practice, Safety and Security Chapter (n 9) Measure 4.2 (and, schematically, Figure 3).

[35] AI Act, art 9(1) and recital 65.

[36] Code of Practice, Safety and Security Chapter (n 9) recital (c); Michèle Finck, *The EU Artificial Intelligence Act: A Commentary* (Oxford University Press 2026) para 4.170.

[37] Bernsteiner and Schmitt, 'Art 55' (n 24) para 8 ['Der Anbieter muss eine Modellbewertung durchführen, um systemische Risiken zu ermitteln. Die so entdeckten Risiken, die sich aus der Entwicklung, dem Inverkehrbringen oder der Verwendung des KI-Modells mit allgemeinem Verwendungszweck mit systemischem Risiko ergeben können, sind in der Folge zu bewerten und durch geeignete Maßnahmen zu mindern.']; Schneider (n 20) para 9, ['Die nach Art. 55 Abs. 1 lit. a identifizierten systemischen Risiken sind nach lit. b zu bewerten und zu mindern.']; also see the European Commission, 'Commission Opinion of 1.8.2025 on the assessment of the General-Purpose AI Code of Practice within the meaning of Article 56 of Regulation (EU) 2024/1689' C(2025) 5361 final ("Commission Opinion"), para 35 ['The commitments of the Safety and Security Chapter further contribute to the proper application of Article 55(1) of the AI Act by specifying that the various obligations are not isolated but complement and feed into each other'].

[38] AI Act, art 9(2).

[39] ISO 31000:2009(E) (n 33) s 2.15; See generally Kevin Paeth and Sean McGregor, 'AI Risk, Safety, and Incident Reporting' in Wei Xu (ed), *Handbook of Human-Centered Artificial Intelligence* (Springer 2025) <https://doi.org/10.1007/978-981-97-8440-0_89-1> accessed 15 May 2026.

[40] AI Act, recital 115; See Code of Practice, Safety and Security Chapter (n 9) app 4.1 (3).

at preventing harmful outcomes arising from the model's capabilities or behaviour,[41] they nonetheless remain part of the broader systemic risk management framework established under Article 55. Accordingly, a systematic and purposive reading supports interpreting the obligations under Article 55(1)(a) through (d) as constituting interconnected elements of a continuous systemic risk management process, beginning with risk identification and extending to risk analysis,[42] evaluation,[43] and (safety and security) mitigation.[44] Such a framing is also consistent with how the systemic risk management process is envisaged in the Safety and Security Chapter of the GPAI Code of Practice.[45]

17. In this vein, it is necessary to address the interpretative value of the Safety and Security Chapter of the GPAI Code of Practice. The Safety and Security Chapter was drafted specifically in relation to the obligations imposed on providers under Article 55(1) AI Act.[46] Although the commitments and accompanying measures contained therein do not expressly distinguish between the individual subparagraphs of Article 55(1), the present chapter will, where appropriate, broadly map them onto the corresponding obligations contained in Article 55(1)(a) through (d). At the same time, it must be emphasised that the Code of Practice remains a voluntary and therefore non-binding instrument.[47] The commitments and measures contained therein do not themselves impose legal obligations on providers and therefore cannot determine the meaning of Article 55(1) as a matter of law.[48]

18. Nevertheless, the interpretative significance of the GPAI Code of Practice should not be understated. The fact that the European Commission has deemed the Safety and Security Chapter as an acceptable means through which providers may demonstrate compliance with Article 55(1) lends the commitments and measures contained therein particular practical and interpretative relevance.[49] In that respect, the Code may be regarded as indicative of how the Commission understands the operationalisation of the obligations imposed under Article 55(1) and thus as carrying interpretative weight for a purposive reading of the provision.[50] The Code of Practice likewise plays a crucial role in the Commission's assessment of 'alternative adequate means' that

---

[41] Zhiqiang Lin, Huan Sun and Ness Shroff, 'AI Safety vs. AI Security: Demystifying the Distinction and Boundaries' (arXiv, 21 June 2025) <https://doi.org/10.48550/arXiv.2506.18932> accessed 15 May 2026, 6; Xiangyu Qi and others, 'AI Risk Management Should Incorporate Both Safety and Security' (arXiv, 29 May 2024) <https://doi.org/10.48550/arXiv.2405.19524> accessed 15 May 2026, 10.

[42] See Section 2.1.2.2.

[43] See Section 2.1.2.2.1.3. on systemic risk acceptance determination in the Code of Practice which corresponds to risk evaluation in established risk management literature; ISO, 'Risk Management — Vocabulary' (ISO 2009) ISO Guide 73:2009 <https://www.iso.org/obp/ui/#iso:std:iso:guide:73:ed-1:v1:en> accessed 15 May 2026, s 3.7.

[44] Jonas Schuett, 'Risk Management in the Artificial Intelligence Act' (2024) 15 European Journal of Risk Regulation 367, 368; Anthony M Barrett and others, 'AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models' (arXiv, 30 June 2025) <https://doi.org/10.48550/arXiv.2506.23949> accessed 15 May 2026, 7; see also Leonie Koessler and Jonas Schuett, 'Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries' (arXiv, 17 July 2023) <https://doi.org/10.48550/arXiv.2307.08823> accessed 15 May 2026.

[45] Code of Practice, Safety and Security Chapter (n 9) Commitment 1, Measure 1.2.

[46] European Commission, 'The General-Purpose AI Code of Practice' (2026) <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai> accessed 15 May 2026.

[47] Also see the commentary on Article 56, Section 2.7. in this work.

[48] See the commentary on Article 56, para 98 in this work.

[49] Also see the commentary on Article 56, Section 2.6. in this work.

[50] Code of Practice, Safety and Security Chapter (n 9) recital (i).

providers that do not adhere to the Code might implement – underlining the Code's *de facto* key role in interpreting the obligations in Article 55(1).[51]

19. This chapter identifies ambiguities arising from the text of Article 55(1), as well as uncertainties concerning its practical operationalisation, and analyses the strongest interpretative approaches to resolving them through textual, systematic, and teleological interpretation of the AI Act. Within that interpretative framework, the Safety and Security Chapter of the GPAI Code of Practice serves as an important interpretative reference point and, in several instances, provides particularly compelling guidance for resolving ambiguities concerning the operationalisation of the obligations imposed under Article 55(1). At the same time, consistent with this commentary's broader commitment to interpretive optionality, alternative interpretations are identified and discussed where they remain legally plausible.

2.1.1. Article 55(1)(a): Conducting state-of-the-art model evaluations, including adversarial testing

20. Article 55(1)(a) requires providers to 'perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks'. Recital 114 further supplies that 'the *necessary* model evaluations' are to be conducted 'in particular prior to its first placing on the market', including through adversarial testing and, where appropriate, 'through internal or independent external testing'.[52]

21. This section proceeds on the basis that Article 55(1)(a) primarily concerns the identification of systemic risk through state-of-the-art model evaluations, while recognising its functional relationship with the broader systemic risk assessment and mitigation obligations imposed under Article 55(1)(b).[53] It should also be reiterated that other practices and techniques aside from model evaluations can be used for identifying systemic risk.[54] Likewise, model evaluations may be used at multiple stages of the risk assessment and mitigation process, including during systemic risk analysis and the development of risk mitigation measures.[55]

22. The analysis of Article 55(1)(a) proceeds through several closely related questions concerning the scope and operationalisation of the obligation. The first question concerns the obligation to conduct *state-of-the-art* model evaluations and in particular how the notion of *state of the art* should be understood as a benchmark for the appropriateness of the evaluations conducted in the context of systemic risk assessment and mitigation. The analysis then turns to the concept of *model evaluation* itself, including the meaning of adversarial testing as a type of model evaluation expressly listed both in Article 55(1)(a) and the supporting Recital 114. A related question concerns whether Article 55(1)(a) contemplates only internal model evaluations or also requires independent external

[51] Also see Section 2.2.2.
[52] AI Act, recital 114.
[53] See para 20.
[54] See Section 2.1.2.2.1.1.; Code of Practice, Safety and Security Chapter (n 9) Commitment 2, Measures 2.1 and 2.2.
[55] See para 39; Code of Practice, Safety and Security Chapter (n 9) Commitment 3, Measure 3.3; Code of Practice, Safety and Security Chapter (n 9) Commitment 5, Measure 5.1 (adversarial methods should be used to assess whether the implemented safety mitigations are appropriate).

evaluations.[56] Finally, this section considers whether the obligation to conduct and document adversarial testing – as expressly required under Article 55(1)(a) – extends beyond the evaluations themselves to encompass broader documentation obligations relating to the systemic risk assessment and mitigation process.

2.1.1.1. The 'state of the art' condition

23. Article 55(1)(a) requires that model evaluations conducted by providers reflect the state of the art. In this context, the *state of the art* functions as a dynamic reference that avoids fixing detailed technical requirements in the text of the provision and instead tethers the threshold of compliance to evolving scientific and technical knowledge.[57] The legally binding obligation therefore remains compliance with the statutory standard established under Article 55(1)(a),[58] while the concrete measures, methodologies, and evaluation techniques capable of satisfying that standard may evolve over time.[59] Outside of Article 55, the AI Act requires providers of high-risk AI systems to ensure compliance with applicable requirements by 'taking into account their intended purpose as well as the *generally acknowledged state of the art* on AI and AI-related technologies'.[60] Any codes of practice and harmonised standards developed as means for providers to demonstrate conformity with the requirements of the AI Act are likewise expected to reflect the *state of the art*.[61]

---

[56] Section 2.1.1.2.3.

[57] Manfred Kohler, 'A New Role for Standards in the EU Regulatory System' (*How to Regulate? The Regulatory Institute's Blog*, 10 July 2025) <https://howtoregulate.org/a-new-role-for-standards-in-the-eu-regulatory-system/> accessed 22 April 2026; Christian Joerges and others, 'European Product Safety, Internal Market Policy and the New Approach to Technical Harmonisation and Standards' (European University Institute Florence 1991) Working Papers LAW 91/10–14 <https://hdl.handle.net/1814/46244> accessed 15 May 2026, 84; Valerie Thomas, Ajda Mihelčič and Manfred Kohler, *How to Regulate: A Handbook Presenting Regulatory Techniques of 47 Jurisdictions and a Basic Universal Method* (2nd edn, Regulatory Institute 2021) <https://www.howtoregulate.org/wp-content/uploads/2024/10/H2R_hanbook_2024.pdf> accessed 15 May 2026, 117; Simon Gerdemann, 'Artikel 8 Einhaltung der Anforderungen' in Jens Schefzig and Robert Killan (eds), *Beck'scher Onlinekommentar KI-Recht* (C.H. Beck, 5th edn, 2025) para 11; Sandra Schmitz, 'Conceptualising the Legal Notion of "State of the Art" in the Context of IT Security' *Privacy and Identity Management. Between Data Protection and Security* (Springer 2022) <https://doi.org/10.1007/978-3-030-99100-5_3> accessed 21 May 2026, 27.

[58] Article 55(1)(a) requires model evaluations to 'reflect' the state of the art, whereas article 8(1) requires that high-risk AI systems should be 'taking into account' the acknowledged state of the art. *Taking into account* had been differentiated from '"compliance", due to the fact that "state of the art" is not a legally defined concept and it involves several dynamic and complex aspects, difficult to be expressed in a single and clear definition.' (see Medical Device Group, 'Guidance on standardisation for medical devices' (2021) MDCG 2021-5 Rev. 1 <https://health.ec.europa.eu/document/download/59ac4cb0-f187-4ca2-814d-82c42cde5408_en> accessed 15 May 2026, 16). Whether requiring evaluations to 'reflect' the state of the art imposes a more stringent obligation than merely 'taking [it] into account' remains interpretatively unresolved, although a textual argument could be made that 'reflect' suggests a stronger degree of alignment or correspondence than mere consideration.

[59] Joerges and others (n 57) 84; See also Yuan Shi, '"State-of-the-Art" in New EU Medical Device Regulations: A Review of Its Development in Medical Device Law, the Interpretations from Stakeholders, Impacts, and Possible Solutions for Implementation' (Master's Thesis, University of Bonn 2022) <https://www.dgra.de/media/masterthesis/1398-master_shi_yuan_2022.pdf> accessed 15 May 2026, 13.

[60] AI Act, art 8(1); Gerdemann (n 57) para 11; AI Act, recital 64; also see AI Act, recital 65, which says that the risk management measures must be developed in light of 'the state of the art in AI' and not the 'generally acknowledged state of the art'.

[61] AI Act, recitals 116 and 121; also see Commission Notice, The 'Blue Guide' on the Implementation of EU Product Rules 2022 [2022] OJ C247/1 ("Blue Guide") s 4.1.2.4, 53, which states, the 'concept of essential requirements is based on the assumption that the harmonised standards reflect generally acknowledgeable state of the art and the CEN, CENELEC or ETSI review standards regularly in accordance with the relevant standardisation request' [emphasis added and rephrased] as example that 'state of the art' and 'generally acknowledged state of the art' seem to be used synonymously in the AI Act; On the difficulties associated with applying traditional harmonised

24. While the concrete risk assessment and mitigation measures capable of satisfying the state-of-the-art condition may evolve over time, the reference itself cannot remain entirely indeterminate.[62] A helpful starting point to contextualise the subsequent analysis is the three-stage theory developed by the German Federal Constitutional Court in the 1978 *Kalkar I* decision.[63] Although not binding on the EU institutions or on the interpretation of EU law,[64] the decision has served as an influential conceptual reference point for distinguishing between different levels of technological development reflected in legal standards.[65] The Court situated the state of the art between, on the one hand, the *generally accepted rules of technology*, reflecting the 'prevailing opinion among technical practitioners',[66] and, on the other hand, the *state of scientific knowledge and research*, encompassing the latest scientific and technical developments irrespective of their practical feasibility.[67] The *generally accepted rules of technology* have also been described as broadly corresponding to *best practices*.[68] Best practices are those measures that have proven themselves *in practice* and are hardly subject to methodological modernisation.[69] Best practices often serve as 'the minimum basis for state of the art,'[70] or, put differently, that the 'state of the art at least corresponds to best available [practices]'.[71]

25. Within this framework, the state of the art occupies an intermediate position between *generally accepted rules of technology* and the *state of scientific knowledge and research*: it exceeds what is merely generally accepted or routinely implemented in practice but does not extend to the furthest frontier of scientific research and development.[72] Nor are the boundaries between these three categories of technical standards impermeable. Measures that initially qualify as representing the state of scientific knowledge and research will, upon market introduction and subsequently being

---

standardisation processes to GPAI models, see also Hadrien Pouget and Ranj Zuhdi, 'AI and Product Safety Standards Under the EU AI Act' (*Carnegie Endowment for International Peace*, 5 March 2024) <https://carnegieendowment.org/research/2024/03/ai-and-product-safety-standards-under-the-eu-ai-act> accessed 15 May 2026 [speaks on why standards are harder to change so having a very dynamic state of the art is a hard threshold to meet for an instrument that does not lend itself to quick updates – see how it takes 3 years to develop].

[62] See, for example, Joerges and others (n 57) 14; Harm Schepel, *The Constitution of Private Governance: Product Standards in the Regulation of Integrating Markets* (Bloomsbury 2005) <https://www.bloomsbury.com/uk/constitution-of-private-governance-9781847311078/> accessed 15 May 2026, 374.

[63] Case 2 BvL 8/77 *Kalkar Case 1* (1978) BVerfGE 49, 89 (135 et seq.).

[64] See Mehrdad Payandeh, 'Constitutional Review of EU Law after "Honeywell": Contextualizing the Relationship between the German Constitutional Court and the EU Court of Justice' (2011) 48 Common Market Law Review 9.

[65] Mark Seibel, 'Abgrenzung der "allgemein anerkannten Regeln der Technik" vom "Stand der Technik"' [2013] Neue Juristische Wochenschrift 3000; Luise Eder and others, 'Determining the State of the Art in General-Purpose AI Risk Management: From Code to Practice' (Oxford Martin AI Governance Initiative 2026) Research Memo <https://aigi.ox.ac.uk/publications/determining-the-state-of-the-art-in-general-purpose-ai-risk-management-from-code-to-practice/> accessed 28 May 2026, 6; Schmitz (n 57) 5.

[66] *Kalkar Case 1* (n 63) (135 et seq.) para 99; See also Joerges and others (n 57) 83.

[67] *Kalkar Case 1* (n 63) (135 et seq.) para 101.

[68] Morad Abou Nasser and others, 'Guideline "State of the Art" in IT Security: Technical and Organisational Measures' (TeleTrusT 2025) <https://www.teletrust.de/fileadmin/user_upload/2025-09_TeleTrusT_Guideline_State_of_the_art_in_IT_security_EN.pdf> accessed 15 May 2026, 14; Michael Robert, 'Standardization and the State of the Art' (Kommission Arbeitsschutz und Normung 2021) KanBrief 2/21 <https://www.kan.de/en/publications/kanbrief/2021/2-21/standardization-and-the-state-of-the-art/> accessed 15 May 2026 ['allgemein anerkannte Regeln der Technik' or 'also known as *generally accepted good practices*'].

[69] Abou Nasser and others (n 68) 14.

[70] Schmitz (n 57) 28.

[71] ibid 27.

[72] *Kalkar Case 1* (n 63) (135 et seq.) para 100.

proven in practice, move into the category of state of the art.[73] With increasing standardisation, distribution, and market recognition, such measures may become widely deployed and routinely reflected in technical standards.[74] As a result, their level of innovation may decline as they come to be recognised as generally accepted rules of technology.[75] Distinguishing when a measure ceases to be state of the art and instead becomes a generally accepted rule of technology also remains challenging since general recognition and practical validation alone are not decisive.[76] Best practices can remain widely used even where their effectiveness has declined, whereas the state of the art excludes measures that no longer provide adequate safety assurance despite continued availability on the market or standardisation.[77]

26. Situating the state of the art within this broader spectrum of dynamic references is not merely conceptual. The position attributed to the state-of-the-art condition under Article 55(1)(a) directly influences the level of diligence expected of providers when conducting systemic risk evaluations and implementing corresponding mitigation measures. Two principal interpretative pathways are explored in this chapter.

27. First, the state-of-the-art condition in Article 55(1)(a) could be understood in line with the definition provided in the Safety and Security Chapter of the GPAI Code of Practice – that is, 'the forefront of relevant research, governance, and technology that goes beyond best practice'.[78] That definition was developed specifically in the context of Article 55 and has been endorsed by the European Commission as an adequate means for demonstrating compliance with the provision.[79] Alternatively, the state-of-the-art condition in Article 55(1)(a) could be interpreted in line with the systematic reading of the AI Act by reference to the *generally acknowledged state of the art* applied to high-risk AI systems under Article 8(1).[80] Depending on which interpretation is adopted, and therefore where the state-of-the-art condition is situated along the spectrum between established best practices and the frontier of technological development,[81] the provision will shape both the level of rigour, breadth, and depth expected of providers when conducting model evaluations and the degree of technical effort and innovation they are expected to invest in advancing AI safety and security practices more broadly.[82]

2.1.1.1.1. State of the art as corresponding to the frontier of scientific knowledge and research

28. As indicated, one approach to interpreting the state-of-the-art condition in Article 55(1)(a) is to locate it closer to the domain of scientific research and technological development, in a reading that would prioritise active innovation over measures that are merely generally recognised and proven in practice. Indeed, the definition of state of the art developed in the GPAI Code of Practice points

[73] Abou Nasser and others (n 68) 14.
[74] ibid.
[75] ibid.
[76] Schmitz (n 57) 6 ['the legal benchmark of what constitutes state of the art [has been] shifted to the front of technical development, since general recognition and practical validation alone are not decisive for the state of the art of a technology.'].
[77] ibid.
[78] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'state of the art'.
[79] AI Act, art 55(2); Commission Opinion (n 37).
[80] AI Act, art 8(1); AI Act, recital 64.
[81] See paras 25-26.
[82] Code of Practice, Safety and Security Chapter (n 9) apps 3.1 and 3.3; see also Carlos Mougan and others, 'The Science and Practice of Proportionality in AI Risk Evaluations' (2026) 391 Science 769.

toward this reading. The Safety and Security Chapter defines state of the art as 'the forefront of relevant research, governance, and technology that goes beyond best practice.'[83] Best practices, in turn, are defined as the 'processes, measures, methodologies, methods, and techniques [...] accepted amongst providers of general-purpose AI models with systemic risk as [those] that best assess and mitigate systemic risks at any given point in time.'[84] Measures that qualify as state of the art would be those that 'demonstrate equal or superior safety or security outcomes through alternative means that achieve greater efficiency [compared to approaches accepted as best practice].'[85] The EU regulator may therefore assess whether measures reflect the state-of-the-art condition not by reference to what is 'generally recognised or established in practice, but what is technically necessary, appropriate and possible, even if commercial practice is not yet in line with it'.[86]

29. A purposive reading of Article 55(1)(a) likewise supports interpreting the state-of-the-art condition as referring to a more dynamic standard rather than merely one calibrated to industry practice. As the German Federal Constitutional Court recognised in *Kalkar I*, where legislation seeks to regulate technologically complex and rapidly evolving risks:

> *it would not only be inappropriate, but actually contrary to the purpose of the regulation, if the legislature were to establish safety requirements through normative provisions oriented towards the status quo of technological development. [...] A legislator who strives in this way to keep pace with technological developments, specifically in the interest of intensifying safety, can hardly be faulted.*[87]

30. There are policy considerations that support this interpretation. Setting the state-of-the-art condition above best practices, and thus pursuing what was described above as state of the art at the forefront of technical advancement, creates conditions under which providers are not only required to invest in safety but also rewarded for doing so. This incentive structure mirrors insights developed in the context of the state-of-the-art defence as found in US liability law, where it has been argued that 'average safety increases when a state of the art defence is based on the technological advancement test.'[88] Under a regime in which the state of the art is equated with industry-wide best practices, providers would have little incentive to invest in costly safety improvements, as placing a safer model on the market does not reduce exposure to regulatory scrutiny or enforcement.[89] However, where the state of the art is linked to technical advancement, a subset of safety leaders may invest in demonstrably higher levels of protection, giving rise to a form of race-to-the-top 'safety contest' that rewards performance relative to peers.[90] While this theory was originally developed in the context

[83] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'state of the art'.
[84] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'best practices'.
[85] Code of Practice, Safety and Security Chapter (n 9) recital (f).
[86] Joerges and others (n 57) 83.
[87] *Kalkar Case 1* (n 63) para 58.
[88] James Boyd and Daniel E Ingberman, 'Should "State of the Art" Safety Be a Defense Against Liability?' (Resources for the Future 1995) Discussion Paper 96–01 <https://media.rff.org/documents/9601.pdf> accessed 16 May 2026, 5; Schmitz (n 57) 26 (The state-of-the-art defence, as laid out in the Product Liability Directive, may be invoked by an economic operator to avoid liability for damage caused by a defective product if the objective state of scientific and technical knowledge at the time when the product was placed on the market or put into service was not such as to enable the defect to be discovered.) Also see Article 11(e) 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 [2024] OJ L 2853/1 ("Product Liability Directive"), art 11(e).
[89] Boyd and Ingberman (n 88).
[90] Boyd and Ingberman (n 88).

of liability law, it can also be discerned through a teleological reading of the Code of Practice and its recitals, which emphasise the objective of fostering a culture of innovation in AI safety and security.[91]

31. Compliance with the state-of-the-art condition under this interpretation is therefore not a one-off exercise but a continuous process.[92] It requires providers to remain attentive to developments in the broader field of model evaluation and update accordingly.[93] Such developments could constitute reasonable grounds to question whether earlier assessments of systemic risks remain valid.[94] Where such grounds arise, providers who are signatories to the Code of Practice are therefore required to revisit their systemic risk assessment and mitigation process and update their Safety and Security Model Report accordingly within a reasonable amount of time.[95] Notably, the definition adopted in the Code of Practice, and the interpretation it endorses, are likely to have implications not only for providers who are signatories to the Code of Practice but for all providers subject to the obligations under Article 55.[96] It is, however, not implausible that a claim could be brought on the basis of a systemic interpretation of the AI Act, arguing that the state-of-the-art condition in Article 55(1)(a) should instead be read in line with the way state of the art has been defined in the context of Chapter III on high-risk AI systems. After all, appropriate legislative drafting demands substantive consistency across the legislation, meaning that defined terms must be used uniformly and that their content should not diverge from the definitions provided;[97] accordingly, the definition of state of the art should be consistent across the AI Act.

2.1.1.1.2. Systematic alignment of state of the art with the 'generally acknowledged state of the art'

32. An alternative approach to interpreting the state-of-the-art condition in Article 55(1)(a) is to understand it, through a systematic reading of the AI Act, as corresponding more closely to what the Regulation refers to as the 'generally acknowledged state of the art' in Article 8(1). The qualifier *generally acknowledged* was introduced during the legislative process to replace an earlier reference in the AI Act draft to the *current state of the art.*[98] The European Parliament had previously proposed to define the state of the art as 'the level of development of technical capabilities at a given point in time with regard to products, processes and services, based on the relevant consolidated knowledge of science, technology and experience'.[99] While that definition retained a clear temporal

[91] Code of Practice, Safety and Security Chapter (n 9) recital (f) says that the Safety and Security Chapter 'encourage[s] providers of general-purpose AI models with systemic risk to advance the state of the art in AI safety and security and related processes and measures.'

[92] Code of Practice, Safety and Security Chapter (n 9) Measure 1.2.

[93] Providers must update their Safety and Security Model Report if developments have occurred that 'materially improve the state of the art of model evaluation methods', Code of Practice, Safety and Security Chapter (n 9) Measure 7.6(5).

[94] Code of Practice, Safety and Security Chapter (n 9) Measure 7.6(5).

[95] Code of Practice, Safety and Security Chapter (n 9) Measure 7.6.

[96] See para 17–18.

[97] European Parliament, Council of the European Union and European Commission, *Joint Handbook for the Presentation and Drafting of Acts Subject to the Ordinary Legislative Procedure* (2023) <https://www.consilium.europa.eu/media/67390/joint_handbook_en_01-october-2023_clean_def_final.pdf> accessed 16 May 2026, 17.

[98] Braun Binder and Catherine Egli, 'Art. 8 Einhaltung der Anforderungen' in Mario Martini and Christiane Wendehorst (eds), *KI-VO: Verordnung über Künstliche Intelligenz* (2nd edn, C.H. Beck, 2026) para 29.

[99] 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

and therefore dynamic element, tethering the standard to developments 'at a given point in time', the introduction of the qualifier 'generally acknowledged' arguably shifts the emphasis away from the frontier of technological development and towards methods and techniques that have achieved broader professional recognition and validation in practice.

33. This understanding has also been reflected in commentaries on Article 8 AI Act,[100] where the 'generally acknowledged state of the art' has been distinguished from the *latest* 'state of the art' as developed by the most innovative industry leaders and instead associated with 'the generally accepted level of technical risk minimization among providers of the same type of AI system'.[101] In effect, this interpretation brings the 'generally acknowledged state of the art' closer to what the *Kalkar I* framework described as the generally accepted rules of technology. In fact, the coherence of the standard of 'generally acknowledged state of the art' has itself been questioned on the basis that 'there is no such thing as a standard of "generally accepted state of the art"'.[102] Rather, the relevant standard may more plausibly refer either to the 'state of the art' or to the 'generally accepted rules of technology', as understood in the *Kalkar I* decision, but not to a combination of the two. Combining both standards risks collapsing the distinction between measures that are merely established and validated in practice and those that reflect the current level of technological advancement.[103]

34. The AI Act reiterates that harmonised standards which providers of high-risk AI systems and GPAI models may rely upon to demonstrate compliance with the obligations must reflect the *state of the art*.[104] In its *Draft Standardization Request in Support of Safe and Trustworthy Artificial Intelligence* for high-risk systems, the European Commission subsequently defines 'state of the art' as meaning 'a developed stage of technical capability at a given time [...] based on the relevant consolidated findings of science, technology and experience and which is accepted as good practice in technology.'[105] In that context, the Commission goes further to say that 'the state of the art does not necessarily imply the latest scientific research still in an experimental stage or with insufficient technological maturity.'[106] Compared to the definition of state of the art adopted in the Code of

---

Intelligence Act) and amending certain Union legislative acts COM (2021) 0206 COD (2021) 0106, Document P9 [2023] 0236 ("Parliament Amendments"), amendment 210.

[100] Binder and Egli, 'Art 8' (n 98) paras 21–22.

[101] Gerdemann (n 57) para 11.

[102] Mark Seibel, 'Differentiation Between the "Generally Accepted Rules of Technology" and the "State of the Art"' (2013) Neue Juristische Wochenschrift 3000.

[103] ibid.

[104] AI Act, recital 121 and art 40(1); Blue Guide (n 61) s 4.1.2.4; Robert (n 68).

[105] European Commission, 'Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence' C(2023) 3215 final Annex, 2.

[106] ibid; This definition has been criticised as circular, on the grounds that it remains unclear whether technical standards are required to conform to an independently determined developed stage of technical capability or whether they themselves will determine what counts as state of the art, see Henry Fraser and José-Miguel Bello y Villarino, 'Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough' (2024) 15 European Journal of Risk Regulation 431, 437; The Commission's definition also closely follows the wording of the ISO standards designed for the risk management in medical devices. In that context, *state of the art* 'embodies what is currently and generally accepted as good practice. The state of the art does not necessarily imply the most technologically advanced solution. The state of the art described here is sometimes referred to as the "generally acknowledged state of the art", see Raje Devanathan and Virginia Anastassova, 'ALARP to AFAP, the MDR and ISO 14971:2019+A11:2021' (*StarFish Medical*, 24 April 2026) <https://starfishmedical.com/resource/medical-device-risk-management-and-the-change-from-alarp-to-afap/> accessed 16 May 2026 and ISO, 'Medical Devices — Application of Risk Management to Medical Devices' (ISO 2007) ISO 14971:2007 <https://www.iso.org/standard/38193.html> accessed 15 May 2026.

Practice, the definition of state of the art as interpreted under Chapter III thus corresponds more closely to what the Code of Practice defines as *best practices*, which describes measures that 'best assess and mitigate systemic risks at any given point in time.'[107] Indeed, where providers of high-risk AI systems do not rely on harmonised standards that reflect the state of the art,[108] it is advisable for them to take account of 'already known industry standards relating to AI systems with similar purpose and algorithmic functionality, and document this accordingly.'[109]

35. If the state-of-the-art condition in Article 55(1)(a) were to be interpreted in line with the 'generally acknowledged state of the art' referred to in Chapter III, the obligation under Article 55(1)(a) could be described as more provider-friendly[110] because it would be sufficient for providers to follow the lead of comparable GPAI model providers rather than respond to the measures adopted by the most innovative industry leaders.[111] This lateral comparison is particularly useful insofar as it reinforces the product-safety-oriented understanding of the state of the art underpinning the regulation of high-risk AI systems. At the same time, however, there are important reasons to question whether equivalent interpretations should automatically be transposed to the GPAI model rules. For one, the extent to which the AI Act can be understood as product-safety legislation when applied to general-purpose AI models remains contested,[112] particularly given that existing AI standards do not directly address GPAI models, thereby leaving open what constitutes state-of-the-art safety and security mitigations in this context.[113]

2.1.1.2. Model evaluations that reflect the state of the art

36. Article 55(1) introduces model evaluations as a mechanism through which systemic risks are to be assessed and mitigated. In the absence of an operative definition of *model evaluation* in the AI Act, reference may be made to the GPAI Code of Practice's functional understanding of model evaluations.[114] The latter defines model evaluations as a 'systemic risk assessment technique that can be used at all stages of systemic risk assessment', which in turn includes all methods of systemic risk identification, analysis and acceptance determination.[115] This definition is ostensibly broad, allowing for different methods and techniques provided that these are 'appropriate for the model and the systemic risk'.[116] Adversarial testing is identified as one example of state-of-the-art model evaluations providers must conduct.[117] Other types of model evaluation techniques acknowledged by the

[107] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'best practices'.

[108] AI Act, recital 121, 'Standardisation should play a key role to provide technical solutions to providers to ensure compliance with this Regulation, in line with the state of the art, to promote innovation as well as competitiveness and growth in the single market. Compliance with harmonised standards as defined in Article 2, point (1)(c), of Regulation (EU) No 1025/2012 of the European Parliament and of the Council (41), which are normally expected to reflect the state of the art, should be a means for providers to demonstrate conformity with the requirements of this Regulation.'

[109] Gerdemann (n 57) para 11.

[110] Binder and Egli, 'Art 8' (n 98), para 29 ['risks always lagging somewhat behind the latest developments. [...] However, providers are free to consider not only the generally accepted but also the current state, which is why the wording [of the provision] should be understood as [favourable] for providers.' (translated from German)].

[111] Gerdemann (n 57) para 11 (for AI systems).

[112] Also see the forthcoming chapter on Product, Model and Entity Regulation in this work.

[113] Pouget and Zuhdi (n 61)

[114] See para 17–18.

[115] Code of Practice, Safety and Security Chapter (n 9) Glossary, definitions of 'model evaluation' and 'systemic risk assessment'.

[116] Code of Practice, Safety and Security Chapter (n 9) Measure 3.2.

[117] AI Act, art 55(1)(a).

European Commission as suitable for the purposes of demonstrating compliance with Article 55(1)(a) include 'Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and/or proxy evaluations for classified materials.'[118]

37. While an industry-wide accepted definition of model evaluation remains elusive, and those that have been formulated are similarly broad to that found in the Code of Practice,[119] technical literature may provide supplementary insight into how such evaluations are typically categorised. For instance, the Frontier Model Forum ("FMF"), an industry-supported organisation whose members include providers that are both signatories to the Code of Practice or otherwise subject to the obligations under Article 55,[120] identifies benchmark evaluations, red-team exercises, and controlled studies as the three main categories that capture most existing evaluation tasks.[121] The evaluation techniques listed under Measure 3.2 can broadly be captured under these three categories.[122] Model evaluations can also be distinguished by their focus, such as capability evaluations and propensity evaluations,[123] where capability evaluations measure whether a model has certain dangerous capabilities, while propensity evaluations capture whether the model has the propensity to harmfully apply those capabilities.[124]

38. Without being exhaustive,[125] the preceding paragraphs already illustrate the breadth and fragmentation of the evaluation landscape. This diversity could create practical challenges for both providers and regulators when determining what constitutes state-of-the-art model evaluations under Article 55. In practice, model evaluations are conducted using a combination of publicly available benchmarks and bespoke, in-house evaluation methods.[126] Evaluations may be developed and run internally by providers, while others are outsourced to specialised third parties with controlled access to the model.[127] As a result, some of the most advanced evaluation techniques could remain

[118] Code of Practice, Safety and Security Chapter (n 9) Measure 3.2.

[119] For example, Frontier Model Forum, 'Frontier Capability Assessments' (Frontier Model Forum 2025) Technical Report <https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/> accessed 16 May 2026, 1 [model definitions defined as 'structured tests of model capabilities in a given domain [...] followed by analysis on the test results.'].

[120] 'Membership' (*Frontier Model Forum*, 2026) <https://www.frontiermodelforum.org/membership/> accessed 16 May 2026; European Commission, 'The General-Purpose AI Code of Practice' (n 46).

[121] Frontier Model Forum, 'Preliminary Taxonomy of Pre-Deployment Frontier AI Safety Evaluations' (Frontier Model Forum 2024) Issue Brief <https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations/> accessed 16 May 2026.

[122] Q&A sets and task-based evaluations fall under benchmark evaluations; red-team exercises include red-teaming and other adversarial testing methods; and controlled studies capture human uplift studies, model organisms, simulations, and proxy evaluations for classified materials. See para 36.

[123] Toby Shevlane and others, 'Model Evaluation for Extreme Risks' (arXiv, 22 September 2023) <https://doi.org/10.48550/arXiv.2305.15324> accessed 16 May 2026, 2; See also on the difference between capability assessments and bottleneck assessments: Frontier Model Forum, 'Frontier Capability Assessments' (n 119) 5.

[124] Shevlane and others (n 123) 2.

[125] See Connor Dunlop, 'General Purpose AI Models with Systemic Risks – Classification and Specific Obligations (Articles 51, 52, 55)' in Gianclaudio Malgieri and others (eds) *The EU Artificial Intelligence Act: A Thematic Commentary* (Hart Publishing 2026) 403, 410.

[126] Elliot Jones, Mahi Hardalupas and William Agnew, 'Under the Radar? Examining the Evaluation of Foundation Models' (Ada Lovelace Institute 2024) <https://www.adalovelaceinstitute.org/report/under-the-radar/> accessed 16 May 2026.

[127] See METR, 'Details about METR's Evaluation of OpenAI GPT-5.1-Codex-Max' (2025) <https://metr.org/evaluations/gpt-5-1-codex-max-report/> accessed 16 May 2026; See OpenAI reporting on third-

proprietary and therefore not generally available on the market,[128] making it difficult for other providers to ensure that their own model evaluation practices reflect the state of the art.[129]

39. The Safety and Security Chapter of the Code of Practice offers guidance to providers in determining what qualifies as state-of-the-art model evaluations, as well as how such evaluations may be conducted in a manner that demonstrates compliance with Article 55(1)(a).[130] It further clarifies that model evaluations are 'integral along the entire model lifecycle.'[131] At the stage of systemic risk identification, model evaluation techniques, such as red-teaming, may be used to reveal unexpected capabilities or limitations of a model and thereby uncover risks inherent to it.[132] At the risk assessment stage, adversarial evaluation techniques draw upon and feed into risk modelling[133] to the extent that it requires providers to foresee and account for how malicious actors could circumvent model safeguards.[134] The results of these evaluations help determine which safety and security mitigations are appropriate.[135] Model evaluation techniques may also be used to test the effectiveness of those mitigations, for example by assessing whether safety mitigations successfully prevent the model from exhibiting previously identified harmful behaviours.[136]

---

party assessments in their system card, OpenAI, 'GPT-4o System Card' (OpenAI 2024) <https://cdn.openai.com/gpt-4o-system-card.pdf> accessed 16 May 2026, s 15.

[128] On limitations around reproducibility, see Patricia Paskov, Lisa Soder and Everett Smith, 'Toward Best Practices for AI Evaluation and Governance: A Proposal for a European Union General-Purpose AI Model Evaluation Standards Task Force' (RAND 2025) <https://www.rand.org/pubs/perspectives/PEA3624-1.html> accessed 16 May 2026; See generally Su Jung Jee and So Young Sohn, 'A Firm's Creation of Proprietary Knowledge Linked to the Knowledge Spilled over from Its Research Publications: The Case of Artificial Intelligence' (2023) 32 Industrial and Corporate Change 876.

[129] On market availability as an objective criterion in assessing whether a technique reflects the state of the art, see Schmitz (n 57) 4; Eder and others (n 65) 10.

[130] AI Act, art 55(2); Code of Practice, Safety and Security Chapter (n 9) Measure 3.2;); Code of Practice, Safety and Security Chapter (n 9) Appendix 3 Model evaluations; See para 43.

[131] Code of Practice, Safety and Security Chapter (n 9) recital (a); Schneider (n 20) para 8: '"Die Formulierung insbesondere vor seinem ersten Inverkehrbringen" in Erwägungsgrund 114 wird so zu verstehen sein, dass die Anbieter während des gesamten Lebenszyklus eines Modells die Bewertungs- und Minderungspflicht erfüllen müssen.'

[132] Subhabrata Majumdar, Brian Pendleton and Abhishek Gupta, 'Red Teaming AI Red Teaming' (arXiv, 7 July 2025) <https://arxiv.org/abs/2507.05538v2> accessed 17 May 2026, 11.

[133] Anusha Sinha and others, 'From Firewalls to Frontiers: AI Red-Teaming Is a Domain-Specific Evolution of Cyber Red-Teaming' (arXiv, 14 September 2025) <https://arxiv.org/abs/2509.11398v1> accessed 17 May 2026, 7 ['AI red-teaming places little emphasis on practical threat modeling. AI Red Teams often fail to engage rigorously in practical threat modeling (c.f. Figure 1), with common failures including focusing solely on the AI model, ignoring easier paths to the same end, or speculating on threat models without grounding them in real-world threat intelligence. Recent AI security research on jailbreaks (forcing generative AI models to generate content against their safety policies) has been criticized for failing to consider alternative ways the content could be produced or found (e.g., via a web search [59]), leaving threat models implicit, and prioritizing marginal gains in attack success rate rather than downstream impact [88, 73]. In generative AI more broadly, there has been some criticism of the focus on future AI risks, rather than current concerns [26, 9].'].

[134] See on how providers use threat modelling to map how they expect their models to be misused by threat actors to cause severe harm, Aaditya Singh and others, 'OpenAI GPT-5 System Card' (arXiv, 19 December 2025) <https://arxiv.org/abs/2601.03267v2> accessed 17 May 2026, 46 ['Informed by our threat modeling efforts, we created a taxonomy of content related to biological threats, for use both in training models to be safe, and in building system-level safeguards [...]'].

[135] Code of Practice, Safety and Security Chapter (n 9) Commitment 5, Measure 5.1 'Signatories will implement safety mitigations that are [...] sufficiently robust under adversarial pressure'; app 1.3.3(5) lists 'vulnerability to adversarial removal of guardrails' as a possible source of systemic risk.

[136] Code of Practice, Safety and Security Chapter (n 9) Commitment 4, Measure 4.2(3).

40. In addition to the state-of-the-art model evaluations, the Code of Practice requires that providers conduct lighter-touch evaluations at appropriate trigger points along the entire model's lifecycle.[137] These lighter-touch evaluations do not have to comply with the requirements set out in its Appendix 3, which provides guidance on the methodology of model evaluations,[138] but must still be appropriate to the purpose of assessing and mitigating systemic risks. For example, providers may conduct automated evaluations, which are tests run programmatically without the involvement of real users.[139] In line with the obligation for providers to take appropriate measures throughout the model's lifecycle and to cooperate with relevant actors along the AI value chain,[140] the Code of Practice thus requires continuous evaluation, albeit with varying levels of intensity.

2.1.1.2.1. State-of-the-art model evaluations

41. Measure 3.2 in the Safety and Security Chapter reiterates that model evaluations must be '*at least* state-of-the-art' and tailored to the modalities relevant to the systemic risk. The latter condition is not explicitly mentioned in the text of Article 55(1)(a) but aims to ensure that the evaluation meaningfully tests the conditions under which the risk could materialise.[141] The Code of Practice's Glossary in turn defines state of the art as 'the forefront of relevant research, governance, and technology that goes beyond best practice.'[142] Whether this definition should also be understood as reflecting the meaning of the state-of-the-art condition in Article 55(1)(a), or whether alternative interpretations remain possible, is discussed above.[143]

42. Providers must design and conduct model evaluations that are not only state of the art but also appropriate to the model and the systemic risk concerned.[144] The selection and design of suitable evaluation techniques should be informed by the model-independent information gathered pursuant to Measure 3.1, including insights into evaluation practices adopted by other providers and the broader research community. In practice, this may require the development of risk-specific evaluation techniques tailored to particular systemic risk categories.[145] Model evaluations must be designed to capture all identified systemic risks and, at a minimum, the specified systemic risks listed in Appendix 1.4.[146] For certain systemic risk categories – particularly chemical, biological, radiological, and nuclear ("CBRN") and cyber-offence risks – a relatively developed body of evaluation methods already exists.[147] Other risks or capabilities, such as harmful manipulation or the

[137] Code of Practice, Safety and Security Chapter (n 9) Measure 1.2.

[138] ibid. See the next section for a more elaborate discussion of this appendix.

[139] Mikaela Grace and others, 'Demystifying Evals for AI Agents' (Anthropic 2026) <https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents> accessed 17 May 2026.

[140] AI Act, recital 114.

[141] Code of Practice, Safety and Security Chapter (n 9) app 3.2;

[142] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'state of the art'.

[143] See Section 2.1.1.1.

[144] Code of Practice, Safety and Security Chapter (n 9) Measure 3.2.

[145] Joaquin Vanschoren, 'The Role of AI Safety Benchmarks in Evaluating Systemic Risks in General-Purpose AI Models' (European Commission Joint Research Centre 2025) JRC143259 <https://doi.org/10.2760/1807342> accessed 17 May 2026, 8 ['For example, a general reasoning benchmark might not reveal a model's ability to synthesise dangerous biological information, but a targeted safety benchmark specifically designed for CBRN risks would.']

[146] Code of Practice, Safety and Security Chapter (n 9) Commitment 3 and app 1.4.

[147] See, for example Frontier Model Forum, 'Frontier AI Biosafety Thresholds' (Frontier Model Forum 2025) Issue Brief <https://www.frontiermodelforum.org/issue-briefs/frontier-ai-biosafety-thresholds/> accessed 17 May 2026; CBRN and cyber-offence capabilities are tracked in OpenAI, 'Preparedness Framework Version 2' (OpenAI 2025) <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf> accessed 17 May 2026, 5; Anthropic, 'Anthropic's Responsible Scaling Policy Version 3.0' (Anthropic) <https://www-

capability to operate autonomously, remain comparatively under-evaluated.[148] This may require providers to design model evaluations from scratch, which in turn requires the AI Office to assess not only the results of those evaluations but also whether the design of the evaluation itself is appropriate for assessing and mitigating the relevant identified systemic risk.

43. Appendix 3 of the Safety and Security Chapter provides further guidance on the methodology of model evaluations.[149] First, model evaluations must meet the quality standard of having a *high degree of scientific and technical rigour*,[150] which in turn means that evaluations must have internal and external validity as well as be reproducible.[151] *Internal validity* refers to the extent to which an evaluation ensures that the results of the evaluation 'are as accurate as scientifically possible in the evaluation setting and are free from methodological shortcomings that could undermine the results.'[152] The AI Office may be able to verify the internal validity of model evaluations on the basis of the information provided in the Safety and Security Model Report.[153] *External validity* concerns the extent to which model evaluations are 'suitably calibrated for results to be used as a proxy for model behaviour outside the evaluation environment.'[154] Developments such as changes in the deployment context of the AI model,[155] the emergence of new misuse techniques, or advances in evaluation methodologies may undermine the external validity of previously conducted model evaluations and may therefore prompt providers to update their Safety and Security Model Reports accordingly.[156] Finally, *reproducibility* requires providers to document the data, techniques, evaluation conditions, and other elements of the evaluation methodology in a manner that allows third parties, such as researchers and engineers, to validate, reproduce, or improve upon the results of the model evaluation.[157]

44. In addition to the model evaluations being rigorous, providers must also ensure that model evaluations are conducted with at least a state-of-the-art level of model elicitation,[158] which refers to

---

cdn.anthropic.com/e670587677525f28df69b59e5fb4c22cc5461a17.pdf> accessed 17 May 2026, 6; Google DeepMind, 'Frontier Safety Framework Version 2.0' (Google DeepMind 2025) <https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0%20(1).pdf>, 2; Meta, 'Advanced AI Scaling Framework Version 2' (Meta) <https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2> accessed 17 May 2026, s 3.

[148] Matteo Prandi and others, 'Bench-2-CoP: Can We Trust Benchmarking for EU AI Compliance?' (arXiv, 7 August 2025) <https://arxiv.org/abs/2508.05464v2> accessed 17 May 2026, 9.

[149] Code of Practice, Safety and Security Chapter (n 9) app 3.

[150] Code of Practice, Safety and Security Chapter (n 9) app 3.1.

[151] Code of Practice, Safety and Security Chapter (n 9) app 3.2; Glossary, definition of 'high scientific and technical rigour'.

[152] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'internal validity'.

[153] For example, providers must include in their Safety and Security Model Reports 'at least five random samples of inputs and outputs from each relevant model evaluation,' and, where requested by the AI Office, a sufficiently large number of additional random samples of inputs and outputs from the relevant evaluation; see Code of Practice, Safety and Security Chapter (n 9) Commitment 7, Measure 7.3(1)(f).

[154] Code of Practice, Safety and Security Chapter (n 9) app 3.2; Glossary, definition of 'external validity'.

[155] AI Security Institute, 'International Consensus and Open Questions in AI Evaluations' (*AI Security Institute*, 12 February 2026) <https://www.aisi.gov.uk/blog/international-ai-network-consensus-and-open-questions> accessed 17 May 2026 ['For external validity, evaluators should design evaluation protocols with external context in mind including developing realistic scaffolding that would be used in real-world applications or cost-performance trade-off parameters mirroring real-world usage.'].

[156] Code of Practice, Safety and Security Chapter (n 9) Commitment 7, Measure 7.6(5).

[157] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'reproducibility'; See also AI Security Institute (n 155).

[158] Code of Practice, Safety and Security Chapter (n 9) app 3.2.

'technical work to systematically enhance a model's capabilities, propensities, affordances, and/or effects, thereby facilitating an accurate measurement of the full range of its capabilities, propensities, affordances, and/or effects that can likely be attained'.[159] In particular, providers are expected to employ techniques that minimise the risks of under-elicitation – that is, situations in which the evaluation setup fails to reveal relevant capabilities – as well as model deception during evaluation, such as through sandbagging.[160] Under-elicitation risks not being able to capture 'an accurate measurement of the full range of its capabilities, propensities, affordances, and/or effects that can likely be attained.'[161]

45. This requirement is particularly relevant in the context of adversarial testing, where red-teaming exercises must be sufficiently probing to reveal capabilities that might otherwise remain hidden and therefore escape meaningful risk identification.[162] Determining the appropriate level of model elicitation requires providers to take into account what is reasonably foreseeable in terms of potential misuse scenarios, the capabilities of likely misuse actors, and the expected deployment context of the model.[163]

2.1.1.2.2. Adversarial testing of the model

46. Providers are expected to conduct *adversarial testing* of the model as part of the set of state-of-the-art model evaluations required under Article 55(1)(a). Adversarial testing refers to techniques of deliberately trying to subvert a model's built-in defences by simulating hostile or manipulative interactions in which the tester assumes the role of an adversary[164] in order to assess whether the model can be induced to produce harmful or otherwise unacceptable outputs.[165] During this exercise, the adversarial tester can 'identify dangerous capabilities, vulnerabilities, or other emergent properties' of the model that might not be apparent under standard evaluation conditions.[166]

47. The AI Act identifies *red-teaming* as one form of adversarial testing that providers must conduct within the suite of state-of-the-art model evaluations for assessing systemic risk.[167] Red-teaming is an

[159] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'model elicitation'.
[160] Code of Practice, Safety and Security Chapter (n 9) app 3.2(2).
[161] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'model elicitation'.
[162] See Section 2.1.1.2.2.
[163] Code of Practice, Safety and Security Chapter (n 9) app 3.2 second paragraph, (1)–(2).
[164] Jessica Ji, 'How to Improve AI Red-Teaming: Challenges and Recommendations' (*Center for Security and Emerging Technology*, 21 March 2025) <https://cset.georgetown.edu/article/how-to-improve-ai-red-teaming-challenges-and-recommendations/> accessed 17 May 2026; 'Adversarial Testing' (*Holistic AI*) <https://www.holisticai.com/glossary/adversarial-testing> accessed 19 February 2026.
[165] Ji (n 164); Anusha Sinha and others, 'What Can Generative AI Red-Teaming Learn from Cyber Red-Teaming?' (Carnegie Mellon University 2025) CMU/SEI-2025-TR-006 <https://doi.org/10.1184/R1/29410136> accessed 17 May 2026, 4; John Halstead, 'Managing Risks from AI-Enabled Biological Tools' (*GovAI*, 5 August 2024) <https://www.governance.ai/analysis/managing-risks-from-ai-enabled-biological-tools> accessed 17 May 2026.
[166] Barrett and others (n 44) 48; On adversarial testing not just as a method for evaluating capabilities but also as a 'measure of human interaction: specifically of the friction a person encounters when trying to use an AI system to malicious ends', see Laura Weidinger and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (arXiv, 18 October 2023) <https://arxiv.org/abs/2310.11986v2> accessed 17 May 2026, 8.
[167] On mentions of red-teaming as a form of adversarial testing in the AI Act, see Annex XI Section 2 ['Where applicable, a detailed description of the measures put in place for the purpose of conducting internal and/or external adversarial testing (e.g. red teaming) [...]']; On the distinction between red-teaming and adversarial testing, Anthropic uses the terms 'red teaming;' and 'adversarial testing' seemingly synonymously ['"Red teaming," or adversarial testing, is a recognized technique to measure and increase the safety and security of systems.'], Anthropic, 'Frontier Threats Red Teaming for AI Safety' (*Anthropic*) <https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety> accessed 19 February 2026; Ji (n 164) ['Participants at CSET's AI testing workshop generally agreed that

expert-driven and scenario-focused exercise that evolved in the cybersecurity defence sector[168] where experts use various tools and techniques to emulate how an adversary would attempt to identify and exploit system vulnerabilities.[169] AI developers have increasingly adopted red-teaming techniques to identify risks and assess the robustness of safety and security mitigations in AI models and systems.[170] Unlike traditional cybersecurity red-teaming,[171] AI red-teaming practices go beyond identifying security flaws and instead probe how a model can be induced to generate 'harmful, unwanted, or policy-violating outputs'[172], with the aim of assessing and managing 'the safety, security, and trustworthiness of these models'[173].

48. Red-teaming raises a structural question concerning the delineation between evaluating a model and evaluating a system.[174] The AI Act draws a conceptual distinction between (general-purpose) AI models and AI systems built on top of them.[175] However, certain systemic risks may only become apparent once a model is integrated into downstream systems, particularly where such systems provide access to tools or other forms of operational scaffolding.[176] The Code of Practice therefore requires providers, when designing and conducting red-teaming exercises, to account within the limits of reasonable foreseeability for both the elicitation capabilities of potential misuse actors and

---

AI red-teaming involves adversarial testing methods. [fn 1:'There is disagreement on whether or not this is the case across the board: some red-teamers and researchers argue that red-teaming doesn't always require adversarial methods because it has come to encompass both security- and safety-focused testing practices which don't necessarily involve emulating an adversary.]', referencing AI Risk and Vulnerability Alliance (ARVA) and others, 'Red-Teaming in the Public Interest' (Data & Society Research Institute 2025) <https://doi.org/10.69985/VVGP4368> accessed 17 May 2026.); The Frontier Model Forum describes '[a]dversarial testing [...] as one approach to "red teaming" where the aim is to discover harmful content or vulnerabilities in the model through a combination of automated or manual probing techniques', Frontier Model Forum, 'What Is Red Teaming?' (Frontier Model Forum 2023) <https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf> accessed 17 May 2026, 3.

[168] Micah Zenko, *Red Team: How to Succeed By Thinking Like the Enemy* (Basic Books 2015) [Red-teaming is fundamentally a team exercise].

[169] Sven Cattell, Avijit Ghosh and Lucie-Aimée Kaffee, 'Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities' *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2024) <https://doi.org/10.1609/aies.v7i1.31635> accessed 17 May 2026; Sinha and others, 'What Can Generative AI Red-Teaming Learn from Cyber Red-Teaming?' (n 165) 18.

[170] See Lama Ahmad and others, 'OpenAI's Approach to External Red Teaming for AI Models and Systems' (arXiv, 24 January 2025) <https://arxiv.org/abs/2503.16431> accessed 17 May 2026; Anthropic, 'Challenges in Red Teaming AI Systems' (*Anthropic*, 12 June 2024) <https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems> accessed 17 May 2026; Majumdar, Pendleton and Gupta (n 132).

[171] Majumdar, Pendleton and Gupta (n 132) 3 ['the goals of AI red teaming are broader than just ensuring secure and safe behavior of AI models, and its means are deeper than narrow technical approaches like pentesting or fuzzing.']; Shayne Longpre and others, 'A Safe Harbor for AI Evaluation and Red Teaming' (arXiv, 7 March 2024) <https://doi.org/10.48550/arXiv.2403.04893> accessed 17 May 2026, 2 ['[Red teaming] has been adopted by the AI community to instead describe penetration testing of a broader set of system flaws than traditional security (The Hacking Policy Council, 2023)'].

[172] Marie-Laure Hicks and others, 'Exploring Red Teaming to Identify New and Emerging Risks from AI Foundation Models: Summary Workshop Report' (RAND Europe 2023) <https://doi.org/10.7249/CFA3031-1> accessed 21 May 2026, 8; Longpre and others, A Safe Harbor for AI Evaluation and Red Teaming' (n 171) 3; 'Red Team - Glossary' (*NIST Glossary*) <https://csrc.nist.gov/glossary/term/red_team> accessed February 2026; Sinha and others, 'What Can Generative AI Red-Teaming Learn from Cyber Red-Teaming?' (n 165) 10.

[173] See Michael Feffer and others, 'Red-Teaming for Generative AI: Silver Bullet or Security Theater?' (arXiv, 27 August 2024) <https://doi.org/10.48550/arXiv.2401.15897> accessed 17 May 2026, 2.

[174] Sinha and others, 'From Firewalls to Frontiers' (n 133) 2–3. ['Attackers do not delineate between arbitrary distinctions of AI vs. non-AI [...]. Red-teaming only AI components or only traditional software components in these systems fails to properly emulate adversaries – the defining feature of red-teaming. [...] adversary emulation requires system-level thinking, encompassing software, AI, and the interaction between the two, and thus indicates a need for a combined approach to red-teaming.'].

[175] AI Act, arts 3(1) and 3(66), See the forthcoming chapter on System vs Model in this work.

[176] AI Act, recital 110.

the expected use context of the model.[177] This includes considering planned or contemplated integrations of the model into an AI system, as well as integrations currently observed for similar models where such uses are known to the provider and cannot reasonably be excluded for their own model.[178]

2.1.1.2.3. Internal and external model evaluations

49. Closely related questions also arise concerning the extent to which Article 55(1)(a) requires providers to conduct independent external adversarial testing or other types of external model evaluations.[179] External model evaluations are understood within the industry as encompassing evaluations conducted by, or with the involvement of, independent external actors.[180]

50. The AI Act distinguishes between internal and external model evaluations in Recital 114, which states that providers should conduct the necessary model evaluations, including, 'as appropriate, through internal or independent external testing.' Where applicable, providers must then document any measures put in place 'for the purpose of conducting internal and/or external adversarial testing'.[181] This distinction reflects meaningful technical and policy differences between the two types of model evaluations. External model evaluations can enhance trust in the results, while also carrying commercial, security, and operational disadvantages that internal model evaluations do not.[182]

51. External independent assessments are commonly mandated across EU product-safety and safety-critical legislation,[183] wherein such obligations are imposed explicitly as freestanding requirements, most notably through notified-body conformity-assessment procedures or explicit audit obligations.[184] The omission of a reference to internal and external testing in the text of Article 55(1)(a) thus leaves the scope of this obligation open to interpretative ambiguity.

---

[177] See the forthcoming commentary on Article 3(65), section on 'reasonable foreseeability'; See Section 2.1.2.1.2.

[178] Code of Practice, Safety and Security Chapter (n 9) app 3.2 second paragraph (2)(a)–(b).

[179] On the question of whether Article 55(1)(a) requires external model evaluations, commentators are divided. Nathalie A Smuha and Karen Yeung, 'The European Union's AI Act: Beyond Motherhood and Apple Pie?' in Nathalie A Smuha (ed), *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence* (Cambridge University Press 2025) <https://doi.org/10.1017/9781009367783.015> accessed 17 May 2026, 243, noting that 'that providers of (systemic risk) GPAI models can conduct their own audits and evaluations, rather than rely on external independent third party audits.'; Theodoros Karathanasis, 'The Regulatory Interplay Between the AI Act and the DSA: Challenges, Burden, and Rationalization for AI Innovation in the EU' (SSRN, 1 July 2025) <https://doi.org/10.2139/ssrn.5332512> accessed 17 May 2026, 14: 'There exists an obligation to obtain independent external systemic risk assessments (including model evaluations) prior to market placement for GPAI models with systemic risks under certain conditions.'; Schneider (n 20) para 8, stating that an obligation to conduct external model evaluations is inferred from recital 114, 'Die Bewertungen sollen bereits vor dem ersten Inverkehrbringen eines Modells erfolgen und können auch im Rahmen interner oder unabhängiger externer Tests durchgeführt werden (vgl. Erwägungsgrund 114).'

[180] See Frontier Model Forum, 'Third-Party Assessments' (Frontier Model Forum 2025) Technical Report <https://www.frontiermodelforum.org/technical-reports/third-party-assessments/> accessed 17 May 2026, 4.

[181] AI Act, annex XI, s 2(2).

[182] Jacob Charnock and others, 'Expanding External Access To Frontier AI Models For Dangerous Capability Evaluations' (arXiv, 17 January 2026) <https://doi.org/10.48550/arXiv.2601.11916> accessed 17 May 2026, 5; Benjamin S Bucknall and Robert F Trager, 'Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements' (AI Governance Initiative 2023) Whitepaper <https://cdn.governance.ai/Structured_Access_for_Third-Party_Research.pdf> accessed 17 May 2026, 18.

[183] Blue Guide (n 61) s 5.1.3. on actors in conformity assessment.

[184] Alejandro Tlaie and Jimmy Farrell, 'Securing External Deeper-than-Black-Box GPAI Evaluations' (arXiv, 13 March 2025) <https://doi.org/10.48550/arXiv.2503.07496> accessed 17 May 2026, 5; Lisa Soder and Amin

52. One interpretive pathway is to read the obligation to conduct state-of-the-art model evaluations as, in certain circumstances, implicitly requiring both internal and external model evaluations. While Article 55(1)(a) itself does not expressly refer to external evaluations, Recital 114 clarifies that providers of GPAI models with systemic risk are to conduct model evaluations, including, 'as appropriate, through internal or independent external testing', as mentioned above. Read together with the *state-of-the-art* requirement, the recital suggests that whether external evaluations are required depends on whether internal evaluations alone are sufficient to meet the level of rigour necessary for adequately assessing and mitigating systemic risk. Put differently, where certain systemic risks can only be assessed through independent external evaluation, reliance exclusively on internal evaluations may fall short of the state-of-the-art standard required under Article 55(1)(a). At the same time, the '*as appropriate*' language in Recital 114 signals that the requirement to conduct external model evaluations may be conditional rather than automatic.

53. The Safety and Security Chapter of the GPAI Code of Practice, as the European Commission's endorsed operationalisation of Article 55(1),[185] lends support to this reading. According to Measure 3.2 and the accompanying Appendix 3.5, providers are expected to conduct independent external model evaluations in addition to internal model evaluations prior to market placement, unless the model can be demonstrated to be 'similarly safe or safer' than existing models or where signatories are unable to 'appoint adequately qualified independent external evaluators'.[186] By making external model evaluations conditional in this way, the Code of Practice appears to reflect an attempt to balance two competing considerations: on the one hand, a purposive reading of Article 55 directed at ensuring appropriate assessment and mitigation of systemic risk, and on the other, the need to maintain proportionality in the compliance measures imposed on providers.[187]

54. A second interpretative pathway is that the omission of external model evaluations from the text of Article 55(1)(a) is deliberate, with the effect that such evaluations fall outside the scope of the provision. As noted above, the only reference to external model evaluations appears in Recital 114, which may provide interpretative guidance for the operative provision but cannot introduce additional substantive obligations into the text of that provision.[188] This is particularly relevant if conducting external model evaluations is understood to be a materially different obligation from conducting internal model evaluations. The second and only other reference to external model evaluations in the AI Act appears in Annex XI, which is an integral and binding part of the AI Act. However, its reference to external model evaluations is framed as part of the technical

---

Oueslati, 'Trust Is Good, Assurance Is Better' (*Interface*, 20 March 2025) <https://www.interface-eu.org/publications/trust-is-good-assurance-is-better> accessed 17 May 2026.

[185] See Section 2.1. above.

[186] Code of Practice, Safety and Security Chapter (n 9) app 3.5, para 1.

[187] Code of Practice, Safety and Security Chapter (n 9) recital (i) on (purposive) interpretation and recital (c) on the principle of proportionality; The *de facto* stringency of these conditions is contested. Some argue that companies can sidestep independent scrutiny simply by claiming the expertise to determine that their model is no riskier than existing 'similarly safe' systems, effectively transforming independent assessment into an optional formality (CeSIA, 'CeSIA's Feedback on the Final Draft of the EU Code of Practice for GPAI' (*CeSIA*, 18 November 2025) <https://cesia.org/en/publications/cesias-feedback-on-the-final-draft-of-the-eu-code-of-practice-for-gpai/> accessed 17 May 2026); Others contend that '[i]n an industry where every release has to be justified on the grounds of enhanced performance or some novel feature, the Code makes external evaluations de facto mandatory' (From Daron Acemoglu and others, 'Ensuring GPAI Rules Serve the Interests of European Businesses and Citizens' (25 June 2025) <https://thefuturesociety.org/wp-content/uploads/2025/06/ProtectingGPAIRules.pdf> accessed 17 May 2026).

[188] 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 2019) Research Paper 2019–31 <https://doi.org/10.2139/ssrn.3445372> accessed 17 May 2026, 3.

documentation to be provided 'where applicable'.[189] This wording suggests that external evaluations are contemplated by the Act and may be relevant in some circumstances, but does not necessarily support the conclusion that Article 55(1)(a) requires all providers of GPAI models with systemic risk to conduct external evaluations as a matter of course. On that reading, interpreting Article 55(1)(a) as implicitly requiring external model evaluations would extend beyond what is supported by the text of the operative provision itself.

55. The tenability of this line of interpretation, however, is not assured. Its persuasiveness depends in part on establishing that requiring external model evaluations would be a disproportionately burdensome reading of the obligation to perform model evaluations. This can be contested in light of the overall risk-based approach and safety objectives of the AI Act.[190] A purposive interpretation of Article 55(1)(a), directed at ensuring effective assessment and mitigation of systemic risk,[191] may support the conclusion that external model evaluations are necessary for appropriate risk management, given their 'unique benefits [such as] broader researcher participation, diversity of subject matter experts, novel approaches, independence, and greater evaluation speed'.[192] The more pertinent question may therefore not be whether external model evaluations fall within the provision's scope at all but rather under which conditions they are appropriate. In that respect, factors such as the absence of prior independent external evaluations, limitations in the provider's internal evaluation capacity, or the inability of internal evaluations alone to demonstrate sufficient robustness or independence may support the conclusion that external evaluations are 'appropriate' within the meaning of Recital 114.

56. The GPAI Code of Practice provides guidance not only on when external model evaluations are appropriate, but also on how they should be conducted.[193] It sets out criteria for identifying adequately qualified independent external evaluators and specifies what constitutes adequate access to the model in the course of an external evaluation.[194] Significantly, the Safety and Security Chapter defines an independent external evaluator as any 'natural or legal person that has no financial, operational, or management dependence on the Signatory [...] and is otherwise free from the Signatory's control in reaching conclusions and/or making recommendations, including through contractual safeguards and suitable conflict of interest policies.'[195]

2.1.1.3. Documentation obligations under Article 55(1)(a) and (b)

57. The text of Article 55(1)(a) explicitly refers to *documentation* only in a specific context, namely by requiring providers to perform and document adversarial testing. This wording gives rise to the next interpretative question on whether documenting the state-of-the-art model evaluations is also a constitutive element of this provision. Legal scholarship appears to be divided between authors who

[189] AI Act, annex XI, s 2(2).
[190] AI Act, art 1(1).
[191] Code of Practice, Safety and Security Chapter (n 9) recital (i).
[192] Shayne Longpre and others, 'In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI' (arXiv, 25 March 2025) <https://doi.org/10.48550/arXiv.2503.16861> accessed 17 May 2026, 3.
[193] On the state of the art of external model evaluations, see Kevin Klyman and others, 'Safeguarding Third-Party AI Research' (Stanford University Human-Centered Artificial Intelligence 2025) Policy Brief <https://hai.stanford.edu/assets/files/hai-policy-brief-safeguarding-third-party-ai-research.pdf> accessed 17 May 2026, 4.
[194] Code of Practice, Safety and Security Chapter (n 9) app 3.5, para 2.
[195] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'independent external'.

consider documentation exclusively in relation to adversarial testing[196] and those who understand compliance with Article 55(1)(a) as requiring both the execution of the model evaluation and its documentation, where the latter includes the results, the tests performed and the criteria applied.[197]

58. Treating documentation as a distinct obligation built into Article 55(1)(a) can be substantiated on the basis of a teleological and systematic interpretation of the provision. From a teleological perspective, an obligation limited to conducting model evaluations without a corresponding requirement to document them would render the duty to identify and mitigate systemic risk having regard to the state of the art largely unenforceable. In the absence of documentation, the AI Office would be unable to verify either the methodologies relied upon or their adequacy in light of the state-of-the-art requirement. A teleological reading of Article 55(1)(a) thus entails that providers must not only conduct model evaluations but also document them in a manner that enables effective supervisory scrutiny. Interpretation to the contrary would undermine the regulatory purpose of Article 55(1)(a), which is to subject models posing increased systemic risks to enhanced oversight in order to safeguard public goods.[198]

59. A systematic reading of Article 55(1) in conjunction with Annex XI Section 2 speaks to the same effect.[199] Notably, the absence of an explicit cross-reference between Article 55(1) and Annex XI, which in itself is a departure from standard guidance on the drafting of EU regulations,[200] cannot be relied upon to negate the normative relationship between the two, given that annexes form an integral part of EU legislative acts and have the same binding legal status as the operative

---

[196] See in this sense by discussing 'documentation' mainly in the context of adversarial testing: Bernsteiner and Schmitt, 'Art 55' (n 24) para 9: 'Documentation of the details (in particular, the method of execution and follow-up measures) of the respective attack tests is a mandatory component of the technical documentation' (translated from German).

[197] Beurskens (n 20) para 5; Schneder (n 20) para 7 characterises the obligations under Article 55(1)(a) as *'Durchführung und Dokumentation von Modellbewertungen*' ['implementation and documentation of model evaluations']. This wording indicates that the commentators interpret article 55(1)(a) as encompassing not only an obligation to carry out state-of-the-art model evaluations, but also an inherent obligation to document those evaluations as part of the provision's content.

[198] On the underlying logic of the provision as additional obligations due to increased risk, see Bernsteiner and Schmitt, 'Art 55' (n 24) para 2; on safeguarding public goods as an objective of the AI Act (see AI Act, art 1); see Davor Petrić, 'The Court of Justice of the E.U.: A Contextualist Court' (2023) 8 University of Bologna Law Review 11, 27; Bohumila Salachová and Bohumil Vítek, 'Interpretation of European Law, Selected Issues' (2013) 61 Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis 2717, 2719; See also Hannes Rösler, 'Interpretation of EU Law' in Jürgen Basedow, Klaus Hopt and Reinhard Zimmermann (eds), *Max Planck Encyclopedia of European Private Law* (2012) <https://max-eup2012.mpipriv.de/index.php/Interpretation_of_EU_Law> accessed 17 May 2026; Nial Fennelly, 'Legal Interpretation at the European Court of Justice' (1996) 20 Fordham International Law Journal 656, 666 ['The context of a legal text is part of the background to its adoption.'](The Court of Justice of the European Union has consistently relied on teleological arguments to ensure that the meaning of a legal provision corresponds to the purpose of said provision, to that of the legal act within which the provision is situated, or the EU primary law as a whole.).

[199] See Petrić (n 198) 22, 'no legal provision is enacted in isolation from other provisions. Rather, every provision is a part of a certain section or chapter of some legislative act'. Also see the commentary on Article 53, Section 2.1.1.2.1. in this work.

[200] European Commission (ed), *Joint Practical Guide of the European Parliament, the Council and the Commission for Persons Involved in the Drafting of European Union Legislation* (2nd edn, Publications Office 2015) <https://doi.org/10.2880/89965> accessed 17 May 2026, 74: 'There must always be a clear reference in the appropriate part of the enacting terms to the link between those provisions and the annex (using phrases such as 'listed in the Annex" or "set out in Annex I").'.

provisions.[201] Indeed, while Article 53 expressly links the technical documentation obligation to Annex XI,[202] Article 55 does not contain an equivalent internal reference, and the title of Annex XI refers only to Article 53. However, the title of Section 2 of Annex XI leaves little room for ambiguity as to its function in supplementing and operationalising the obligations laid down in Article 55.

2.1.1.3.1. The scope of information to be compiled under Article 55(1)(a) and (b)

60. The scope of the documentation obligation under Article 55(1)(a) and (b) is specified through Annex XI Section 2, which sets out the categories of information that providers of GPAI models with systemic risk may be required to supply to the AI Office.[203] This includes technical documentation containing, *inter alia*, a detailed description of model evaluation strategies, evaluation results, and methodologies.[204] Annex XI also requires, where applicable, a description of the measures put in place for conducting internal or external adversarial testing, as well as a description of the model's system architecture.[205]

61. Providers who are also signatories to the GPAI Code of Practice are required to compile and keep up to date a Safety and Security Model Report containing detailed information on their systemic risk assessment and mitigation processes to be shared with the AI Office before placing a model on the market.[206] For providers that are signatories to the Code of Practice, it is conceivable that the Commission may rely primarily on the Safety and Security Model Report when requesting information,[207] rather than also requesting the provision of the information compiled under Annex XI, insofar as the information required by the Annex is already covered by the Safety and Security Model Report.[208] For providers that are not signatories to the Code of Practice, requests for information compiled under Annex XI Section 2 are likely to be informed by the structure and scope of information reflected in the Safety and Security Model Report, given its role as an operational benchmark for demonstrating compliance with Article 55.[209]

---

[201] Notably, the CJEU has previously relied on systemic interpretation to determine the meaning of a provision consistently with the provisions from an annex to the same legal act; See Case C-881/19, *Tesco Stores ČR a.s. v. Ministerstvo zemědělství* [2024] ECLI:EU:C:2022:15, paras 34–39; Petrić (n 198) 23.

[202] Also see the commentary on Article 53, Section 2.1.1. in this work.

[203] See the commentary on Article 53, Section 2.1.1.2.2. on the contents of annex XI, s 2.

[204] AI Act, annex XI, s 2(1).

[205] AI Act annex XI, s 2(2) and (3).

[206] Code of Practice, Safety and Security Chapter (n 9) Commitment 7, Measure 7.1.

[207] See the forthcoming commentary on Article 91 in this work.

[208] See Code of Practice, Safety and Security Chapter (n 9) recital (c) on the principle of proportionality to systemic risk ['The Signatories recognise that while systemic risk assessment and mitigation is iterative and continuous, they need not duplicate assessments that are still appropriate to the systemic risks stemming from the model; see also Code of Practice, Safety and Security Chapter (n 9) Commitment 7, ['If Signatories have already provided relevant information to the AI Office in other reports and/or notifications, they may reference those reports and/or notifications in their Model Report.].

[209] See Section 2.2. and the commentary on Article 56 in this work.

2.1.1.3.2. The extent of required documentation

62. Article 55(1)(a) says that state-of-the-art model evaluations must be conducted and documented[210] with a view of identifying and mitigating systemic risk.[211] One interpretative question that arises in this context relates to the extent and detail of documentation required.

63. The Safety and Security Chapter of the GPAI Code of Practice emphasises that the level of detail in documentation and reporting should be proportionate to the systemic risks.[212] Having an appropriate level of detail in the supplied documentation is therefore important for demonstrating the technical rigour of model evaluations.[213] Subpar documentation practices, or evidence that other providers have identified and implemented more rigorous approaches to documentation, could render a provider's own documentation practices inadequate for the purposes of demonstrating compliance.[214] This, in turn, would prompt providers to adjust their documentation practices accordingly. In practice, the documentation landscape remains fragmented, although there are emerging efforts to converge towards higher standards.[215] Moreover, there is arguably a spillover from the state-of-the-art requirement as extending to documentation practices. Given how fundamental documentation is to the risk assessment and mitigation process, this could contribute to a similar upward dynamic where providers are incentivised to invest greater technical effort and innovation into '[advancing] the state of the art in AI safety and security and related processes'.[216]

64. The documentation compiled by providers must, at a minimum, be sufficiently detailed to enable the AI Office to assess whether the provider has adequately assessed and mitigated systemic risks.[217] Documentation therefore performs an evidentiary function, in the sense that it must render the evaluation process, its methodology, and its results legible to the EU regulator.[218] At the same time, the breadth and depth of documentation are constrained by considerations of proportionality.[219] The AI Act does not require providers to document their processes in a manner that would impose disproportionate burdens, nor to disclose more information than is necessary for assessing compliance.[220] This creates a tension between, on the one hand, ensuring that the AI Office can meaningfully evaluate the adequacy of systemic risk assessment and mitigation, and, on the other, avoiding excessive documentation requirements.

---

[210] See para 57.

[211] AI Act, art 55(1)(a).

[212] Code of Practice, Safety and Security Chapter (n 9) recital (c).

[213] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'reproducibility'.

[214] Code of Practice, Safety and Security Chapter (n 9) Commitment 7, Measure 7.6(5).

[215] See, for example Tegan McCaslin and others, 'STREAM (ChemBio): A Standard for Transparently Reporting Evaluations in AI Model Reports' (arXiv, 3 September 2025) <https://doi.org/10.48550/arXiv.2508.09853> accessed 17 May 2026; Ruchira Dhar and others, 'EvalCards: A Framework for Standardized Evaluation Reporting' (SSRN, 15 September 2025) <https://doi.org/10.2139/ssrn.5444574> accessed 17 May 2026.

[216] Code of Practice, Safety and Security Chapter (n 9) recital (f).

[217] See AI Act, art 91 on the information necessary to assess compliance; Code of Practice, Safety and Security Chapter (n 9) on *adequate* reduction of risk.

[218] See McCaslin and others (n 215).

[219] See Mougan and others (n 82).

[220] See Section 2.2. for analysis on AI Act, art 78(2).

2.1.2. Article 55(1)(b): Assessment and mitigation of possible systemic risks at Union level

65. Article 55(1)(b) requires providers to 'assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk.'[221] The following analysis will tackle several key interpretative questions pertinent to the interpretation of this obligation. It begins by examining the type of *possible* risks that are subject to the risk assessment and mitigation process. It then outlines the broader risk assessment and mitigation framework that spans Article 55(1)(a), (b) and (d),[222] including how providers can demonstrate compliance along each step of this iterative process. In carrying out these obligations, providers are required to adopt and implement appropriate measures to ensure that systemic risks are reduced to an acceptable level. The meaning of *appropriate* and *acceptable* is examined in Section 2.1.2.3. Finally, the analysis considers at which stages of the AI model lifecycle the AI Act requires risk assessment and mitigation measures.

2.1.2.1. 'Possible' systemic risks at Union level

66. Article 55(1) AI Act limits risk assessment and mitigation to *possible* systemic risks that have an effect at the Union level and that 'may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk'.[223] The explicit reference to systemic risks having an effect at the Union level is somewhat curious, given that having 'a significant impact on the Union market' already constitutes an essential characteristic of the definition of *systemic risk*.[224] Indeed, for risks to qualify as systemic under Article 3(65), and for providers to qualify as providers of models with systemic risk under Article 51, the systemic risk they present must already have a significant impact on the Union market so as to justify EU intervention on the basis of Article 114 TFEU.[225]

67. The possible systemic risks to be identified are those that 'may stem from the development, the placing on the market, or the use' of the GPAI model.[226] The term 'may' signals that, while systemic risks can emerge at various stages – development, placing on the market, and use – this formulation does not necessarily exclude the possibility that such risks arise elsewhere along the AI model's lifecycle,[227] which providers are required to monitor as part of their ongoing risk management obligations.[228] In practice, this has been described as imposing a potentially 'boundless' obligation, given that, by their very nature as being general-purpose, GPAI models may be used across a wide

[221] AI Act, art 55(1)(b).

[222] This is also pertinent to article 55(1)(c) to the extent that risk mitigation includes security mitigations.

[223] Systemic risk by definition will have an effect on the Union market – it is one of its defining characteristics; therefore, this qualifier will be dismissed as a tautology and not analysed further.

[224] AI Act, art 3(65); Code of Practice, Safety and Security Chapter (n 9) app 1.2.1. Also see the forthcoming commentary on Article 3(65) in this work.

[225] Philipp Hacker, Atoosa Kasirzadeh and Lilian Edwards, 'AI, Digital Platforms, and the New Systemic Risk' (arXiv, 22 September 2025) <https://doi.org/10.48550/arXiv.2509.17878> accessed 17 May 2026, 26.

[226] Schneider (n 20) para 9: 'Die nach Art. 55 Abs. 1 lit. a identifizierten systemischen Risiken sind nach lit. b zu bewerten und zu mindern. Dies betrifft nur systemische Risiken, die sich auf der Unionsebene auswirken und aus der Entwicklung, dem Inverkehrbringen oder der Verwendung des KI-Modells mit allgemeinem Verwendungszweck ergeben können.'

[227] See forthcoming chapter on Modifications, Section 2.2.1. in this work on the notion of 'lifecycle'.

[228] AI Act, recital 114 ['providers of general-purpose AI models with systemic risks should continuously assess and mitigate systemic risks, [...] taking appropriate measures along the entire model's lifecycle and cooperating with relevant actors along the AI value chain.'].

range of contexts and must therefore be monitored in a very broad set of scenarios, thereby potentially challenging legal certainty.[229] Section 2.1.2.4. below examines how the obligations of assessment and mitigation span the entire model lifecycle, from development through to downstream use.

68. The question of what amounts to a *possible* systemic risk is closely linked to the level of effort that providers are expected to invest in anticipating risks at the identification stage. Notably, the AI Act itself does not systematically employ the term *possible systemic risks* in other contexts, which leaves its precise scope open to interpretation. The GPAI Code of Practice differentiates between *specified systemic risks*, which are those listed in Appendix 1.4, and all other *potential systemic risks* that 'could stem from the model and be systemic'.[230] The notion of *possible systemic risks* most plausibly encompasses both the specified risks identified in the Code of Practice and the broader category of potential systemic risks envisaged therein.

69. A useful point of comparison is the risk management framework applicable to high-risk AI systems under Article 9, which requires providers to identify 'known and reasonably foreseeable risks' that may stem from the intended use or reasonably foreseeable misuse of the system.[231] This formulation anchors the obligation in a standard of foreseeability, even if the AI Act does not further specify how that standard is to be operationalised.[232] By contrast, Article 55 refers to 'possible' systemic risks, raising the question of whether this notion extends beyond reasonably foreseeable risks.

70. Possible systemic risks should, at a minimum, be understood as encompassing both known and reasonably foreseeable risks. Known risks are those that refer to 'harm [that] has occurred in the past or is certain to occur in the future.'[233] Risks can become *known* not just by virtue of the subjective knowledge of a specific provider, but also if these risks have been subject to significant media attention or entered into a recognised incident database and thus can be assumed to be known, even if the provider does not use the database in question or has overlooked the specific entry.[234]

71. The key interpretative question is whether the notion of *possible* systemic risks extends beyond the standard of reasonable foreseeability. While the use of the term *possible* could signal a deliberate departure from the wording of Article 9 and be read as encompassing more remote or even speculative risks, such an interpretation may be difficult to reconcile with the principles of proportionality and legal certainty.[235] Requiring providers to account for purely hypothetical or

---

[229] Christoph Krönke, 'Das europäische KI-Gesetz: Eine Verordnung mit Licht und Schatten' (2024) Neue Zeitschrift für Verwaltungsrecht 529, para 534 ['Die Verpflichtung für Anbieter von GPAI-Modellen zur Bewertung und Minderung diffuser „Fernrisiken" reicht weit in die Regulierung der Verwendung der Systeme hinein und gibt den Anbietern nahezu unkalkulierbare Verpflichtungen auf. Es ist höchst zweifelhaft, ob diese Anforderungen dem auch unionsrechtlich verbindlichen allgemeinen Grundsatz der Rechtssicherheit (Art. 6 III EUV) gerecht werden.'].

[230] Code of Practice, Safety and Security Chapter (n 9) 11, figure 4.

[231] AI Act, art 9(2)(a).

[232] Nadja Braun Binder and Catherine Egli 'Art. 9 Risikomanagementsystem' in Mario Martini and Christiane Wendehorst (eds), *KI-VO: Verordnung über Künstliche Intelligenz* (2nd edn, C.H. Beck, 2026) para 22.

[233] AI Act, art 9(2). Binder and Egli (n 232) para 20 ['The risk management process pursuant to Article 9 must comprise four steps: the identification and analysis of known and reasonably foreseeable risks (point a)']; see also Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1 ("DSA"), art 34(1)(b).

[234] Simon Gerdemann, 'Artikel 9 Risikomanagementsystem' in Jens Schefzig and Robert Killan (eds), *Beck'scher Online-Kommentar KI-Recht* (C.H. Beck, 5th edn, 2025) para 38 (the analysis on *known risks* is in context of article 9 on high risk AI systems); Schuett (n 44) 376.

[235] Krönke (n 229) 534.

unforeseeable risks would risk rendering the obligation effectively unbounded.[236] At the same time, given the potentially more severe and wide-ranging impacts of general-purpose AI models with systemic risk, providers may be expected to adopt a more rigorous and forward-looking approach to identifying and analysing risks. Indeed, where the potential harm is severe or even catastrophic, the gravity of the consequences may outweigh a low probability of occurrence.[237] Accordingly, while the standard of reasonable foreseeability remains the appropriate baseline, its application in this context is likely to require a more precautionary assessment than under Article 9.[238]

2.1.2.1.1. Possible systemic risk corresponds to reasonably foreseeable risk

72. There are several arguments that support reading *possible* systemic risks to correspond to *reasonably foreseeable* risks. First, the challenge of setting boundaries for foreseeability is fundamentally one of legal certainty: providers must be able to determine when 'they are allowed to stop looking for new risks'.[239] Framing the notion of *possible* systemic risk in terms of a more functionally useful legal concept of *reasonable foreseeability* could help to address this concern.[240] Second, the principle of proportionality, both as a guiding principle for the interpretation of risk management obligations, and as a motivation behind the EU legislature's drafting choices, supports the view that the identification of possible systemic risks must remain constrained by a standard of reasonable foreseeability.[241] The European Parliament and Council amended the Commission's version of Article 9 to include the notion of *reasonableness* 'to keep the burden of risk management in proportion to risk.'[242] A third argument in favour of this interpretation draws on considerations of systematic consistency across the AI Act. The risk management process set out in Article 9 and the risk assessment and mitigation process in Article 55(1) share the same definition of *risk* in Article 3(2), are directed at protecting the same public interests – including public health, safety, and fundamental rights – and draw on established risk management frameworks reflected in the ISO/IEC guidelines.[243] Indeed, historical versions of the negotiated AI Act proposal reveal that 'the Council has suggested extending Article 9 to "general purpose AI systems"',[244] at a time when the term

[236] ibid.

[237] Schuett (n 44) 'For example, it should be extremely difficult for a provider to credibly assure that a catastrophic risk was unforeseeable.'

[238] Fabian Teichmann, 'Risk, Reasonableness and Residual Harm under the EU AI Act: A Conceptual Framework for Proportional Ex-Ante Controls' [2026] European Journal of Risk Regulation 1, 7 ['By tying obligations to what is "reasonably foreseeable," the Act aligns with the concept of fault in tort law, where the foreseeability of harm is a key factor in determining negligence. However, the AI Act's regime is *ex ante* and does not wait for harm to occur and be litigated; it proactively requires the producer to think like a "reasonable risk manager." This aspect could be seen as importing a negligence standard into regulatory compliance: failing to address a foreseeable misuse risk could render the AI system non-compliant (and possibly defective under product liability rules).'].

[239] Schuett (n 44) 376; Krönke (n 229) para 534.

[240] See on the established role of *reasonable foreseeability* in tort law, Roderick Bagshaw, 'What Is "Reasonable Foreseeability"?' in Kylie Burns and others (eds), *Torts on Three Continents: Honouring Jane Stapleton* (Oxford University Press 2024) <https://doi.org/10.1093/oso/9780198889748.003.0008> accessed 18 May 2026.

[241] Gerdemann , 'Art 9' (n 234) para 40 ['If further risk identification measures are not expected to uncover additional risks, or if any identifiable risks are unlikely to be relevant to the probability or severity of damage, the provider is not required to undertake these measures.'] (translated from German).

[242] Fraser and Bello y Villarino (n 106) 437.

[243] Schuett (n 44).

[244] Council of the European Union, 'Draft Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) - General Approach' (2022) 2021/0106(COD) ("Draft Regulation General Approach"); The Council defines the term 'general purpose AI system' as 'an AI system that – irrespective of how it is placed on the market or put into service, including as open source software – is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question

general-purpose AI model had not yet been introduced.[245] These shared structural features support a consistent interpretation of the amount of effort providers need to invest in identifying and analysing reasonably foreseeable risks.

73. Based on the above, the way in which reasonable foreseeability is interpreted and applied for the purposes of risk identification and analysis under Article 9(2)(a) can inform how providers of GPAI models with systemic risk identify and analyse *possible* systemic risks under Article 55(1). A risk is *foreseeable* if it 'has not yet occurred but can already be identified.'[246] *Reasonableness* has been interpreted as an objective standard that allows for the strengths of a provider, such as their 'specific knowledge [to] individually tighten the standard of care for foreseeability [while] subjective grounds [...] such as lack of knowledge, insufficient training, or lack of experience remain irrelevant.'[247] For providers of GPAI models with systemic risk, the standard of *reasonableness* would be calibrated in reference to factors including their expertise, the foreseeability of the damage, 'the availability and the costs of precautionary or alternative methods,' and 'the nature and value of the protected interest involved,'[248] which, in this case, includes risk to public health, safety, public security, fundamental rights, or the society as a whole.[249]

2.1.2.1.2. Reasonably foreseeable systemic risks

74. While the notion of *possible* may thus correspond to *reasonably foreseeable*, the risk management process under Article 55(1) may be read as requiring providers of GPAI models with systemic risk to undertake more extensive efforts in identifying and analysing risks than those imposed on providers of high-risk AI systems.

75. It is possible to justify this interpretation by reference to the nature and characteristics of systemic risk,[250] as well as the 'potential significantly negative effects' that warrant such models being subject to the relevant obligations under the AI Act,[251] even where exemptions would otherwise apply to GPAI models without systemic risk.[252] The GPAI Code of Practice confirms that *systemic risk*, as

answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems'.

[245] See Schuett (n 44) [During the drafting of Article 9, 'the Council has suggested extending Article 9 to "general purpose AI systems".'] At the time of drafting, the term *general-purpose AI model* was not yet being used and a *general-purpose AI system* was defined as a 'system that may be used in a plurality of contexts and be integrated in a plurality of other AI systems'. Draft Regulation General Approach (n 244).

[246] Schuett (n 44) 376.

[247] Binder and Egli, 'Art 9' (n 232) para 23 ['Dementsprechend wird auf den objektiven, verständigen Durchschnittsbeobachter abzustellen sein, wobei besondere subjektive Stärken wie spezifische Kenntnisse des Anbieters den Sorgfaltsmaßstab für die Vorhersehbarkeit individuell verschärfen können. Hingegen bleiben subjektive Entlastungsgründe wie Unkenntnis, zu geringe Ausbildung oder fehlende Erfahrung unberücksichtigt.].

[248] 'Principles of European Tort Law (PETL)' (*European Group on Tort Law*) <https://www.egtl.org/PETLEnglish.html> accessed 18 May 2026, art 4:102.

[249] AI Act, art 3(65); also see the forthcoming commentary on Article 3(65) in this work.

[250] Andrea Palumbo, 'Systemic Risk Management and the Constitutional Limits of Delegating Political Discretion: An Analysis of the DSA and the AI Act' [2025] European Journal of Risk Regulation 1, 3 ['These regimes create a specific risk category that revolves around the adjective "systemic," i.e., their rationale is that additional and more stringent provisions are warranted for providers of services and products that pose risks presenting a systemic character.'].

[251] AI Act, recital 97: 'Considering their potential significantly negative effects, the general-purpose AI models with systemic risk should always be subject to the relevant obligations under this Regulation.'

[252] AI Act, recital 104 [Exceptions as regards the transparency-related requirements imposed on GPAI models released under an open-source license do not apply to GPAI models with systemic risk that remain bound to obligations under this Regulation.].

defined in Article 3(65), should be interpreted in light of both the probability and severity of harm reflected in the definition of *risk* in Article 3(2),[253] while also taking account of additional characteristics that confer its systemic nature. These include compounding or cascading effects, high velocity, and the fact that a small number of actors or events could trigger the materialisation of the systemic risk, which may in turn be difficult or impossible to reverse.[254] Given the ways in which systemic risk may materialise, it is reasonable to argue that providers of GPAI models with systemic risk should be required to assess and mitigate not only *reasonably foreseeable risks*, but also a broader category of *possible* risks, subject to a more stringent standard of effort rather than the reasonableness threshold providers of high-risk AI systems are expected to adhere to. As has been noted elsewhere, 'the greater the potential impact of the risk, the more effort an organisation needs to put into foreseeing it. [...] it should be extremely difficult for a provider to credibly assure that a catastrophic risk was unforeseeable'.[255]

76. Evidence supporting this interpretation can be drawn from the relationship between reasonable foreseeability and the state-of-the-art requirement, both of which function as standards that structure the scope and intensity of providers' obligations. In the context of Article 9, commentators have argued that, in assessing both known and reasonably foreseeable risks, the objective standard of care must also take account of the *generally acknowledged state of the art*. This implies that neither purely theoretical nor entirely improbable risks are covered.[256] Rather, the effort required to identify and assess reasonably foreseeable risks must be calibrated to what is generally acknowledged as the state of the art, which in practice corresponds to established best practices.[257]

77. Article 55(1) similarly requires providers of GPAI models with systemic risk to employ methods and tools that reflect the *state of the art* for the purposes of assessing and mitigating systemic risks. As discussed in Section 2.1.1.1. above, the notion of state of the art mentioned in Article 55(1)(a) arguably sits at a higher threshold than the *generally acknowledged state of the art* used elsewhere in the AI Act as it requires measures that go beyond established best practices.[258] Therefore, where the assessment and mitigation of systemic risks must reflect this higher standard, limiting *possible* risks to *reasonably foreseeable* risks may not be sufficient if the state of the art demands a greater degree of effort from providers.

78. This reading appears to be endorsed by the Code of Practice's Safety and Security Chapter's Recital (g) on the precautionary principle.[259] The recital recognises that the lack and subpar quality of data surrounding systemic risk may impede the assessment of systemic risk and instead compels providers to extrapolate from current adoption rates and research and development trajectories of models when identifying systemic risks. This forward-looking approach to risk identification is also reflected in Measure 2.2 of the Code of Practice, which requires providers to develop *appropriate*

[253] Code of Practice, Safety and Security Chapter (n 9) recital (i).
[254] Code of Practice, Safety and Security Chapter (n 9) app 1.2.2, 35.
[255] Schuett (n 44) 376.
[256] Braun Binder and Egli, 'Art 9' (n 232) para 23; Gerald Spindler, 'Anforderungen an Hochrisiko-KI-Systeme (außer Transparenz)' in Eric Hilgendorf and David Roth-Isigkeit (eds) *Die neue Verordnung der EU zur Künstlichen Intelligenz* (2nd edn, C. HJ. Beck 2025) para 8.
[257] See Section 2.1.1.1.
[258] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'state of the art'.
[259] It reads: 'The Signatories recognise the important role of the Precautionary Principle, particularly for systemic risks for which the lack or quality of scientific data does not yet permit a complete assessment. Accordingly, the Signatories recognise that the extrapolation of current adoption rates and research and development trajectories of models should be taken into account for the identification of systemic risks.'

systemic risk scenarios as a basis for future risk modelling.[260] This exercise necessarily involves specifying ways in which systemic risks stemming from a model might materialise.[261] For the systemic risk scenarios to be *appropriate*, the providers will have to rely on 'best practices, the state of the art, or other more innovative processes, measures, methodologies, methods, or techniques that go beyond the state of the art.'[262] As a result, where providers deploy appropriate measures that go beyond the state of the art, they may become able to and also enable other providers to identify and analyse systemic risks that extend beyond what is merely reasonably foreseeable.[263] The *state-of-the-art* condition effectively lowers the threshold of foreseeability,[264] as advances in methods and techniques make it possible to capture risks that would previously have fallen outside the scope of reasonable foreseeability.

79. Further support for interpreting possible systemic risks as extending beyond reasonable foreseeability can be drawn from the structure of Article 9 and the limiting clauses embedded in that provision. Article 9 limits the risk management system to only those risks that can be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or through the provision of adequate technical information.[265] The absence of a comparable limiting clause in Article 55(1) suggests a broader range of *possible risks* subject to assessment and mitigation. To this point, the safety mitigations listed in the Code of Practice's Safety and Security Chapter are not confined to the design and development stage[266] but extend into deployment,[267] governance,[268] and building a safe ecosystem around the model.[269]

80. In addition, Article 9 limits identification and analysis of risks to those arising from the use of high-risk AI system in accordance with its intended purpose and under conditions of reasonably foreseeable misuse.[270] By contrast, GPAI models are characterised by their significant generality and by the fact that they may form the basis for a wide range of downstream systems, uses and applications.[271] Combined with the potentially greater severity of harm and the broader range of

[260] Risk modelling exercises are listed in Code of Practice, Safety and Security Chapter (n 9) Measure 3.3.

[261] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'systemic risk scenario'.

[262] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'appropriate'; See Section 2.1.1.1.

[263] In OpenAI, 'Preparedness Framework' (n 147) 4–7, OpenAI distinguishes between *Research Categories* and *Tracked Categories.* The Tracked Categories include the following capabilities: CBRN, cybersecurity, and AI self-improvement. The Research Categories include capabilities that could pose risks of severe harm but do not yet meet the five criteria required for designation in the Tracked Categories, which include being plausible, measurable, severe, net new, and instantaneous or irremediable. It remains unclear how the concept of 'plausible' harm maps onto the legal standard of (reasonable) foreseeability.

[264] Teichmann (n 238) 9 ['As new threats emerge or new solutions are invented, the acceptable "residual risk" threshold effectively tightens because more can be done to mitigate risk.']

[265] AI Act, art 9(3). See Section 2.1.2.2.2.

[266] Code of Practice, Safety and Security Chapter (n 9) Commitment 5.

[267] Code of Practice, Safety and Security Chapter (n 9) Commitment 5, Measure 5.1(4) on staging the access to the model.

[268] Code of Practice, Safety and Security Chapter (n 9) Commitment 8 on systemic risk responsibility allocation, Measure 8.1.

[269] Code of Practice, Safety and Security Chapter (n 9) Commitment 5, Measure 5.1(7) on enabling a safe ecosystem of AI agents.

[270] AI Act, art 9(2)(a) and (b).

[271] Recital 101 AI Act; Claire Boine and David Rolnick, 'Why The AI Act Fails to Understand Generative AI' (2025) 26 Minnesota Journal of Law, Science and Technology 61, 96–97; See Michael Veale and João Pedro Quintais, 'The Obligations of Providers of General-Purpose AI Models' in Gianclaudio Malgieri and others, *The EU Artificial Intelligence Act: A Thematic Commentary* (Hart Publishing 2026) 361, 361; see also, David Fernández-Llorca and others, 'An Interdisciplinary Account of the Terminological Choices by EU Policymakers

public interests at stake,[272] this structural difference suggests that the assessment and mitigation of *possible* systemic risks under Article 55 cannot be confined to what is *reasonably foreseeable* within the narrower risk management process under Article 9.

2.1.2.1.3. Sources of possible systemic risks

81. Systemic risk is defined as being specific to high-impact capabilities and as increasing 'with model capabilities and model reach, [and] can arise along the entire lifecycle of the model'.[273] The wording 'specific to' has been discussed as meaning that systemic risk stems from high-impact capabilities, or, in other words, that high-impact capabilities are the main source of systemic risk.[274] It has also been posited that, on an alternative interpretation, systemic risk is identified in models that have high-impact capabilities but does not stem exclusively from them.[275] Instead, systemic risk may arise in the most advanced GPAI models while also being shaped by additional factors, including 'conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails and other factors.'[276]

82. In identifying the sources of systemic risk, providers should therefore take account not only of factors internal to the model, such as its capabilities, but also of factors and conditions external to the model. This reading finds support in Recital 110 and is, more importantly, expressly reiterated in Appendix 1.3 of the Code of Practice's Safety and Security Chapter. The latter enumerates selected model capabilities,[277] model propensities,[278] model affordances,[279] and contextual factors as part of a non-exhaustive list of potential sources of systemic risk for the purposes of systemic risk identification.

2.1.2.2. Assessment and mitigation of possible systemic risks

83. Providers of GPAI models with systemic risk are required to assess and mitigate possible systemic risks.[280] The terms *assessment* and *mitigation* are not defined in the AI Act. To this end, providers may rely on the GPAI Code of Practice for guidance on how to implement the risk assessment and mitigation process. In particular, the Safety and Security Chapter describes *risk assessment* as encompassing the steps of systemic risk identification,[281] systemic risk analysis,[282] and systemic risk acceptance determination.[283] Systemic risk *mitigation* requires providers to implement both safety and security measures.[284] The full systemic risk assessment and mitigation process is continuous and

---

Ahead of the Final Agreement on the AI Act: AI System, General Purpose AI System, Foundation Model, and Generative AI' (2025) 33 Artificial Intelligence and Law 875, 880.

[272] AI Act, arts 3(65) and 9(2)(a).

[273] AI Act, art 3(65) and recital 110.

[274] See the forthcoming commentary on Article 3(65).

[275] ibid.

[276] AI Act, recital 110.

[277] Code of Practice, Safety and Security Chapter (n 9) app 1.3.1.

[278] Code of Practice, Safety and Security Chapter (n 9) app 1.3.2.

[279] Code of Practice, Safety and Security Chapter (n 9) app 1.3.3.

[280] AI Act, art 55(1)(b).

[281] Code of Practice, Safety and Security Chapter (n 9) Commitment 2 and Glossary, definition of 'systemic risk assessment'.

[282] Code of Practice, Safety and Security Chapter (n 9) Commitment 3.

[283] Code of Practice, Safety and Security Chapter (n 9) Commitment 4.

[284] Code of Practice, Safety and Security Chapter (n 9) Commitment 5.

iterative,[285] spanning the model's entire lifecycle,[286] and updated until systemic risk is brought to an *acceptable* level.[287]

84. The Code of Practice's structuring of the risk assessment and mitigation process is also consistent with the way in which the AI Act envisions risk management for providers of high-risk AI systems.[288] Article 9(2) explicitly lists risk identification, risk analysis, and risk evaluation as distinct and sequential steps within a 'continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system'.[289] Arguments can thus be levied in support of a systematic reading of the AI Act,[290] under which the risk assessment and mitigation process is understood to follow a similar structure across providers of high-risk AI systems and providers of GPAI models with systemic risk.

85. Interpretative support may also be drawn from international risk management standards, in particular ISO/IEC Guide 51 and the related standards ISO 31000 and ISO 73.[291] While these standards are not expressly referenced in the AI Act, the terminology and risk management structure adopted in the Code of Practice (and in Article 9)[292] mirrors their risk management framework.[293] The following analysis therefore draws on the ISO standards only insofar as their terminology and structure support the interpretation of Article 55(1) and the corresponding commitments in the GPAI Code of Practice.[294] Where the term *assessment* is used throughout the rest of this chapter, it should thus be understood as encompassing the identification, analysis, and acceptance determination of possible systemic risks.

[285] Code of Practice, Safety and Security Chapter (n 9) recital (c).

[286] Code of Practice, Safety and Security Chapter (n 9) recital (a); AI Act, recital 110.

[287] Code of Practice, Safety and Security Chapter (n 9) Commitment 1, para 1 ['The purpose of the Framework is to outline the systemic risk management processes and measures that Signatories implement to ensure the systemic risks stemming from their models are acceptable.'].

[288] Schuett (n 44) 377–378.

[289] AI Act, art 9(2).

[290] On systemic methods of interpretation, see Petrić (n 198) 19.

[291] ISO and IEC, 'Safety Aspects — Guidelines for Their Inclusion in Standards' (ISO/IEC 2014) ISO/IEC Guide 51:2014 <https://www.iso.org/obp/ui/#iso:std:iso-iec:guide:51:ed-3:v1:en> accessed 18 May 2026, s 6; These guidelines inform the application of safety principles within the broader risk management process outlined in ISO 31000:2009(E) (n 33); European Commission, '1st European AI Office Webinar on Risk Management Logic of the AI Act and Related Standards' (*European Commission*, 30 May 2024) <https://digital-strategy.ec.europa.eu/en/events/1st-european-ai-office-webinar-risk-management-logic-ai-act-and-related-standards> accessed 18 May 2026, 9 ['This definition [of risk] aligns with the definition of risk in other NLF legislation and e.g. with Safety Risk Management per ISO Guide 51.']; Schuett (n 44) 375 ['But as the risk management process in the AI Act seems to be inspired by ISO/IEC Guide 51, I use or adapt many of their definitions.'].

[292] Schuett (n 44) 377–378.

[293] Herwig CH Hofmann, 'The Integration of Global Standards into the EU as "Regulatory Union"' (University of Luxembourg 2022) Law Research Paper 2022–006 <https://doi.org/10.2139/ssrn.4240982> accessed 18 May 2026, 14: 'European Standardisation Organisations are not isolated from international standard setting. They incorporate international standards and international best practices and take these into account.'.

[294] AI Act, recital 121; Code of Practice, Safety and Security Chapter (n 9) recital (d): 'The Signatories also recognise that they may be able to rely on international standards to the extent they cover the provisions of this Chapter.'

2.1.2.2.1. Risk assessment

2.1.2.2.1.1. Systemic risk identification

86. Risk identification is the starting point of the risk assessment and mitigation process,[295] at which stage providers engage in 'finding, recognising and describing' all possible risks stemming from the model.[296] The AI Act does not provide specific methods for identifying risks. Providers may therefore draw on established risk identification techniques and methodologies,[297] including those described in the GPAI Code of Practice.

87. Under the Code of Practice's Safety and Security Chapter, systemic risk identification proceeds in two steps. First, providers need to follow a structured process to identify and compile a list of *potential*[298] systemic risks identified through a broader information-gathering exercise about the model,[299] and the *specified* systemic risks that have been pre-identified and listed in Appendix 1.4.[300]

88. For the purposes of identifying potential systemic risks, signatories are expected to draw on a set of five 'distinct but in some cases overlapping types of risks' listed in Appendix 1.1. These include risks to public health, safety, public security, fundamental rights, and society as a whole. This categorisation reflects a broad set of protected public interests and largely mirrors the definition of systemic risk set out in Article 3(65).[301] Appendix 1.1 further provides a non-exhaustive list of examples falling within these categories, including risks of major accidents; risks affecting critical sectors or infrastructure; impacts on public mental health; and risks to fundamental rights such as freedom of expression and information, non-discrimination, privacy, and the protection of personal data. It also includes risks to the environment, non-human welfare, economic security, and democratic processes, as well as risks arising from the concentration of power and from illegal, violent, hateful, radicalising, or false content, including child sexual abuse material ("CSAM") and non-consensual intimate images ("NCII").

89. Based on the types of risks listed in Appendix 1.1, signatories are required to consider a range of information sources, including (i) model-independent information; (ii) relevant information concerning the model and similar models, including information derived from post-market monitoring, serious incidents, and near misses; and (iii) any other relevant information communicated to providers by the AI Office, the Scientific Panel, or other relevant initiatives.[302] Providers are then required to analyse relevant characteristics of the compiled risks,[303] such as their

[295] Schuett (n 44) 375.
[296] ISO 31000:2009(E) (n 33) s 2.15 on risk identification.
[297] Braun Binder and Egli, 'Art. 9' (n 233) para 24 ['Risikoermittlung bedeutet die systematische Nutzung verfügbarer Informationen zur Identifizierung von Gefahren. Die Gefahr kann dabei als potenzielle Schadensquelle definiert werden. Da die KI-Verordnung nicht vorschreibt, wie die Anbieter Risiken ermitteln sollen, müssen sie sich auf bestehende Techniken und Methoden stützen.'].
[298] See Section 2.1.2.1. on difference between potential and possible systemic risks.
[299] Code of Practice, Safety and Security Chapter (n 9) Commitment 2, Measure 2.1.
[300] Code of Practice, Safety and Security Chapter (n 9) Commitment 2, Measure 2.1(2).
[301] Also see the forthcoming commentary on Article 3(65) in this work.
[302] Code of Practice, Safety and Security Chapter (n 9) Measure 2.1(1).
[303] Code of Practice, Safety and Security Chapter (n 9) Measure 2.1(1)(b).

nature and sources,[304] and to identify which of them constitute systemic risks stemming from the model.[305]

90. Alongside potential systemic risks, providers are by default required to assess whether their model displays any of the specified systemic risks listed in Annex 1.4 of the Code of Practice.[306] These include risks relating to (1) chemical, biological, radiological and nuclear harms; (2) loss of control; (3) cyber-offence; and (4) harmful manipulation.[307] The *specified* systemic risks are *based on* and may correspond to the types of risks listed in Annex 1.1,[308] which in turn include risks listed in Recital 110.

91. For each identified systemic risk, providers are then required to develop appropriate systemic risk scenarios, including by determining the number of such scenarios and the level of detail at which they are described.[309] The Glossary defines a *systemic risk scenario* as a scenario in which a systemic risk stemming from a model might materialise.[310] These systemic risk scenarios are to form the basis for the systemic risk modelling that signatories must undertake as part of the subsequent systemic risk analysis stage of the full assessment and mitigation process.[311]

2.1.2.2.1.2. Systemic risk analysis

92. Signatories are required to engage in risk analysis to develop as complete an understanding as possible of the identified risks, including their nature, sources, and level.[312] In situations of high uncertainty, the ISO 31000 guidelines recommend using a combination of qualitative and quantitative techniques to assess the probability of risk occurring and the magnitude or level of risk should it materialise.[313] Any determinations as to the level of risk made at this stage will inform the next step of systemic risk acceptance determinations. Providers may, and signatories should, rely on the Code of Practice for guidance on what methods and techniques to use for the purposes of systemic risk analysis.

93. More specifically, Commitment 3 of the Safety and Security Chapter requires providers to analyse each identified systemic risk, with the outcome of that analysis informing the subsequent determination of systemic risk acceptance.[314] At this stage, providers must first gather model-independent information relevant to the identified systemic risk using methods such as web searches and literature reviews, market analyses that involve assessing the capabilities of other models, reviews of training data for indications of data poisoning or tampering, and analyses of historical

[304] See Code of Practice, Safety and Security Chapter (n 9) apps 1.2 and 1.3.
[305] Code of Practice, Safety and Security Chapter (n 9) Measure 2.1(1)(c).
[306] Code of Practice, Safety and Security Chapter (n 9) Measure 2.1(2).
[307] Code of Practice, Safety and Security Chapter (n 9) app 1.4.
[308] For example, CBRN risks may arise where models lower the barriers to entry for malicious actors in the design, development, acquisition, distribution, or use of chemical or biological weapons. Such risks may manifest in threats to public health, the environment, and non-human welfare, as reflected in Code of Practice, Safety and Security Chapter (n 9) app 1.1. Systemic risks to public security may include cyber-offence risks, such as enabling large-scale or sophisticated cyber-attacks, including attacks targeting critical systems.
[309] Code of Practice, Safety and Security Chapter (n 9) Measure 2.2.
[310] Code of Practice, Safety and Security Chapter (n 9) Glossary.
[311] Code of Practice, Safety and Security Chapter (n 9) Commitment 3, Measure 3.3.
[312] Koessler and Schuett (n 44) 2; IEC and ISO, 'Risk management — Risk assessment techniques' (IEC and ISO 2019) IEC 31010:2019 <https://www.iso.org/standard/72140.html> accessed 18 May 2026.
[313] ISO 31000:2009(E) (n 33) ss 2.21 and 2.23.
[314] Code of Practice, Safety and Security Chapter (n 9) Commitment 3.

incident data and incident databases.[315] The breadth and depth of this information-gathering exercise depend on the probability and severity of harm.[316] Providers then need to conduct model evaluations that are at least state of the art and appropriate to both the model and the systemic risk in question.[317] The third step requires providers to conduct systemic risk modelling,[318] which involves specifying the pathways through which a systemic risk may materialise.[319]

94. Providers should then use at least state-of-the-art risk methods to estimate both the probability and the severity of harm for each identified systemic risk.[320] Estimates of systemic risk could be expressed using formats such as risk scores, risk matrices,[321] probability distributions, or other suitable representations, and may be quantitative, semi-quantitative, or qualitative in nature. The fifth and final element of the systemic risk analysis process is post-market monitoring.[322] This process involves gathering information about the model's capabilities, propensities, affordances, and/or effects over the period from when the model is placed on the market until the retirement of the model from being made available on the market.[323] During post-market monitoring, providers commit to providing adequate access to the model to an adequate number of independent external evaluators.[324] Providers are exempt from having to grant access to external evaluators where the model qualifies as a similarly safe or safer model with regard to the same systemic risk, as specified in Appendix 2.2.[325]

2.1.2.2.1.3. Systemic risk acceptance determination

95. Risk acceptance determination, which also corresponds to risk evaluation in established risk management practices,[326] is the next step in the risk assessment process, at which it is necessary to determine whether the analysed risks are *acceptable* or whether they exceed a level that cannot be

[315] Code of Practice, Safety and Security Chapter (n 9) Commitment 3, Measure 3.1.
[316] Code of Practice, Safety and Security Chapter (n 9) Measure 1.2(1)(d).
[317] See Sections 2.1.1.1. and 2.1.1.2.
[318] Code of Practice, Safety and Security Chapter (n 9) Commitment 3, Measure 3.3.
[319] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'systemic risk scenario' as 'a scenario in which a systemic risk stemming from a model might materialise.'
[320] Code of Practice, Safety and Security Chapter (n 9) Commitment 3, Measure 3.4.
[321] Risk matrices are one of the most common risk evaluation techniques, where '[a] risk matrix, also known as heat map or consequence/likelihood matrix, is a table that contains consequence and likelihood ratings of different risks, often on a scale from 1 to 5. Each cell represents a specific combination of consequence and likelihood. Different risks can be plotted on the matrix to determine the need and priority of addressing them (IEC, 2019). Risk matrices are one of the most common risk evaluation techniques.' Koessler and Schuett (n 44) 25; IEC 31010:2019 (n 312); definition of 'risk matrix' in ISO Guide 73:2009 (n 43) s 3.3.5.8 as a 'tool for ranking and displaying risks by defining ranges for consequence and likelihood'.
[322] Code of Practice, Safety and Security Chapter (n 9) Commitment 3, Measure 3.5.
[323] Code of Practice, Safety and Security Chapter (n 9) Glossary defines 'appropriate' as 'suitable and necessary to achieve the intended purpose of systemic risk assessment and/or mitigation, whether through best practices, the state of the art, or other more innovative processes, measures, methodologies, methods, or techniques that go beyond the state of the art.'
[324] See Section 2.1.1.2.3.
[325] Code of Practice, Safety and Security Chapter (n 9) app 2; See also Meta (n 147) s 2.1.1, 4. [In Meta's three-stage governance approach, the first step, *Anticipate*, consists in identifying comparable models to serve as a reference class. Meta compares its models against those available externally on the basis of anticipated capabilities, supported modalities, intended uses, and projected compute requirements. This comparison is used to identify an estimated reference class of comparable models, which is then relied upon throughout the development process to track the model's performance and to inform both the evaluations to be conducted and the mitigations to be implemented. Where it is expected that a model may significantly exceed current frontier capabilities, Meta will conduct *ex ante* threat modelling exercises to assess whether the model may give rise to novel risks.].
[326] e.g. in ISO 31000:2009(E) (n 33).

tolerated and instead require mitigation.[327] This is achieved by comparing the results of the risk analysis against previously established risk criteria and determining where risk reduction measures are needed.[328]

96. Commitment 4 of the Safety and Security Chapter covers systemic risk acceptance determination. This commitment requires providers to specify systemic risk acceptance criteria, which are then used to determine whether each identified systemic risk, as well as the overall systemic risk profile, are acceptable.[329] These criteria must incorporate a safety margin to account for potential limitations, uncertainties, or changes relating to the source of the systemic risk, the systemic risk assessment, and the effectiveness of the mitigation measures.[330] The signatories have discretion to develop acceptance criteria suitable for the systemic risk at issue unless criteria are prescribed for specified systemic risks pursuant to Appendix 1.4.[331] In evaluating the specified systemic risks as listed in Appendix 1.4, signatories commit to using appropriate systemic risk *tiers* defined in terms of model capabilities, and may additionally incorporate model propensities, risk estimates,[332] and other metrics.[333] The tiers must be measurable and include at least one systemic risk tier that has not yet been reached.[334]

97. Only where, on the basis of the systemic risk acceptance criteria, each identified systemic risk and the overall systemic risk are determined to be acceptable may providers proceed with the development, placing on the market, and/or use of the GPAI model.[335] Should the results of the systemic risk acceptance determination reveal that the systemic risks stemming from the model are unacceptable, or are reasonably foreseeable to soon no longer be determined acceptable, the providers must refrain from making the model available on the market, or, where necessary, restrict its availability, withdraw it, or recall it.[336] Signatories are required to then implement *appropriate* safety and security mitigations pursuant to Commitments 5 and 6 respectively and conduct another round of systemic risk identification, systemic risk analysis, and systemic risk acceptance determination until the systemic risks are deemed to be acceptable.[337]

2.1.2.2.2. Risk mitigation

98. At the stage of risk mitigation, providers are required to select and implement measures to address and mitigate identified systemic risks.[338] Risk mitigation is an iterative process involving the selection and implementation of mitigation measures, the assessment of whether the resulting residual risk

[327] ISO/IEC Guide 51:2014 (n 291) 2; For the purposes of the ISO/IEC Guide 51, the terms 'acceptable risk' and 'tolerable risk' are considered to be synonymous.
[328] Koessler and Schuett (n 44) 24; IEC 31010:2019 (n 312) s 6.4.4 on risk evaluation; ISO Guide 73:2009 (n 43) s 3.3.6., definition of *risk evaluation* as 'process of comparing the results of risk analysis against risk criteria to determine whether the level of risk (3.3.5.10) is acceptable or tolerable'.
[329] Code of Practice, Safety and Security Chapter (n 9) Measure 4.1.
[330] ibid.
[331] Code of Practice, Safety and Security Chapter (n 9) Measure 4.1(1)(b).
[332] See Leonie Koessler, Jonas Schuett and Markus Anderljung, 'Risk Thresholds for Frontier AI' (arXiv, 20 June 2024) <https://arxiv.org/abs/2406.14713v1> accessed 18 May 2026.
[333] Code of Practice, Safety and Security Chapter (n 9) Measure 4.1(a)(i).
[334] Code of Practice, Safety and Security Chapter (n 9) Measure 4.1(a)(ii) and (iii).
[335] Code of Practice, Safety and Security Chapter (n 9) Measure 4.2.
[336] Code of Practice, Safety and Security Chapter (n 9) Commitment 4, Measure 4.2(1).
[337] Code of Practice, Safety and Security Chapter (n 9) Commitment 4, Measure 4.2(2) and (3).
[338] Code of Practice, Safety and Security Chapter (n 9) Commitment 5; More broadly on 'risk management', see ISO Guide 73:2009 (n 43) s 3.8.1 defining *risk treatment*.

has been brought to an acceptable level, and, where it has not, the identification and implementation of further measures.[339] Given that the AI Act does not prescribe how risk mitigation is to be conducted or what specific mitigation measures must be implemented for the purposes of complying with Article 55(1)(b), the primary responsibility lies with providers to identify and make the case for what they consider to be appropriate mitigations.[340] At the same time, the AI Office retains the authority to assess the appropriateness of chosen mitigation measures and may, through the structured dialogue mechanism under Article 93(2), signal what it considers to constitute appropriate risk mitigation.[341] The GPAI Code of Practice offers further guidance on how providers of GPAI models with systemic risk may implement such measures for the purposes of demonstrating compliance with their obligations.[342]

99. The Safety and Security Chapter of the GPAI Code of Practice requires providers of GPAI models with systemic risk to implement appropriate safety and security mitigations along the model's entire lifecycle.[343] Security mitigations are framed in terms of ensuring an adequate level of cybersecurity protection for the model and its physical infrastructure, with a view to preventing systemic risks that may arise from unauthorised access, release, and/or model theft.[344] Security mitigations are particularly relevant to the obligation set out in Article 55(1)(d).[345] Safety mitigations are directed at ensuring that the systemic risks stemming from the model are brought to an acceptable level.[346] Measure 5.1 of the Safety and Security Chapter contains a non-exhaustive list of safety mitigations providers may implement in the course of bringing systemic risk to an acceptable level.[347] This may include, for example, technical measures at the level of the AI model itself (such as filtering inputs or outputs), technical instructions to downstream providers to provide certain information or notices, or monitoring and reviewing the model for risky behaviour.

100. The requirement that mitigation measures be 'appropriate' implies that they must be sufficiently robust under adversarial conditions, taking into account, *inter alia*, the model's release and distribution strategy, which may itself constitute a source of systemic risk.[348] Model evaluations are particularly relevant at the risk mitigation stage,[349] as they can be employed to test the effectiveness of the safety mitigations by revealing, for example, whether the model remains susceptible to

[339] Code of Practice, Safety and Security Chapter (n 9) Measure 4.2; ISO 31000:2009(E) (n 33) s 6.5.1.

[340] Schneider (n 20) para 9; Providers may refer to shared taxonomies and databases for AI risk mitigations, such as the preliminary AI Risk Mitigation Database and Taxonomy, which together provide an empirical and conceptual foundation for a more coordinated, comprehensive approach to mitigating AI risks, Peter Slattery and others, 'MIT AI Risk Repository' (*MIT AI Risk Initiative*) <https://airisk.mit.edu/> accessed 18 May 2026; see also Alexander K Saeri and others, 'Mapping AI Risk Mitigations: Evidence Scan and Preliminary AI Risk Mitigation Taxonomy' (arXiv, 12 December 2025) <https://doi.org/10.48550/arXiv.2512.11931> accessed 18 May 2026, 3 ['Some focus on adapting established mitigations from cybersecurity or safety-critical industries (e.g., incident response, system shutdown; Koessler & Schuett, 2023), while others introduce novel approaches specific to AI (e.g., alignment techniques, model interpretability; Ji et al., 2023)'.].

[341] AI Act, art 93 and recital 164.

[342] Code of Practice, Safety and Security Chapter (n 9) Commitments 5 and 6; See also Frontier Model Forum, 'Frontier Mitigations' (Frontier Model Forum 2025) <https://www.frontiermodelforum.org/technical-reports/frontier-mitigations/> accessed 18 May 2026 for an overview of emerging industry practices for implementing and assessing frontier mitigations.

[343] Code of Practice, Safety and Security Chapter (n 9) Commitments 5 and 6.

[344] Code of Practice, Safety and Security Chapter (n 9) Commitment 6.

[345] See Section 2.1.4.

[346] Code of Practice, Safety and Security Chapter (n 9) Commitment 5.

[347] Code of Practice, Safety and Security Chapter (n 9) Commitment 5, Measure 5.1.

[348] Code of Practice, Safety and Security Chapter (n 9) Measure 5.1 and app 1.3.3(3).

[349] See Section 2.1.1.2.

jailbreaking or other adversarial attacks.[350] The results of the model evaluations will determine whether improved mitigation measures must be implemented, or whether the systemic risks stemming from the model are acceptable and providers may thus proceed with the development, making available on the market, or use of the model.[351]

101. Further interpretative guidance may also be drawn from the way in which the AI Act frames risk mitigation in the context of high-risk AI systems.[352] The risk management process outlined in Article 9, similarly to the framework constructed around GPAI models with systemic risk, imposes a similar sequence of risk identification, risk analysis and risk management upon providers of high-risk AI systems. Notably, unlike Article 9, Article 55 does not impose comparable limitations on the scope of risk mitigation measures that providers of GPAI models with systemic risk are required to implement.[353] Under Article 9(3), risk mitigation is confined to risks that can be adequately mitigated or eliminated through system design and development, or through the provision of appropriate technical information. In other words, the scope of risk management under Article 9 is confined to risks that can be addressed through the measures listed in Article 9(5)(a)–(c), with the consequence that where residual risks cannot be reduced to an acceptable level the system may not be placed on the market.[354] The absence of an equivalent limitation in Article 55(1)(b) suggests that the scope of mitigation for GPAI models with systemic risk may indicate that providers are required to engage with systemic risks more broadly, including in situations where such risks cannot be fully mitigated through design, or technical measures, or provision of information alone.

2.1.2.3. 'Appropriate' measures for 'acceptable' risk

2.1.2.3.1. Appropriate risk assessment and mitigation measures

102. The objective of the risk assessment and mitigation process is to ensure that systemic risks stemming from the model are brought to an *acceptable* level.[355] Across this process,[356] the measures adopted must be *appropriate*,[357] that is, 'suitable and necessary to achieve the intended purpose of systemic risk assessment and/or mitigation, whether through best practices, the state of the art, or other more innovative processes, measures, methodologies, methods, or techniques that *go beyond the state of the art*.'[358] In this context, suitability, as one limb of the proportionality test,[359] requires that a given measure be capable of effectively contributing to the objective of reducing systemic risk to an acceptable level. The suitability of a measure may also be assessed by reference to how targeted it

[350] Code of Practice, Safety and Security Chapter (n 9) Measure 5.1.
[351] Code of Practice, Safety and Security Chapter (n 9) Measure 4.2.
[352] AI Act, art 9.
[353] See Section 2.1.2.2.2.
[354] Finck (n 36) para 4.190, citing Schuett (n 44) 377 'Art 9 does not say this explicitly, but it seems to be a logical consequence of the process'.
[355] Code of Practice, Safety and Security Chapter (n 9) Commitment 1.
[356] The GPAI Code of Practice requires that all measures taken at any step across the full systemic risk assessment and mitigation process be appropriate.
[357] Under article 9(5), providers of high-risk AI systems are required to adopt '[s]uitable and targeted risk management measures.
[358] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'appropriate' (emphasis added).
[359] Takis Tridimas, 'The Principle of Proportionality' in Robert Schütze and Takis Tridimas (eds), *Oxford Principles Of European Union Law: The European Union Legal Order*, vol 1 (Oxford University Press 2018) <https://doi.org/10.1093/oso/9780199533770.003.0010> accessed 18 May 2026; See also Matthias Klatt and Moritz Meister, *The Constitutional Structure of Proportionality* (Oxford University Press 2012) <https://doi.org/10.1093/acprof:oso/9780199662463.001.0001> accessed 18 May 2026.

is, that is, the extent to which it relates to the identified risk.[360] This is reflected in the GPAI Code of Practice's preference for providers to adopt targeted measures that address specific risks without unduly impairing beneficial model capabilities.[361]

103. *Necessity*, in turn, requires that equal or superior safety or security outcomes cannot be achieved through alternative means that are less burdensome or more efficient.[362] In this respect, the notion of *appropriateness* under Article 55(1)(b) aligns with the reading of *appropriate* as used in Article 9(4),[363] where it is understood as comprising two interrelated dimensions: first, *effectiveness*, in the sense that the measure must be capable of mitigating the risk in light of the current state of technology, and second, *proportionality*, in that the burden imposed by the measure must not be grossly disproportionate to the level of risk reduction achieved.[364]

104. In addition to being informed by the principle of proportionality, the notion of appropriateness is also closely linked to the state-of-the-art condition.[365] In selecting appropriate measures for assessing and mitigating systemic risk, the GPAI Code of Practice requires that providers consider 'best practices, the state of the art, or other more innovative processes, measures, methodologies, methods, or techniques that go beyond the state of the art.'[366] If a provider adopts only minimal measures while peers in the field typically implement more robust safeguards, such measures should be regarded as inappropriate. Improvements in developing mitigation measures corroborate that *appropriateness* is a dynamic standard,[367] given that if more robust training techniques or improved testing practices become established in the industry, a provider should incorporate them or have a compelling reason as to why not.[368]

2.1.2.3.2. Acceptable levels of systemic risk

105. Providers of GPAI models with systemic risk must adopt appropriate risk assessment and mitigation measures until each identified systemic risk and the overall systemic risk is deemed to be acceptable.[369] This formulation presupposes that some degree of residual risk will remain following

---

[360] Code of Practice, Safety and Security Chapter (n 9) recital (f); König, 'Art. 9 Risikomanagementsystem' in David Bomhard, Fritz-Ulli Pieper, and Susanne Wende (eds), *KI-VO Verordnung über künstliche Intelligenz* (1st edn, Deutscher Fachverlag 2025) para 24; Gerdemann, 'Art 9' (n 234) para 56 [there is an expectation that mitigation measures be specifically directed at identified risks, rather than reducing risk in an undifferentiated manner.].

[361] Code of Practice, Safety and Security Chapter (n 9) recital (f).

[362] ibid.

[363] Fraser and Bello y Villarino (n 106) 438.

[364] Teichmann (n 238) 8 ['For instance, if a slight software tweak can prevent a serious failure mode, it is appropriate and expected to implement it, but if addressing a very marginal risk would require an enormous expense or fundamentally alter the system's utility, it might be beyond reasonably practicable and thus not mandated.']; see also Mónica Álvarez Fernández, 'Risk Management System (Article 9)' in Alejandro Huergo Lora (ed) and Gustavo Manuel Díaz González (coord), *The EU Regulation on Artificial Intelligence: A Commentary* (Wolters Kluwer Italia 2025) 175.

[365] See Fraser and Bello y Villarino (n 106).

[366] Code of Practice, Safety and Security Chapter (n 9) Glossary, definition of 'appropriate'.

[367] In the context of high-risk AI systems, see AI Act, recital 65: 'The risk-management system should adopt the most appropriate risk-management measures in light of the state of the art in AI.'

[368] Code of Practice, Safety and Security Chapter (n 9) recital (a) ['the Signatories recognise that implementing appropriate measures will often require Signatories to adopt at least the state of the art, unless systemic risk can be conclusively ruled out with a less advanced process, measure, methodology, method, or technique.'].

[369] Code of Practice, Safety and Security Chapter (n 9) Measure 4.2; see also AI Act, art 9(5), which requires that the risk management measures referred to under article 9(2)(d) have to be such that the relevant residual risk is deemed 'acceptable'; See also Gerdemann, 'Art 9' (n 234) para 66 ['Accordingly, in the final assessment of the acceptability of residual risks, providers cannot simply resort to making general statements about the acceptable risk exposure of

the implementation of mitigation measures. Residual risks are those risks that are left over after mitigations have been implemented.[370] What qualifies as an acceptable level of residual risk is, in the first instance, determined by providers themselves, in particular through the thresholds they establish to assess whether a given risk is acceptable or requires further mitigation.[371]

106. *Acceptability* is defined elsewhere in terms of the two core dimensions of risk, namely the severity and the probability of harm.[372] Accordingly, the higher the probability of harm and the more serious the nature and extent of the potential damage, the less likely it is that a risk will be considered acceptable.[373] Further insight into what constitutes acceptable or tolerable residual risk can be derived from ISO Guide 51. According to this guidance, a number of factors may be taken into account when determining whether risk is *tolerable* or *acceptable*, including prevailing societal values, the need to strike a balance between the ideal of absolute safety and what is achievable, the demands to be met by the product or system, and considerations such as suitability for purpose and cost-effectiveness.[374] Providers may draw also on these factors in assessing whether residual risk is acceptable in light of the objectives of the AI Act,[375] which require balancing the promotion of innovation with a high level of protection of health, safety, and fundamental rights.[376]

107. The inclusion of risks to fundamental rights within the notion of systemic risk, and thus into the process of risk acceptance determination,[377] introduces a further layer of normative complexity.[378] Specifically, it highlights the tension of trust in providers to independently assess how much encroachment on fundamental rights may be considered acceptable as residual risk.[379] In this

---

the AI system, but must instead relate to specific risks or dangers to particular legal interests arising from the specific AI system [...]']; See also Peter Sebelius, 'Policy for Establishing Criteria for Risk Acceptability ISO 14971:2019' (*Medical Device HQ*, 5 December 2023) <https://medicaldevicehq.com/articles/policy-for-establishing-criteria-for-risk-acceptability-according-to-iso-149712019/> accessed 18 May 2026 ['The requirement on having a policy for establishing criteria for risk acceptability was added to the ISO 14971:2019 version of the standard. The requirement is particularly important to meet MDR and IVDR requirements on risk management. The reason for the addition of the requirement of having a policy for establishing criteria for risk acceptability in the ISO 14971:2019 version of the standard was that the concept was often misunderstood in the previous 2007 version of the standard. The "risk policy" was often replaced with only a risk evaluation matrix as seen below. This was not the intent of the standard. [...] management must define and document a policy that is the starting point for the determination of criteria for risk acceptability. Thus, the risk acceptability criteria should be derived from the policy'].

[370] ISO/IEC Guide 51:2014 (n 291) s 3.8 on definition of residual risk.

[371] Code of Practice, Safety and Security Chapter (n 9) Measure 4.1 on systemic risk acceptance determination.

[372] AI Act, art 3(2) on the definition of 'risk'.

[373] Gerdemann, 'Art 9' (n 234) para 63.

[374] ISO/IEC Guide 51:2014 (n 291) s 6.2.1; The ISO guide also treated the terms 'tolerable risk' and acceptable risk as synonymous (see s 3.15); See also, Gerdemann, 'Art 9' (n 234) para 64.

[375] Braun Binder and Egli, 'Art 9' (n 232) para 33 ['Die Risikominimierungspflicht der KI-VO wird folglich dadurch relativiert, dass von den Anbietern kein völlig fehlerfreies Hochrisiko-KI-System erwartet wird. Vielmehr geht der europäische Gesetzgeber davon aus, dass gewisse verbleibende Restrisiken vertretbar sind.'].

[376] AI Act, art 1(1); see the forthcoming commentary on Article 1 in this work; Gerdemann, 'Art 9' (n 234) para 64.

[377] Sebelius (n 369) ['The requirement on having a policy for establishing criteria for risk acceptability was added to the ISO 14971:2019 version of the standard. The requirement is particularly important to meet MDR and IVDR requirements on risk management. The reason for the addition of the requirement of having a policy for establishing criteria for risk acceptability in the ISO 14971:2019 version of the standard was that the concept was often misunderstood in the previous 2007 version of the standard. The "risk policy" was often replaced with only a risk evaluation matrix as seen below. This was not the intent of the standard. [...] management must define and document a policy that is the starting point for the determination of criteria for risk acceptability. Thus, the risk acceptability criteria should be derived from the policy'].

[378] Carsten Orwat and others, 'Normative Challenges of Risk Regulation of Artificial Intelligence and Automated Decision-Making' (arXiv, 11 November 2022) <https://arxiv.org/abs/2211.06203v1> accessed 18 May 2026, 23.

[379] Finck (n 36) para 4.194, citing Schuett.

respect, the regulatory framework implicitly accepts that certain degrees of interference with fundamental rights may be tolerated.[380] The AI Act does not provide a clear margin of appreciation for such harms, instead relying on general principles such as proportionality and the state of the art.

108. Indeed, one consequence of linking acceptability to the state-of-the-art condition is that the level of mitigation expected for a given capability or risk profile may increase over time as available safety and security practices evolve.[381] As more effective risk mitigation techniques become available, the continued presence of certain risks may no longer be acceptable, meaning that the threshold of acceptable risk evolves alongside other dynamic references in the AI Act, including measures that are appropriate and reflect the state of the art. In this respect, the AI Act does not require absolute safety, but rather a level of relative risk reduction that reflects what is currently achievable.[382]

2.1.2.4. Timing of risk assessment and mitigation measures

109. Article 55(1)(b) requires providers to assess and mitigate possible systemic risks that may arise across the model's entire lifecycle, giving particular notice to those systemic risks that arise during development, placing on the market, and use of the model.

2.1.2.4.1. Development

110. Although not defined by the AI Act, *development* can, in the context of Article 55, reasonably be understood to encompass the initial design and creation of the model, including data compilation, training, and fine-tuning.[383] Systemic risks may stem from choices made during this stage, even if their harmful effects only materialise after the model has been placed on the market. In that sense, risks both arise from the development stage and may already be identifiable at that stage.

111. While Article 2(8) excludes research, testing and development activities prior to market placement from the scope of the regulation,[384] its impact is nuanced and does not preclude providers from being required to conduct risk assessment and mitigation before placing a model on the market.[385] To ensure that, at the point of market placement, any residual systemic risk has been reduced to an acceptable level, providers must undertake the necessary risk assessment and mitigation processes during development.[386] The exclusion of development activities as such does not remove the obligation to ensure compliance at the moment of placing the model on the market.[387]

---

[380] Carsten Orwat and others (n 378) 23.

[381] Fraser and Bello y Villarino (n 106) 432. If, however, capabilities increase, an evolving state of the art may still imply that the total level of risk might not diminish.

[382] Gerdemann, 'Art 9' (n 234) para 64.

[383] See also the forthcoming commentary on Article 2 in this work.

[384] AI Act, art 2(8).

[385] See, more extensively, the forthcoming commentary on Article 2 in this work.

[386] Also see the forthcoming chapter on Product, Model and Entity Regulation in this work.

[387] For more on market access being conditional upon compliance as a feature of product safety, see the forthcoming commentary on Article 2 in this work; AI Act, recital 97, 'It should be understood that the obligations for the providers of general-purpose AI models should apply once the general-purpose AI models are placed on the market. [...] Considering their potential significantly negative effects, the general-purpose AI models with systemic risk should always be subject to the relevant obligations under this Regulation.'

2.1.2.4.2. Placing on the market

112. A GPAI model may be placed on the market in a variety of ways, including through 'libraries, application programming interfaces (APIs), as direct download, or as physical copy.'[388] In requiring providers to assess and mitigate risks that may stem from the placing on the market, attention should be paid both to systemic risks that emerge *after* market placement and to risks that arise *because of* the manner in which the GPAI model is placed on the market.[389]

113. As to the former, the AI Act recognises that the full range and nature of systemic risks may become clear to providers only after market placement and in the course of the model's interaction with users.[390] To this end, providers are required to conduct post-market monitoring in order to gather information about the model's capabilities, propensities, affordances and effects, which may in turn inform whether the systemic risk should be considered acceptable.[391] Post-market monitoring may reveal that the model's capabilities, propensities or affordances have materially changed, such as through further post-training, access to additional tools, or increased inference compute.[392] It may also reveal developments that materially undermine the external validity of model evaluations previously conducted, materially improve the state of the art in evaluation methods, or otherwise suggest that the systemic risk assessment carried out was materially inaccurate.[393]

114. Systemic risk may also be influenced by the manner in which the GPAI model is placed on the market.[394] Indeed, in assessing whether a GPAI model could present systemic risk as part of the classification process under Article 51, the AI Office 'could take into account the way the model will be placed on the market or the number of users it may affect.'[395] The Safety and Security Chapter of the Code of Practice also requires providers to implement safety mitigations that take account of the model's release and distribution strategy, for example by 'staging the access to the model, e.g. by limiting API access to vetted users, gradually expanding access based on post-market monitoring, and/or not making the model parameters publicly available for download initially.'[396]

2.1.2.4.3. Use of the model

115. Providers of GPAI models with systemic risk are also required to assess and mitigate possible systemic risks that may arise during the use of the model.[397] A systematic reading of this obligation in light of the risk management process established for high-risk AI systems and the product-safety

---

[388] AI Act, recital 97.

[389] Code of Practice, Safety and Security Chapter (n 9) app 1.3.3, listing release and distribution strategy as a possible source of systemic risk.

[390] AI Act, recital 111: 'The full range of capabilities in a model could be better understood after its placing on the market or when deployers interact with the model.'; Beurskens (n 20) para 6: 'some systemic will "only become apparent upon placing the AI on the market (e.g., through inquiries from downstream providers.)"'.

[391] Post-market monitoring is defined in the Code of Practice Glossary as 'the monitoring of a model in the time span from when it is placed on the market until the retirement of the model from being made available on the market' and required by Code of Practice, Safety and Security Chapter (n 9) Measure 3.5. Post-market monitoring is also a component of the risk assessment process for high-risk AI systems, see AI Act, art 9(2)(c).

[392] Code of Practice, Safety and Security Chapter (n 9) Measure 7.6(2).

[393] Code of Practice, Safety and Security Chapter (n 9) Measure 7.6(5).

[394] AI Act, recital 110. Also see the forthcoming commentary on Article 3(9) in this work.

[395] AI Act, recital 110 and annex XIII(f).

[396] AI Act, art 111; Code of Practice, Safety and Security Chapter (n 9) app 1.3.3, 36.

[397] In this chapter, the notion of 'use' of the model is used loosely, also clearly capturing instances where the model is used through a system. For a discussion on whether models can be used directly, also see the forthcoming commentary on Article 2 in this work.

logic underpinning the AI Act's risk-based approach[398] indicates that Article 55(1)(b) requires providers to also consider systemic risks that may stem from the *misuse* of their models.[399] Notably, the AI Act does not define what it means to *use* or *misuse* a (GPAI) model.[400] It does, however, speak to what it means to *use* and *misuse* a high-risk AI system.

116. Article 9(2)(a) requires providers of high-risk AI systems to assess risks that may arise when the AI system is used in accordance with its intended purpose. The concept of (intended) *use* can be traced back to the ISO Guide 51 on the inclusion of safety aspects in standards,[401] which defines *intended use* as 'use in accordance with information provided with a product or system, or, in the absence of such information, by generally understood patterns of usage.' ISO Guide 51 distinguishes between *intended use*, which is used synonymously with reasonably foreseeable use, and reasonably foreseeable misuse.[402] While the AI Act does not define intended 'use' as such, it similarly frames the use of an AI system by reference to the intentions of the supplier, or, in the terminology of the AI Act, the intended purpose as envisaged by the provider. Article 3(12) defines *intended purpose* as 'the use for which an AI system is intended by the provider, including the specific context and conditions of use'.[403] The intended purpose of a system may be communicated by the provider or deduced from the instructions for use, promotional or sales materials and statements, and the technical documentation prepared by the provider.[404]

117. To meaningfully fulfil their obligations under Article 55(1)(b), providers must assess and mitigate risks that may stem from the ways their GPAI model is used. However, given the very nature of such models as general-purpose and capable of being deployed across a potential vast range of contexts, GPAI models do not easily lend themselves to having an *intended purpose* as envisioned in Article 3(12).[405] It is likewise unclear whether establishing such an intended purpose is a necessary precondition for the applicability and enforceability of Article 55(1). Even if a GPAI model cannot be said to have a clearly delineated intended purpose, it may nevertheless have an intended use, however broadly framed, as communicated by the provider through system cards, model cards, and other public-facing documentation. Intended use may also be inferred from generally understood patterns of usage, which providers are likely to monitor through post-market surveillance and market monitoring practices. The absence of a clearly defined intended purpose to which usage scenarios can be tethered, may indeed prompt providers to exercise greater effort in identifying, analysing and mitigating reasonably foreseeable systemic risks. In this context, the requirement of

[398] AI Act, art 9(2), recital 9 and recital 46; See Alessandro Mantelero, 'Conformity Assessment, Quality and Risk Management Systems (Articles 8, 9, 17, 42, 43, 46)' in Gianclaudio Malgieri and others (eds) *The EU Artificial Intelligence Act: A Thematic Commentary* (Hart Publishing 2026) 237, 242.

[399] AI Act, art 28: 'Aside from the many beneficial uses of AI, it can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices.'

[400] Also see the forthcoming commentary on Article 2 in this work.

[401] ISO/IEC Guide 51:2014 (n 291) s 3.6 on *intended use* as 'use in accordance with information provided with a product or system, or, in the absence of such information, by generally understood patterns of usage'.

[402] See Blue Guide (n 61) s 2.8.

[403] AI Act, art 3(2).

[404] See Anthropic, 'Usage Policy' (*Anthropic*, 2025) <https://www.anthropic.com/legal/aup> accessed 18 May 2026; a company's usage policy dictates what uses of its AI systems are acceptable or unacceptable. Usage policies generally prohibit inputs that elicit a range of undesirable model outputs, beyond what is already illegal. Longpre and others (n 171) 2

[405] See Boine and Rolnick (n 271).

reasonableness shifts the burden onto the provider to 'research and understand the user environment and likely failure modes.'[406]

118. Providers would also have to assess and mitigate reasonably foreseeable misuses, which includes use of a model in a way not intended by the provider but which can result from reasonably foreseeable human behaviour.[407] This means that providers have to consider ways their model can be intentionally or otherwise used in a manner not intended by the provider.[408] While Article 55(1) does not explicitly refer to misuse, Recital 110 makes clear that 'systemic risks [...] are influenced by conditions of misuse,' noting that international approaches have identified the need to pay particular attention to risks arising from potential intentional misuse or from unintended issues of control relating to alignment with human intent. The Safety and Security Chapter of the Code of Practice requires providers to consider both *intentional misuse* and *unintended model behaviour.*[409]

119. Intentional misuse of GPAI models involves the use of the model 'by malicious actors to enable or scale harmful activities.'[410] The Code of Practice lists instances of intentional misuse, including, but not limited to, cyberattacks, development and use of CBRN capabilities, and large-scale disinformation.[411] In identifying reasonably foreseeable risks arising from intentional misuse, the Code of Practice requires providers to consider the potential number, capacity, and motivation of malicious actors to misuse a model as a potential source of systemic risk.[412] This is further reflected in the requirements concerning model elicitation, under which providers must undertake model evaluation that at least matches the elicitation capabilities of misuse actors relevant to the systemic risk scenario.[413]

120. Unintended model behaviour refers to instances in which systemic risk arises from the use of the model in ways not anticipated by the provider, but not because the model is being deliberately used contrary to the provider's intentions or instructions.[414] Rather, the model may behave 'in ways that developers and users did not intend, or be unsafe in ways that could plausibly cause large-scale harm. This includes highly consequential accidents caused by inadequate capabilities, alignment, or

[406] Teichmann (n 238) 7 ['For instance, when deploying an AI system in a critical infrastructure setting, it is reasonably foreseeable that human operators might misuse it under pressure or that malicious actors might attempt to manipulate it.'].

[407] On high-risk AI systems, AI Act, recital 65 says providers should consider risks that may stem from 'from readily predictable human behaviour'.

[408] ISO/IEC Guide 51:2014 (n 291) para 3.7.

[409] Code of Practice, Safety and Security Chapter (n 9) app 1.4(2), definition of 'loss of control' as 'Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems. See also Miles Brundage and others, 'Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies' (arXiv, 7 February 2026) <https://doi.org/10.48550/arXiv.2601.11699> accessed 18 May 2026, 26.

[410] Brundage and others (n 409) 90.

[411] Code of Practice, Safety and Security Chapter (n 9) app 1.4. Other types of intentional misuse that may not qualify as systemic risk include 'violent and criminal activity; fraud; and the generation of child sexual abuse material or nonconsensual intimate imagery', Brundage and others (n 409) 90; U.S. AI Safety Institute, 'Managing Misuse Risk for Dual-Use Foundation Models' (National Institute of Standards and Technology 2024) NIST AI 800-1 ipd NIST AI 800-1 ipd <https://doi.org/10.6028/NIST.AI.800-1.ipd> accessed 18 May 2026.

[412] Code of Practice, Safety and Security Chapter (n 9) app 1.3.3(9), 36.

[413] Code of Practice, Safety and Security Chapter (n 9) app 3.2.

[414] Brundage and others (n 409) 25.

safeguards.'[415] The Code of Practice recognises such instances of unintended model behaviour as falling within the broader category of loss of control.[416]

2.1.3. Article 55(1)(c): Handling of serious incidents

2.1.3.1. General remarks

121. Article 55(1)(c) addresses the obligations of providers of GPAI models with systemic risk regarding the handling of serious incidents. According to this provision, providers of GPAI models with systemic risk must 'keep track of, document, and report without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them'.

2.1.3.1.1. Rationale

122. The purpose of this provision is to ensure transparency regarding relevant incidents vis-à-vis the AI Office as the competent authority.[417] In addition, the provision aims to enable coordinated responses to serious incidents by the AI Office and providers of GPAI models, thereby preventing escalation after an incident, restoring (or preserving) the capacity to act and preventing further harm.[418] Furthermore, the provision facilitates the AI Office's accumulation of knowledge, as incident reports provide up-to-date insights – particularly with respect to potential attack vectors[419] as well as potentially harmful patterns.[420] Additionally, Article 55(1)(c) also promotes the exchange and dissemination of knowledge and thereby reduces potential information asymmetries between government and industry.[421] This also helps to inform future regulatory efforts as well as the development of best practices.[422] Finally, the provision ultimately helps to build public trust in AI technologies in general by enhancing proper oversight.[423]

2.1.3.1.2. Incident reporting obligations in EU law

123. The obligation to report serious incidents to authorities has been established across European legislation well before the AI Act. In particular, similar provisions can be found in many EU

[415] ibid.

[416] Code of Practice, Safety and Security Chapter (n 9) app 1.4.

[417] See similarly with regard to the parallel obligation for high-risk AI system providers in article 73, Sarah Hartmann 'Art. 73 Meldung schwerwiegender Vorfälle' in Mario Martini and Christiane Wendehorst (eds), *KI-VO: Verordnung über Künstliche Intelligenz* (2nd edn, C.H. Beck, 2026) para 1.

[418] With regard to article 73, Christian Djeffal, 'Art.73 Meldung schwerwiegender Vorfälle' in Jens Schefzig and Robert Kilian (eds), *Beck'scher Online-Kommentar KI-Recht* (4th edn, C.H. Beck 2025) para 1.

[419] AI Act, art 64(1): 'The Commission shall develop Union expertise and capabilities in the field of AI through the AI Office'.

[420] With regard to article 73, European Commission, 'AI Act: Commission Issues Draft Guidance and Reporting Template on Serious AI Incidents, and Seeks Stakeholders' Feedback' (*European Commission*, 26 September 2025) <https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks> accessed 26 September 2025 ("Draft Guidance on Article 73"), para 2.

[421] Furthermore, information sharing may be facilitated by the (additionally applicable) obligation of providers under article 53(3) to cooperate with the Commission and national competent authorities.

[422] Rishi Bommasani and others, 'The California Report on Frontier AI Policy' (arXiv, 17 June 2025) <https://doi.org/10.48550/arXiv.2506.17303> accessed 27 September 2025, 31; see also Kevin Wei and Lennart Heim, 'Designing Incident Reporting Systems for Harms from General-Purpose AI' (arXiv, 2025) <https://doi.org/10.48550/arXiv.2511.05914> accessed 23 February 2026.

[423] Draft Guidance on Article 73 (n 420) para 2.

Regulations and Directives – for example in Article 87(1) Medical Devices Regulation ("MDR")[424], Article 23(1) NIS2 Directive ("NIS2")[425], Article 19(1) Digital Operational Resilience Act ("DORA")[426], and Article 14(1) Cyber Resilience Act ("CRA")[427]. Article 33(1) GDPR likewise contains an at least comparable obligation.[428]

124. Although these various provisions pursue different objectives in general,[429] (and also use slightly different wording[430]) one might still be able to draw inspiration from them for the interpretation of Article 55(1)(c).[431] This is why the following discussion will, where relevant, refer to the reporting obligations under the aforementioned instruments.

2.1.3.1.3. Internal systematics of the AI Act and interaction with other EU legal instruments

125. The obligation under Article 55(1)(c) entails interactions both within the AI Act and with other EU legal instruments. The most important interaction within the AI Act is that between Article 55(1)(c) and Article 3(49), as the latter defines the term 'serious incident' that is referred to in Article 55(1)(c). By its wording, however, the definition found in Article 3(49) refers only to AI *systems*, and not GPAI models (with systemic risk). The Commission appears to assume the existence of a certain link between the two provisions.[432] The relationship between the two provisions will be examined in more detail below.

---

[424] 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) 178/2002 and Regulation (EC) 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Medical Devices Regulation) [2017] OJ L 117/1 ("MDR").

[425] Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) [2022] OJ L 333/80 ("NIS2").

[426] Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011 [2022] OJ L 333/1 ("DORA").

[427] Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act) [2024] OJ L 2847/1 ("CRA").

[428] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1 ("GDPR").

[429] The MDR 'aims to ensure the smooth functioning of the internal market as regards medical devices, taking as a base a high level of protection of health for patients and users, and taking into account the small- and medium-sized enterprises that are active in this sector. It also 'sets high standards of quality and safety for medical devices in order to meet common safety concerns as regards such products' (recital 2); the NIS2 Directive aims 'to build cybersecurity capabilities across the Union, mitigate threats to network and information systems used to provide essential services in key sectors and ensure the continuity of such services when facing incidents, thus contributing to the Union's security and to the effective functioning of its economy and society' (recital 1); the DORA aims to 'achieve a high common level of digital operational resilience' for financial entities (art 1) and the CRA 'aims to set the boundary conditions for the development of secure products with digital elements by ensuring that hardware and software products are placed on the market with fewer vulnerabilities and that manufacturers take security seriously throughout a product's lifecycle' (recital 2).

[430] MDR, art 87 addresses 'serious incidents'; NIS2, art 23 addresses 'significant incidents'; DORA, art 19 addresses 'major ICT-related incidents'; and CRA, art 14 addresses 'severe incidents'.

[431] In detail, see the forthcoming chapter on Interpreting the AI Act through Systematic Analogies in this work.

[432] Commission Guidelines (n 16) para 100 pointing towards a link between the term 'serious incident' in article 55(1)(c) and in article 3(49) by saying that '[a]part from this, the AI Office considers a "serious incident" in the context of Chapter V AI Act as any incident or malfunctioning of a general-purpose AI model that directly or

126. Additionally, attention should be given to Article 73 of the AI Act on serious incident reporting in the context of high-risk AI systems. That provision contains a very similar obligation to report serious incidents for providers of high-risk AI systems, though its content is considerably more detailed. The Commission has published its draft guidance on the reporting obligation under Article 73, to which this contribution will partly refer.[433] Although that draft guidance expressly states that it is not intended to apply to Article 55(1)(c),[434] it nevertheless contains some valuable indications on how the Commission interprets the term 'serious incident' in Article 3(49), to which Article 55(1)(c) appears to refer.

127. Notably, Article 73(9) governs some of the interaction with other EU legal instruments more precisely. The provision limits the reporting obligation for high-risk AI systems by providers that are subject to Union legislative instruments that lay down their own reporting obligations equivalent to those in the AI Act. Under Article 73(1), those providers only remain obliged to report serious incidents that directly or indirectly lead to an infringement of obligations under Union law intended to protect fundamental rights (Article 3(49)(c)).[435] The reporting obligations under Article 19 DORA, Article 23 NIS2 and Article 14 CRA qualify as equivalent in that sense.[436] High-risk AI system providers subject to these obligations are therefore partially protected by Article 73(9) against duplicate reporting duties with regard to serious incidents and thereby relieved of some administrative burden.[437]

128. By contrast, no such rule exists under Chapter V in general or under Article 55 specifically for providers of GPAI models with systemic risk. Therefore, providers of GPAI models with systemic risk may be subject to parallel reporting obligations under the CRA, the NIS2 Directive or the DORA under certain circumstances, resulting in an increased compliance burden. This means that, in any case, providers of GPAI models with systemic risk must always report to the AI Office – even if they already reported the same incident pursuant to another instrument to another institution.[438]

---

indirectly leads to any of the events listed in the corresponding definition for AI systems in Article 3(49), points (a) to (d), AI Act'.

[433] Draft Guidance on Article 73 (n 420).

[434] ibid para 4.

[435] Hartmann (n 417) para 17; Djeffal (n 418) para 67; see also Susanne Wende, 'Art. 73 Meldung schwerwiegender Vorfälle' in David Bomhard, Fritz-Ulli Pieper & Susanne Wende (eds), *KI-VO: Verordnung über künstliche Intelligenz* (Deutscher Fachverlag, 2025) para 30.

[436] In detail, Djeffal (n 418) para 70 ff; agreeing to this Hartmann (n 417) para 17; see similarly Finck (n 36) para 10.53 and Draft Guidance on Article 73 (n 420) para 56 ff.

[437] Djeffal (n 418) para 67.

[438] It cannot be ruled out that the Digital Omnibus, in its final version, will also streamline the reporting obligations under the AI Act. At present, however, the proposal to establish a 'single entry point' by ENISA does not cover reporting under the AI Act; see European Commission, 'Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725 and (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150 and (EU) 2022/868 and Directive (EU) 2019/1024 (Digital Omnibus)' COM (2025) 837 final, art 6.

2.1.3.2. Relevant information about serious incidents

2.1.3.2.1. Serious incident

129. As indicated earlier, Article 3(49) seems, at first glance, to define what constitutes a serious incident under the AI Act.[439] According to this provision, 'serious incident' means 'an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following:

> (a) the death of a person, or serious harm to a person's health;
> (b) a serious or irreversible disruption of the management or operation of critical infrastructure;
> (c) the infringement of obligations under Union law intended to protect fundamental rights;
> *(d)* serious harm to property or the environment.'

130. As mentioned before, a difficulty arises from the fact that the definition expressly refers only to incidents or malfunctionings of AI *systems*, but not of (general-purpose) AI *models*. This opens three different pathways for interpretation. First, the definition could be understood to mean that an incident or a malfunctioning of a standalone GPAI model with systemic risk can never give rise to a serious incident within the meaning of the AI Act – meaning that such an event could only occur if the GPAI model with systemic risk has been integrated into an AI system.[440] Second, one could assume that this is simply a drafting error, with Article 3(49) to be read as if it stated: 'an incident or malfunctioning of an AI system or GPAI model with systemic risk'.[441] Third, one might conclude that the definition in Article 3(49) is not exclusively determinative for Article 55(1)(c) but serves as the structural basis for a broader understanding of a *serious incident* aligned with Article 55's overarching objective to assess and mitigate systemic risks.

2.1.3.2.1.1. Need for integration into an AI system?

131. As explained above, the definition in Article 3(49) could be read to mean that standalone GPAI models with systemic risk cannot cause serious incidents within the meaning of the AI Act as long as they are not integrated into AI systems.[442] Recital 115, however, speaks against this interpretation, by referring to situations in which 'the development or use of the model causes a serious incident'. This recital therefore appears to assume that models themselves can also cause serious incidents. Further, the GPAI Code of Practice supports an understanding under which GPAI models with systemic risk alone can cause serious incidents, as it defines a 'resolved serious incident' as a 'serious incident *of a model*...'.[443] Moreover, both the Commission's and the AI Board's adequacy

[439] Wei and Heim (n 422) note that '[s]ome jurisdictions have recognized the distinctions between incident types' but that 'the EU AI Act establishes a single incident reporting requirement, with the definition of an incident including both rights and safety incidents'.
[440] Nynke Elske Vellinga and Jeanne Mifsud Bonnici, 'Article 55 Obligations of Providers of General-Purpose AI Models with Systemic Risk' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act* (Wolters Kluwer, 2025) 868.
[441] Schneider (n 20) para 14.
[442] ibid; similarly Eric Hilgendorf & Johannes Härtlein, 'Art. 55 Pflichten der Anbieter von KI-Modellen mit allgemeinem Verwendungszweck mit systemischem Risiko' in Eric Hilgendorf & Johannes Härtlein (eds), *Verordnung über künstliche Intelligenz: KI-VO* (C.H. Beck, 2025) para 4; also see Beurskens (n 20) para 7 who appears to presuppose the transmission of the information along the AI value chain and the corresponding acquisition of knowledge by the provider of the GPAI model with systemic risk as a necessary precondition.
[443] Code of Practice, Safety and Security Chapter (n 9) 32, emphasis added.

assessments of the GPAI Code of Practice also expressly refer to 'serious incidents *of the model*'.[444] Finally, from a teleological perspective, interpreting the term serious incident as exclusive to AI systems could exclude some systemic risks, such as loss of control over the model,[445] from the reporting obligation, thereby undermining the rationale of Article 55 in assessing and mitigating systemic risks.

132. On the other hand, it could be argued that a serious incident cannot occur without having people interact with the GPAI model with systemic risk and that this generally requires integration into an AI system.[446] This reading could be supported by the fact that other obligations in Article 55 explicitly establish a direct reference to the respective GPAI model with systemic risk (point (a): 'perform *model* evaluation [...] including conducting and documenting adversarial testing *of the model*' (emphasis added); point (b): 'assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use *of general-purpose AI models* with systemic risk' (emphasis added); point (d): 'ensure an adequate level of cybersecurity protection *for the general-purpose AI model* with systemic risk') (emphasis added), whereas this reference is missing in point (c) – indicating that the most direct point of reference in 55(1)(c) is not the model itself, but the respective AI system it is integrated in.

2.1.3.2.1.2. Application of Article 3(49) serious incident definition?

133. There are arguments to support a reading under which Article 3(49) is exclusively determinative for the concept of a serious incident in Article 55(1)(c). One can argue that, in substance, Article 55(1)(c) seeks to ensure transparency in situations where particularly severe consequences have occurred and these consequences were caused by objects regulated by the AI Act. From a teleological point of view – having the aim of the AI Act to ensure a high level of health, safety and fundamental rights protection in mind – it should make no difference whether an AI system or a GPAI model with systemic risk caused the serious consequences.[447] Moreover, a divergent definition of the term 'serious incident' within the AI Act could potentially lead to reporting gaps in cases where a GPAI model with systemic risk is integrated into an AI system and where it is unclear whether the incident was (directly or indirectly) caused by the GPAI model or by the AI system. That could be the case in instances where the model is integrated into an AI system that is not considered to be a high-risk system,[448] and there is resultantly no obligation for system providers to

[444] Commission Opinion (n 37) para 37; European Artificial Intelligence Board, 'Conclusion of the Artificial Intelligence Board on the Assessment of the General-Purpose AI Code of Practice pursuant to Article 56 of Regulation 2024/1689 (Artificial Intelligence Act or "AI Act")' (*European Commission*) <https://digital-strategy.ec.europa.eu/en/policies/ai-board\> accessed 12 February 2026, 10 (emphasis added).

[445] From a comparative perspective it might be worth mentioning that California SB-53, 'Artificial intelligence models: large developers' [2025], s 22757.11(d)(3) encompasses '[l]oss of control of a frontier model causing death or bodily injury' as a '[c]ritical safety incident'; also see Section 2.1.3.2.1.4.

[446] The AI Act seems to largely presuppose such integration in an AI system before a GPAI model can be used. For a discussion on whether an AI model can be 'put into service' (more directly), see the forthcoming chapter on Article 2 in this work.

[447] Similarly, Schneider (n 20) para 13; also see Michael Chatzipanagiotis, 'Incident Reporting and Investigation under the AI Act: Some Insights from Aviation' (2026) 34 International Journal of Law and Information Technology eaaf019.

[448] Systems built on GPAI models with systemic risk should not automatically qualify as high-risk AI systems, see, in detail, Moritz Stilz, 'KI-Systeme mit allgemeinem Verwendungszweck: automatisch Hochrisiko?'(2026) Künstliche Intelligenz und Recht 39.

report serious incidents.[449] Additionally, in cases where the model is integrated into a high-risk AI system, the provider of the high-risk AI system could argue that it was not the system but only the model that caused the serious incident.

134. Diverging definitions of the notion of a 'serious incident' could, in such cases, create additional uncertainty and result – in the worst case, especially if both concepts of a serious incident were to be interpreted narrowly – in neither the provider of the potentially involved GPAI model nor the provider of the potentially involved AI system reporting the event. Moreover, pursuant to Article 73, providers of high-risk AI systems do not need to directly report serious incidents to the AI Office, but rather to the national competent authorities.[450] To avoid reporting gaps and to ensure the development of expertise and capability in the AI Office,[451] one could argue that a divergent interpretation of the term in Article 3(49) and in Article 55(1)(c) should therefore be avoided and that the definition found in Article 3(49) should also be exclusively determinative for Article 55(1)(c).

2.1.3.2.1.3. Article 3(49) as the structural basis for the serious incident definition in Article 55(1)(c)?

135. It could also be argued that GPAI models with systemic risk can cause serious incidents, but that such incidents are not to be understood exclusively in the sense defined by Article 3(49).

136. It seems likely that the Commission does not hold the definition provided in Article 3(49) to be exclusively determinative for Article 55(1)(c). The Commission 'considers a "serious incident" in the context of Chapter V AI Act as any incident or malfunctioning of a general-purpose AI model that directly or indirectly leads to any of the events listed' in Article 3(49)(a)–(d).[452] In this respect, the result of the Commission's interpretation seems to be the same as if the definition in Article 3(49) were assumed to apply to GPAI models as well. At the same time, however, the Commission appears to adopt a broader understanding than determined in Article 3(49), as it sees 'serious cybersecurity breaches' as falling under the concept of serious incidents of GPAI models under Article 55(1)(c) 'due to their possible implications for the obligations provided for in Article 55(1), points (b) and (d)'.[453]

137. The Code of Practice likewise suggests that the definition contained in Article 3(49) does not fully determine Article 55(1)(c)'s reporting obligation. Rather, the Code of Practice seems to indicate that what amounts to a *serious incident* should be interpreted in light of the systemic risks that Article 55 seeks to capture. To this end, the Safety and Security Chapter of the GPAI Code of

[449] Chatzipanagiotis (n 447) 36 proposes to extend reporting obligations to 'all developers and deployers of AI systems, whether they are classified as high-risk or not' because otherwise 'potentially valuable information and experiences might go unnoticed'.

[450] Which, according to article 73(11), shall in turn notify the Commission in accordance with article 20 of the Market Surveillance Regulation; pursuant to article 75(1), however, AI system providers will arguably be required to report to the AI Office where the high-risk AI system is based on a GPAI model and the model and the system are developed by the same provider; see the forthcoming commentary on Article 75 in this work. This is now also expressly addressed in article 75(1ab) of the agreed text of the AI Omnibus proposal: Council of the European Union, 'Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence – Letter sent to the European Parliament' (13 May 2026) ST 9247/26.

[451] AI Act, art 64.

[452] Commission Guidelines (n 16) para 100.

[453] ibid.

Practice, under Measure 9.3, like the Commission Guidelines,[454] also refers to 'a serious cybersecurity breach' as a situation that triggers the reporting obligation under Article 55(1)(c) – a category that does not appear as a distinct one in Article 3(49).

138. Additionally, the GPAI Code of Practice states that, for the purpose of assessing a model's systemic risk under the Safety and Security Framework which signatories are required to set up, not only 'serious incidents' but also 'near misses' can be relevant. For instance, near misses may serve as indicators that the Safety and Security Framework requires updating.[455] On its face, one could argue that this only indicates that near misses need to be documented for the purposes of fulfilling Article 55(1)(a) and (b)'s obligations. According to the GPAI Code of Practice's Glossary, however, a 'near miss' is defined as a situation in which 'a serious incident could have, but ultimately did not, materialise'.[456] This could be read as indicating that the materialisation of any harm is not necessary under the Code's definition of a serious incident, given that the concept of a serious incident forms part of the near miss definition itself. In relation to GPAI models with systemic risk, the Code therefore could be understood to interpret the term 'serious incident' in a manner similar to the earlier Commission's proposal, which, by using the formulation 'might have led', explicitly encompassed such near misses.[457] One might therefore conclude that, while the term 'serious incident' continues to follow the definition in Article 3(49) for high-risk AI systems, it should be understood more broadly for GPAI models with systemic risk to also include near misses. This reading is, however, open to challenge on two counts.

139. An argument against this latter interpretation can be found in the GPAI Code of Practice itself, which clearly distinguishes between near misses and serious incidents in (all) other instances.[458] Additionally, Measure 9.3 of the Safety and Security Chapter explicitly only sets out reporting timelines for situations in which the involvement of the provider's model '(directly or indirectly) led' – and not 'might have led' – to specific outcomes.[459] Accordingly, as further discussed below,[460] there likely exists no reporting obligation pursuant to Article 55(1)(c) for near misses under the Code of Practice and Article 55(1)(c).

140. However, the legislative history of Article 3(49)[461] seems to imply that Article 3(49) should not necessarily serve as the structural basis for the concept of serious incidents under Article 55(1)(c). The definition of a 'serious incident' in the initial Commission proposal did not contain the wording that the incident must be that 'of an AI system'.[462] This addition was only introduced in the Council's

---

[454] ibid.

[455] In cases where the near miss involves the providers model or a similar model and the near miss is 'likely to indicate that the systemic risks stemming from at least one of their models are not acceptable have occurred', see Code of Practice, Safety and Security Chapter (n 9) Measure 1.3(2).

[456] Code of Practice, Safety and Security Chapter (n 9) 31.

[457] European Commission 'Proposal for a Regulation laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act)' COM (2021) 206 final' ("AI Act Proposal"), art 3(44).

[458] See Code of Practice, Safety and Security Chapter (n 9) Measure 1.3(2) 'serious incidents and/or near misses', Measure 2.1(1)(a)(ii) 'serious incidents and near misses), Measure 7.6(4) 'serious incidents and/or near misses', Measure 9.2(9) 'any patterns [...] that can reasonably assumed to be connected to the serious incident, such as (...) data on near misses'.

[459] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.

[460] See Section 2.1.3.2.1.4.

[461] Also see Schneider (n 20) para 12.

[462] AI Act Proposal (n 457) art 3(44).

proposal[463] and ultimately found its way into the final legal text of the AI Act. Because the first rules on GPAI (at that time called 'general-purpose AI systems') were also incorporated into the AI Act for the first time during the Council's proposal stage,[464] the simultaneous addition of 'of an AI system' instead of 'of an AI system or GPAI model' into the definition might seem deliberate.[465] However, it seems equally plausible that the absence of the wording 'or a GPAI model' may have been a drafting oversight following the introduction of the first special rules on general-purpose AI systems. It can additionally be argued that, at the time of adding provisions on GPAI, the wording 'of an AI system' was also intended to cover GPAI, which at that stage in the legislative process – as stated above – was referred to as 'general purpose AI systems'.[466]

2.1.3.2.1.4. Synthesis

141. Some ambiguity clearly persists. The most convincing arguments, however, appear to support a partial reliance on the definition of a 'serious incident' in Article 3(49) as a structural basis, while recognising that Article 55(1)(c) ultimately addresses different risks than the ones found in high-risk AI system contexts (Article 73 AI Act). For the practical application of Article 55(1)(c), and based on a purposive reading of it, the following approach seems appropriate: as a general rule, the definition set out in Article 3(49), interpreted so as to include GPAI models with systemic risk, should apply.[467] In view of the AI Act's objective to 'ensure a high level of protection of health, safety and fundamental rights', it makes sense to establish a reporting obligation concerning the consequences listed in Article 3(49), irrespective of whether they result from a high-risk AI system or a GPAI model with systemic risk[468] (or both) and to keep the AI Office informed in this regard. This approach prevents uncertainty and reporting gaps in cases where a GPAI model with systemic risk is integrated into a non-high-risk AI system and encompasses the cases in which a model was not integrated into a system.

142. At the same time, however, there are good reasons to argue that the definition of a 'serious incident' should be extended in the context of Article 55(1)(c), given the risks inherent to GPAI models presenting systemic risk. This is also indicated by both the Commission GPAI Guidelines and the GPAI Code of Practice.[469] The most obvious example being the 'serious cybersecurity breach', which both the Code and the Guidelines refer to.[470]

---

[463] Draft Regulation General Approach (n 244).

[464] ibid art 3(44).

[465] Schneider (n 20) para 14.

[466] Note, however, that the concept of a general-purpose AI system is now defined in article 3(66) of the enacted AI Act as 'an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems'. The Council's proposal defined a general purpose AI system as 'an AI system that - irrespective of how it is placed on the market or put into service, including as open source software - is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems'. The latter definition seems to be more closely related to what a GPAI model is under the enacted AIA, rather than the current definition of a GPAI system under Art 3(66).

[467] Similarly, Finck (n 36) para 6.117 ['in the absence of another definition, Article 3(49) should also be relied on for the interpretation of Article 55(1)(c)'].

[468] Schneider (n 20) para 13.

[469] See Section 2.1.3.2.1.3.

[470] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3(2) and Commission Guidelines (n 16) para 100.

143. Beyond that, but probably less convincing,[471] one could argue that an even broader interpretation appears justified, encompassing additional consequences insofar as they materially increase the systemic risk posed by the model – as serious cybersecurity breaches arguably do – or they constitute a materialisation of a systemic risk. More generally, insofar as the obligation to keep track of serious incidents serves to inform the provider's systemic risk assessment,[472] there also seem to be reasons to require the provider to keep track of, document and report all incidents that materially increase or lead to the materialisation of systematic risk – and not just serious cybersecurity breaches, as the Commission GPAI Guidelines and the GPAI Code of Practice propose.[473] Expanding Article 55(1)(c) in this way arguably seems important because incidents that materially increase systemic risk are not captured by Article 3(49)'s definition – which is tailored to downstream harm caused by AI systems. Ultimately, the reporting obligation serves not only the purpose of making the AI Office and the provider aware of previously unidentified risks – expanding the obligation to report to include material increases of the systemic risk posed by the model ensures that both also become aware of the (near-)materialisation of already known risks which, in light of the incident, may no longer be seen as adequately mitigated.

144. There are also policy considerations that might support this broader interpretation. California Transparency in Frontier Artificial Intelligence Act ("SB-53") – which, as the first comprehensive frontier AI safety statute enacted in the United States undoubtedly carries international significance – similarly addresses the inherent risks of highly capable models within its reporting provisions.[474] It is particularly noteworthy in this regard that the scope of the reporting obligation in SB-53 is confined to large frontier developers and models[475] – and thus differs structurally from the general incident reporting framework of the AI Act, which was conceived primarily with high-risk AI systems in mind. The definition of a 'critical safety incident' under SB-53 encompasses, for instance, '[h]arm resulting from the *materialization of a catastrophic risk*' (emphasis added) and cases in which a 'frontier model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates *materially increased catastrophic risk*' (emphasis added).[476] It is precisely this model-specific tailoring of SB-53's incident reporting concept that makes the comparison structurally apt.

145. One might be tempted to introduce these policy considerations – at least indirectly – into the legal interpretation of the term 'serious incident' in Article 55(1)(c) via Article 56(1), which mandates the AI Office to 'take into account international approaches'. However, this line of argument faces several objections. First, the obligation to take international approaches into account binds the AI Office only in the specific context of encouraging and facilitating the drawing up of codes of practice

[471] See Section 2.1.3.2.2.5. on possible counter-arguments.

[472] Code of Practice, Safety and Security Chapter (n 9) Measure 1.3(2).

[473] Commission Guidelines (n 16) para 100; Code of Practice, Safety and Security Chapter (n 9) Measure 9.3(2).

[474] SB-53 (n 445) s 22757.11(i)(1) defines 'frontier model' as 'a foundation model that was trained using a quantity of computing power greater than 10^26 integer or floating-point operations'.

[475] SB-53 (n 445) s 22757.13(c)(1) establishing an obligation for large frontier developers to report critical safety incidents pertaining to their frontier models. Also consider that SB-53, s 22757.13(c)(4) prescribes that developers are 'encouraged, but not required, to report critical safety incidents pertaining to foundation models that are *not* frontier models' (emphasis added).

[476] SB-53 (n 445) s 22757.11(d); see also Mariami Tkeshelashvili, Ritika Verma and Steven M Kelly, 'AI Loss of Control Risk: Indications & Warning' (Institute for Security and Technology 2026) <https://securityandtechnology.org/wp-content/uploads/2026/02/AI-Loss-of-Control-Risk-1.pdf> accessed 20 March 2026, propose a framework to monitor loss of control risks and also give examples of documented instances of frontier models deploying deceptive techniques to subvert operator controls.

– not in the interpretation of the AI Act itself.[477] Second, even if one were to argue that the GPAI Code of Practice – having been recognised by the Commission as adequate – informs the interpretation of Article 55(1)(c), this reasoning must ultimately be abandoned in light of the chronological sequence: SB-53 was enacted after the GPAI Code of Practice had already been finalised. In the end, this comparative argument is therefore one that carries no normative force – though it might retain persuasive value from a policy perspective.

146. There is also a teleological argument to extend the concept of a 'serious incident' within the meaning of Article 55(1)(c) to cover near misses as well.[478] This expansion is suggested by the GPAI Code of Practice's definition of a near miss and would also be supported by a purposive reading of Article 55(1)(c). In view of the expected information asymmetry between providers of GPAI models with systemic risk and the AI Office, and the AI Office's need to be able to detect harmful patterns and emerging risk vectors at an early stage, it may be inappropriate in such a sensitive area as systemic risk assessment and mitigation to wait until risks have materialised.[479]

147. However, several arguments weigh against including near misses into the definition of a serious incident. First, earlier drafts of the definition of a serious incident included wording that would have covered near misses ('might have led'). This wording was removed during the drafting process, signalling the EU legislature's intent to keep the two concepts separate.[480] Likewise, the GPAI Code of Practice refers only to reporting obligations in cases where the involvement of the model *led* to a particular outcome.[481] It also clearly distinguishes – as noted above, notwithstanding the broader definition in the Glossary – between near misses and serious incidents, suggesting that near misses fall outside the serious incident definition.[482] The argument that (all)[483] near misses need to be reportable to inform the AI Office at an early stage is also not particularly persuasive. That is because providers may need to document and assess near misses under Article 55(1)(a) and (b),[484] enabling the AI Office to request that information pursuant to Article 91. Further, near misses trigger an update of a signatory's Safety and Security Framework[485] in cases where they are 'likely to indicate that the systemic risks stemming from at least one of [the signatories] models are not acceptable have occurred.'[486] Such updates must be notified to the AI Office within five business days.[487] As a result, the view that near misses are not captured by Article 55(1)(c)'s reporting obligation is more compelling than the alternative.

---

[477] Also see the commentary on Article 56 in this work, paras 14–15.

[478] See Section 2.1.3.2.1.3.

[479] Also see, more generally, Wei and Heim (n 422) on why near misses are a 'valuable source of data for safety learning' and stating that 'the vast majority of safety incidents are not harm events but rather near misses'; similarly Ren Bin Lee Dixon and Heather Frase, 'AI Incidents: Key Components for a Mandatory Reporting Regime' (Center for Security and Emerging Technology 2025) <https://doi.org/10.51593/20240023> accessed 23 February 2026 ['near misses should be included within the scope of mandatory reporting. Reporting near misses can enhance incident data collection, as these events exhibit similar characteristics to incidents, apart from their outcomes. In addition to aiding early detection of novel AI risks, tracking near misses could reveal vital conditions that prevented harm from occurring, which can be leveraged to strengthen safety measures'].

[480] AI Act Proposal (n 457) art 3(44) as well as Parliament Amendments (n 99) art 3(44); see also Chatzipanagiotis (n 447) 6.

[481] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.

[482] Code of Practice, Safety and Security Chapter (n 9) Measures 1.3(2), 2.1(1)(a)(ii), 7.6(4) and 9.2(9).

[483] See, regarding possible overlaps between near misses and material increases in systemic risk, Section 2.1.3.2.2.5.

[484] See Section 2.1.3.2.1.3.

[485] Code of Practice, Safety and Security Chapter (n 9) Measure 1.3(2).

[486] ibid.

[487] ibid Measure 1.4; Chatzipanagiotis (n 447) 35 proposes to establish voluntary reporting systems for near misses.

148. Since, as demonstrated above, the interpretation of the term 'serious incident' in Article 55(1)(c) is built on the definition set out in Article 3(49) as its structural basis, the following analysis is structured in accordance with that definition. Particular attention will be drawn at the relevant points to the specific considerations that arise for GPAI models with systemic risk in this context.

2.1.3.2.1.5. Incident or malfunctioning

149. Article 3(49) defines a serious incident as 'an incident or malfunctioning'. The Commission likewise refers to this phrase in its GPAI Guidelines.[488] However, what is to be understood by 'incident or malfunctioning' is not further defined in the AI Act.[489] There seem to be two possible pathways for interpretation.

150. On the one hand, because the reporting of an incident does not amount to an admission of wrongdoing[490] and because the AI Office needs to be kept up to date regarding existent and emerging systemic risks within the Union, one could – from a teleological point of view – argue that what matters is merely whether the GPAI model with systemic risk was somehow involved.[491] The term 'incident' would then not carry an additional independent meaning as long as the causality requirement is fulfilled. This understanding also seems to be indicated in the GPAI Code of Practice, which refers more generally to the 'involvement' of the model (directly or indirectly) leading to a specified outcome, rather than an incident or malfunctioning of the model.[492]

151. On the other hand, it also seems reasonable to interpret the term 'incident' as entailing an additional limiting criterion – such that not every causal connection between a GPAI model and a specified outcome would be sufficient to trigger the reporting obligation.[493] Likewise, the Commission, in its draft guidance on the serious incident reporting obligation for providers of high-risk AI systems, states that an incident 'is a not planned/programmed deviation in the characteristics of performance'.[494] Under this – more narrow – interpretation, to better understand the notion of an 'incident' it is useful to look at several other EU instruments that contain a definition of that term.

152. A definition of the term 'incident' can be found in several other EU instruments that contain similar reporting obligations.[495] Looking at those instruments – notably the MDR, the NIS2 Directive and the DORA – one might be able to carefully draw some interpretative inspiration for the AI Act.[496] First, translated to the governance of GPAI models with systemic risk, the MDR indicates that an incident could be said to occur only where the characteristics or performance of the GPAI model

[488] Commission Guidelines (n 16) para 100.
[489] Also see Draft Guidance on Article 73 (n 420) para 7.
[490] Code of Practice, Safety and Security Chapter (n 9) recital (j); see, in detail, Section 2.1.3.6.4.
[491] See similarly Chatzipanagiotis (n 447) 6 ['the term 'incident' encompasses situations related to the development or operation of a high-risk AI system or general-purpose AI model irrespective of whether they involve a malfunction'].
[492] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.
[493] Similarly for the definition in Article 3(49), Susanne Wende, 'Art. 3 Begriffsbestimmungen' in David Bomhard, Fritz-Ulli Pieper & Susanne Wende (eds), *KI-VO: Verordnung über künstliche Intelligenz* (Deutscher Fachverlag, 2025) para 375.
[494] Draft Guidance on Article 73 (n 420) para 8.
[495] ibid.
[496] See the forthcoming chapter on Interpreting the AI Act through Systematic Analogies in this commentary.

with systemic risk deteriorate.[497] The fact that use-errors are explicitly mentioned in the MDR but not in the AI Act seems to be a knife that cuts both ways. One could argue that the fact that it is expressly mentioned in the MDR speaks against adopting this approach in the AI Act. On the other hand, one could argue that this is a general principle within reporting obligations throughout the EU legal order and it was just expressly mentioned for clarification in the MDR. Second, the NIS2 Directive's definition[498] likewise refers to a deterioration of certain characteristics of a service (or data). Third, the DORA[499] implies that only unplanned events are captured and that also a series of linked events can suffice for the definition to be applicable.[500] Similarly, the draft guidance on the reporting obligation under Article 73 lists some examples of incidents and malfunctions that largely feature unplanned events, namely misclassifications, significant drops in accuracy, temporary system downtime and unexpected system behaviour.[501]

153. At the same time, however, applying 'unplanned' as a delimiter to GPAI models appears more difficult because their behaviour is hard to predict and they exhibit no genuine intended purpose,[502] so there is, in a sense, no baseline against which deterioration could be measured. What also remains less clear under this approach is why – compared to the Commission's proposal of the AI Act[503] – the term 'malfunctioning' was added as an alternative element of the definition.[504] In the MDR, for example, a malfunction is merely a subcategory of an incident.[505] The added term in the AI Act likely primarily serves a clarifying function, for example, intended to make clear that not only events triggered by malicious actors (as the term 'incident' could be understood to imply) are meant to be captured.[506]

[497] The MDR defines an incident in article 2(64) as 'any malfunctioning or deterioration in the characteristics or performance of a device made available on the market, including use-error due to ergonomic features, as well as any inadequacy in the information supplied by the manufacturer and any undesirable side-effect'.

[498] The NIS2 Directive, in article 6(6), defines an incident as 'an event compromising the availability, authenticity, integrity or confidentiality of stored, transmitted or processed data or of the services offered by, or accessible via, network and information systems'.

[499] The DORA defines an 'ICT-related incident' in article 3(8) as 'a single event or a series of linked events unplanned by the financial entity that compromises the security of the network and information systems, and have an adverse impact on the availability, authenticity, integrity or confidentiality of data, or on the services provided by the financial entity'.

[500] Support for the view that an 'accumulation of smaller AI incidents could lead to a serious AI incident' can also be found in Karine Perset and Luis Aranda, 'Defining AI Incidents and Related Terms' (OECD 2024) OECD Artificial Intelligence Papers 18 <https://doi.org/10.1787/d1a8d965-en> accessed 20 September 2025, 12; see also Commission Implementing Regulation (EU) 2024/2690 of 17 October 2024 laying down rules for the application of Directive (EU) 2022/2555 as regards technical and methodological requirements of cybersecurity risk-management measures and further specification of the cases in which an incident is considered to be significant with regard to DNS service providers, TLD name registries, cloud computing service providers, data centre service providers, content delivery network providers, managed service providers, managed security service providers, providers of online marketplaces, online search engines, social networking services platforms and trust service providers [2024] OJ L 2690, art 4 ('Recurring incidents').

[501] ibid para 12

[502] See similarly Boine and Rolnick (n 271) 93.

[503] AI Act Proposal (n 457).

[504] See similarly Chatzipanagiotis (n 447) 5.

[505] See MDR, art 2(64).

[506] See similarly Draft Guidance on Article 73 (n 420) para 11: 'In practice, the distinction between "incident" and "malfunction" in the AI Act should not be understood as a strict distinction, but rather as an emphasis on the importance of malfunctions in the context of incident monitoring.'

2.1.3.2.1.6. Causal connection

154. For the reporting obligation under Article 55(1)(c) to be triggered, a causal connection must exist between the GPAI model with systemic risk concerned and one of the qualifying outcomes. The precise manner in which that causal connection is to be established, and between which specific elements it must exist, remains unclear.

155. If one follows the approach that the concept of 'serious incident' in Article 3(49) is to be treated as the structural basis for Article 55(1)(c), the *incident or malfunction* would need to be in a causal connection ('directly or indirectly leads to') to at least one of the outcomes specified in Article 3(49) (or to another situation in which a systemic risk has materially increased or materialised). The Commission also partly refers to this same wording in its GPAI Guidelines – stating that apart from aforementioned cybersecurity breaches, it considers 'a "serious incident" in the context of Chapter V AI Act as any incident or malfunctioning of a general-purpose AI model that directly or indirectly leads to any of the events listed in the corresponding definition for AI systems in Article 3(49), points (a) to (d), AI Act.'[507]

156. The GPAI Code of Practice requires signatories to report 'if the *involvement* of their model (directly or indirectly) led to' (emphasis added) the further specified outcomes.[508] Therefore, the Code of Practice seems to indicate that the causal connection need not exist between an incident or malfunctioning of the model and a specified outcome, but rather it suffices if the model is in some way causally involved. The outcome of this interpretation would be the same as under the previously discussed interpretation, if one does not understand the 'incident or malfunction' component of the definition to be a limiting criterion.[509]

157. Apart from that ambiguity with regard to the causal link between the model and a specified outcome, a different understanding and concept of the causality requirement seems to be implied both in the GPAI Code of Practice and in the Commission's GPAI Guidelines with regard to *serious cybersecurity breaches*. Measure 9.3(2) of the Safety and Security Chapter of the GPAI Code of Practice requires for the reporting obligation to be triggered that the involvement of the signatory's model 'led to ... a serious cybersecurity breach, including the (self-)exfiltration of model weights and cyberattacks'.[510] This formulation largely resembles the Commission's GPAI Guidelines, which provide that 'cybersecurity breaches related to the model or its physical infrastructure, including the (self-)exfiltration of model parameters and cyberattacks' are to be reported.[511] The latter formulation clarifies that it suffices that the model is somehow involved ('*related to* the model or its physical infrastructure', emphasis added). This implies that certain outcomes – especially those increasing the systemic risk posed by the model – do not necessarily need to be caused *by* the model but can also be happening *to* the model.

158. Regardless of whether one requires a link between the model's involvement or an incident or malfunction of the model and a qualifying outcome, and following the above, the interpretation of the causation concept embedded in the 'directly or indirectly leads to' formulation seems to become particularly relevant with regard to the outcomes enumerated in Article 3(49) as well as

[507] Commission Guidelines (n 16) para 100.
[508] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.
[509] See Section 2.1.3.2.1.5.
[510] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3 according to which also a suspicion of 'reasonably likelihood' can suffice.
[511] Commission Guidelines (n 16) para 100.

materialisations of systemic risk. This causal relationship will in many cases be an indirect one, as a model will rarely be capable of bringing about the outcomes referred to in Article 3(49) without prior integration into an AI system. In its draft guidance on the reporting obligation under Article 73, the Commission defines an incident or malfunction as directly or indirectly causal 'if, without it, the harm in its concrete form would not have occurred (or reasonably likely respectively more probable not to have occurred).'[512] The draft guidance further clarifies that secondary effects suffice, possibly even multiple steps later in the chain – an example being an AI system that provides incorrect medical imaging, leading to an incorrect diagnosis, ultimately leading to harm to a patient.[513] Finally, the Commission intends to limit causations to cases of system use corresponding to its intended purpose[514] or reasonably foreseeable misuse.[515]

159. This definition set out in the draft guidance on Article 73 comprises two components – a factual one and a normative one – of which, however, only one is directly transposable in the GPAI model context. What may be directly transposable is the *factual component*,[516] understood in the sense of the *conditio sine qua non* formula familiar to most EU legal systems.[517] The relevant question to be asked is therefore whether the harm in its concrete form – so, especially the outcomes mentioned in Article 3(49) – would not have occurred (or reasonably likely respectively more probable not to have occurred) had a different model been used instead or had the model not experienced an incident or a malfunction.[518]

160. The second component, however, appears to be a *normative* one.[519] The limitation in the draft guidance on Article 73 to cases in which the AI system is used in accordance with its *intended purpose* or *reasonably foreseeable misuse* seems to rest on the normative premise that the risk management obligations imposed on providers of high-risk AI systems under Article 9 are themselves confined to risks arising from use in accordance with the system's intended purpose and under conditions of reasonably foreseeable misuse.[520] Providers might therefore be subject to reporting obligations under Article 73(1) only in relation to materialisations of the risks they are required to assess and mitigate under Article 9. That logic cannot be directly transposed to GPAI models. Given their significant generality and ability to form the basis for a wide range of downstream systems and applications,[521] the concepts of intended purpose and reasonably foreseeable misuse might not map neatly onto GPAI models.[522] The risks for which a provider of a

[512] Draft Guidance on Article 73 (n 420) para 13.

[513] ibid.

[514] Defined in article 3(12) as 'the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation'.

[515] Draft Guidance on Article 73 (n 420) para 14; Chatzipanagiotis (n 447) 9 argues that a restriction to intended purpose and reasonably foreseeable misuse is misplaced in a reporting context but rather 'makes sense in a liability context'. This is because, according to him, 'reporting harms due to unforeseeable misuse provides valuable safety insights'. For the definition of reasonably foreseeable misuse see article 3(13) ['the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems, including other AI systems'].

[516] Jens Kleinschmidt, 'Causation' in Jürgen Basedow, Klaus Hopt and Reinhard Zimmermann (eds), *The Max Planck Encyclopedia of European Private Law* (Oxford University Press 2012) <https://max-eup2012.mpipriv.de/index.php/Causation> accessed 19 May 2026, 156.

[517] ibid.

[518] On the corresponding discussion of the scope and function of this element, see Section 2.1.3.2.1.5.

[519] See Kleinschmidt (n 516).

[520] See AI Act, art 9(2).

[521] See para 80.

[522] See Section 2.1.2.1.2.

GPAI model with systemic risk bears responsibility are, however, defined by Article 55(1)(a), (b) and (d), which perform an at least similar structural function as Article 9(2)(a) and (b) in the high-risk AI system framework.[523] Consistent with that rationale, the reporting obligation under Article 55(1)(c) should not extend to every serious incident that is factually connected to the development or use of a GPAI model. Rather, it should be limited to incidents that materialise within the possible systemic risks that the provider was required to assess and mitigate under Articles 55(1)(a), (b) and (d).[524] This conclusion is supported by the wording of Recital 115, which conditions the reporting obligation on situations in which, '*despite efforts to identify and prevent risks related to a general-purpose AI model that may present systemic risks*, the development or use of the model causes a serious incident' (emphasis added). Additionally, one might argue that a purposive reading suggests that incidents that do not fulfil this normative criterion do not fall within the purpose and objective of Article 55(1)(c) to inform the systemic risk identification, assessment and mitigation pursuant to Article 55(1)(a) and (b).

2.1.3.2.2. Specified outcome

161. According to the definition in Article 3(49), the obligation to report a serious incident applies if one of the specified outcomes actually occurs.

2.1.3.2.2.1. Death or serious harm

162. The first outcome covered is the death of a person or serious harm to a person's health. In most cases, a person's death (as well as the further defined harms in points (b) to (d)) will likely be directly caused by an AI system rather than a GPAI model with systemic risk – for example, where the system's output triggers a mechanical reaction.[525] However, cases cannot be ruled out in which it can be demonstrated that it was not the AI system but rather (or additionally) the underlying GPAI model with systemic risk that was at least indirectly responsible for a person's death.[526]

163. The AI Act does not expressly define what exactly constitutes *serious harm* in the sense of Article 3(49). To begin with, the open wording suggests that not only physical harm but also mental health harm is likely to be covered.[527] Some authors suggest that the parameters for determining whether harm qualifies as serious could include the intensity, extent and severity or duration of the harm as well as the consequences for the person's daily life.[528] To ensure at least some normative anchor, it also seems possible to draw guidance from Article 2(58) MDR, which provides a more detailed definition of the term 'serious adverse event' and bears certain resemblances with the AI Acts definition in Article 3(49)(a)[529]. According to that provision, a serious adverse event is any adverse event that led to the death or a 'serious deterioration in the health of the subject' resulting in further

[523] See Section 2.1.2.1.
[524] See Section 2.1.2.1. on the meaning of 'possible systemic risks at Union level'.
[525] Christiane Wendehorst, 'Art. 3 Begriffsbestimmungen' in Mario Martini and Christiane Wendehorst (eds), *KI-VO: Verordnung über Künstliche Intelligenz* (2nd edn, C.H. Beck, 2026) para 365.
[526] Opposing opinion seems to be found at Finck (n 36) para 6.118 ['This criterion is much more difficult to apply in relation to models than systems as systems may cause such harm due to their physicality, but models cannot'].
[527] Lukas Feiler and Beat König, 'Article 3 Definitions' in Ceyhun Necati Pehlivan, Nikolaus Forgó and Peggy Valcke (eds), *The EU Artificial Intelligence (AI) Act* (Wolters Kluwer, 2025) 76; this is also supported by Code of Practice, Safety and Security Chapter (n 9) Measure 9.3, 27 ['serious harm to a person's health (mental and/or physical)'] as well as app 1.1. listing mental health risks; also see Perset and Aranda (n 500) 11.
[528] Jonathan Kirschke-Biller and Anna Lena Füllsack, 'Artikel 3 Begriffsbestimmungen' in Jens Schefzig and Robert Kilian (eds), *Beck'scher Online-Kommentar KI-Recht* (5th edn, C.H. Beck 2025) para 572.
[529] Also see the forthcoming chapter on Interpreting the AI Act through Systematic Analogies in this work.

defined outcomes. Although this definition concerns the consequences of a serious health deterioration, it nevertheless seems plausible to qualify at least the listed outcomes as serious harm to a person's health within the meaning of the AI Act. These outcomes are: '(i) life-threatening illness or injury, (ii) permanent impairment of a body structure or body function, (iii) hospitalisation or prolongation of patient hospitalisation, (iv) medical or surgical intervention to prevent life-threatening illness or injury or permanent impairment to a body structure or a body function (v) chronic distress'.[530]

2.1.3.2.2.2. Disruption of critical infrastructure

164. Article 3(49) also defines 'serious incidents' to include 'serious and irreversible disruption of the management or operation of critical infrastructure'.

165. Article 3(62) defines the term *critical infrastructure* with reference to Article 2(4) of the Critical Entities Resilience Directive ("CERD")[531] which defines 'critical infrastructure' as 'an asset, a facility, equipment, a network or a system, or a part of an asset, a facility, equipment, a network or a system, which is necessary for the provision of an essential service', with an essential service defined as 'a service which is crucial for the maintenance of vital societal functions, economic activities, public health and safety, or the environment' (Article 2(5)).

166. The notion of 'serious and irreversible disruption' in Article 3(49) remains unclear, however. Particular attention must be paid to the fact that both elements must be fulfilled cumulatively ('and'). The question of when a disruption is serious is answered inconsistently in legal literature. Some authors suggest that the decisive factor is whether the effects of the disruption are or could be severe.[532] Others propose looking at factors such as intensity, magnitude and extent, duration and reversibility, impairment of the supply situation (types of goods affected, number of persons affected), threats to public safety and order, and societal impacts.[533] Although both approaches seem plausible, they both lack a clear normative anchor. Some authors therefore recommend drawing on Article 15(1) CERD, which provides a non-exhaustive list of parameters that might be used to assess whether a disruption of the provision of an essential service was significant.[534] This approach appears most persuasive given the AI Act's explicit reference to the CERD.[535] Article 15(1) CERD lists: '(a) the number and proportion of users affected by the disruption; (b) the duration of the disruption; (c) the geographical area affected by the disruption, taking into account whether the area is geographically isolated'.

167. Additionally, in its draft guidance on the reporting obligation pursuant to Article 73, the Commission gives some examples of disruptions that should be considered serious. Those are (i) 'The disruption might result in an imminent threat to life or the physical safety of a person, including

[530] Also see Draft Guidance on Article 73 (n 420) drawing from Medical Device Coordination Group, 'MDCG 2023-3 Rev.2 - Q&A on Vigilance Terms and Concepts as Outlined in the Regulation (EU) 2017/745 and Regulation (EU) 2017/746' (Medical Device Coordination Group 2023) MDCG 2023-3 Rev. 2 <https://health.ec.europa.eu/document/download/af1433fd-ed64-4c53-abc7-612a7f16f976_en?filename=mdcg_2023-3_en.pdf> accessed 15 May 2026, 10.

[531] Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities [2022] OJ L 333/164 ("CERD").

[532] Wendehorst (n 525) para 366.

[533] Kirschke-Biller and Füllsack (n 528) para 574.

[534] Feiler and König (n 527) 76.

[535] On the other hand, it could be argued that, in the context of creating the cross-reference, a reference to article 15(1) could also have been included, had it been intended to be covered.

through serious harm to the provision of basic supplies to the population or the exercise of the core function of the State', as well as (ii) 'Destruction of key infrastructure' and (iii) 'Disruption in social or economic activities'.

168. It is important to note that the parameters listed in Article 15(1) CERD address only the significance of the disruption in the *provision* of an essential service.[536] Unlike the CERD, however, the definition in Article 3(49) also covers disruptions in the *management* of critical infrastructure.[537] Where such disruptions likewise affect the operation of the infrastructure, the parameters mentioned above might nonetheless be applied. What remains unclear, however, is how to assess the serious disruption of the management of critical infrastructure if it does not also impact the operation. One possible approach would be to consider management disruptions as serious if – absent intervention by the provider – they could subsequently lead to a serious disruption in the operation of the critical infrastructure.[538]

169. Moreover, the disruption must not only be serious, but also irreversible.[539] The concept of irreversibility is not reflected in current EU legislation on critical infrastructure, and it is also unclear how it differs from destruction.[540] Furthermore, it remains unclear why providers of GPAI models with systemic risk are, pursuant to Article 55(1)(c), required to keep track of, document and report possible corrective measures with respect to disruptions that are, by definition, irreversible anyway. One possible approach would be that the assessment of the seriousness already encompasses considerations of reversibility. This, however, is countered by the wording,[541] which clearly presupposes the existence of disruptions that are serious but not irreversible. To give the term an independent meaning, it might be understood as to denote a disruption that cannot be remedied through ordinary maintenance work for which providers are expected to provide at all times.[542] Another line of reasoning would hold that any disruption, once it has occurred, is never fully reversible and that the focus should therefore be on the consequences. An illustrative example would be a large-scale power grid failure leading to a prolonged blackout across an entire Member State. Such an event would probably qualify as a serious disruption. It would, however, be considered irreversible under the latter interpretation only if, for example, as a result of the blackout, persons in a nearby hospital were to die. In other words: a disruption might be regarded as irreversible when, despite its eventual rectification, further negative effects remain.

170. The latter view also appears to be shared by the Commission in its draft guidance on Article 73, which states that, to qualify as a serious disruption, the following aspects should be taken into account: (i) 'The disruption requires rebuilding of physical infrastructure or destroys specialized equipment which is not readily available' as well as (ii) 'Contamination of water, soil or air' and (iii) 'Loss or corruption of essential records – such as patient data, civil registries, or financial transactions – that cannot be reliably restored or reconstructed'. Also mentioned are (iv) 'Permanent disablement of a critical node, such as a rail junction, power substation, or landing station, that cannot be repaired or replaced without years-long lead times' and (v) 'Loss of a space-based asset (e.g. Global Navigation Satellite System or communications satellite) whose destruction vacates its

[536] Similar to Wendehorst (n 525) para 366.
[537] Which is alien to EU critical infrastructure legislation, see Feiler and König (n 527) 76.
[538] See similarly ibid.
[539] AI Act, art 3(49)(b).
[540] Feiler and König (n 527) 76.
[541] See AI Act, art 3(49)(b): 'a serious *and* irreversible disruption' (emphasis added).
[542] Wendehorst (n 525) para 366; similarly Wende, 'Art 3' (n 493) para 380; also see Chatzipanagiotis (n 447) 7.

orbital slot or frequency and cannot be replaced without an extended replacement procedure that typically lasts years.'[543]

2.1.3.2.2.3. Infringement of EU law to protect fundamental rights

171. Furthermore, covered by Article 3(49) is 'the infringement of obligations under Union law intended to protect fundamental rights'. In the European Parliament's earlier proposal,[544] a narrower version of the *serious incident* definition covered only 'a breach of fundamental rights protected under Union law'.[545] The enacted definition is, by its wording, considerably wider and risks becoming amorphous, as it captures not only infringements of fundamental rights themselves but of all 'obligations' under Union law intended to protect fundamental rights.[546]

172. This might lead to the challenge[547] that obligations under, in particular, the GDPR qualify, with the result that every violation of GDPR obligations, for example every personal data breach within the meaning of Article 4(12) GDPR, could, in theory, constitute a serious incident.[548] This would produce the contradictory outcome that a personal data breach would always trigger the reporting obligation under Article 55(1)(c), while it might, by contrast, not even require notification under the GDPR itself because of the exception in Article 33(1) GDPR (i.e. 'unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons').[549]

173. More generally, it makes little sense to capture only serious health injuries or, for example, only serious and irreversible disruptions of critical infrastructure on the one hand, while at the same time, on the other hand, allowing *any* violation of ordinary statutory provisions to suffice elsewhere – provided that they are 'intended to protect fundamental rights' and relate to (another) fundamental right. To avoid such contradictory outcomes, either a certain degree of severity of the infringement should be required[550] or the term 'intended to protect' should be interpreted narrowly.[551] In assessing the severity of the infringement, it would be reasonable to take guidance from the other categories in Article 3(49), not least because it could create an inconsistency if certain obligations were recognised as aiming to protect 'health' or 'property' as fundamental rights, and breaches of those obligations would therefore fall within Article 3(49)(c) yet would fail to meet the higher thresholds required in Article 3(49)(a) ('death' and 'serious harm') and (d) ('serious harm'). After all, an overly broad interpretation would also run counter to the very purpose of the reporting obligation, as it would force the AI Office to expend resources unnecessarily and could impair its ability to respond appropriately to genuinely serious incidents. Such an information overload should be avoided through a restrictive interpretation.[552]

174. In its draft guidance for Article 73, the Commission gives some examples for infringements covered by the definition. These include (i) 'An AI based recruitment system excludes candidates based on

[543] Draft Guidance on Article 73 (n 420) para 21.
[544] Parliament Amendments (n 99) art 3(1)(44).
[545] ibid.
[546] Also see Wendehorst (n 525) para 367.
[547] Chatzipanagiotis (n 447) 8 notes that it also remains unclear how an infringement of such obligations should be established and whether formal determination would be required.
[548] Feiler and König (n 527) 77; Wendehorst (n 525) para 367.
[549] See also similarly Finck (n 36) para 2.341.
[550] Wendehorst (n 525) para 367; also see Finck (n 36) para 2.341 and Chatzipanagiotis (n 447) 8.
[551] Also see the Draft Guidance on Article 73 (n 420) paras 24, 25.
[552] ibid; also see Bommasani and others (n 422) 35 noting that '[p]ast efforts in other domains have sometimes suffered from overly inclusive reporting criteria, which risk drowning signal in noise'.

ethnicity or gender' as well as (ii) 'A credit scoring system excludes certain categories of persons, such as those having a name from a certain region or living in certain neighbourhoods' and (iii) 'A biometric identification system frequently wrongly identifies people of different ethical background.'[553]

2.1.3.2.2.4. Serious harm to property or environment

175. Lastly, Article 3(49) covers 'serious harm to property or the environment'.

176. The AI Act provides no further guidance on what is to be understood under this term. Some authors propose considering 'the amount of damage' for property damage and 'the financial cost of removal, but also, for example, [...] the danger to humans, animals, plants and their habitat' for environmental damages.[554] According to others, relevant factors include, for example, the value of the affected goods, the irreversibility of the damage, the number of persons affected (including future generations for environmental damage), and the extent of the impact on natural ecosystems.[555] Other authors argue that one may look, *inter alia*, at the intensity, magnitude, and extent of the damage; the goods affected, the duration and reversibility of the damage, the number of persons affected as well as the broader societal consequences.[556] To assess the seriousness of environmental harm, one could also draw on the ISO 14001 standard for environmental management systems, which contains categories for both 'minor environmental impact' and 'major environmental impact'.[557]

177. The Commission's draft guidance on Article 73 further sets out the following parameters for assessing the seriousness of harm to property:[558] (i) 'The economic impact, including cost of repair or replacement. The damage to property is deemed to be serious if the damage or destruction impairs the intended usability or substance of the property to such an extent that it can no longer be used for its intended purpose. The amount of damage, the cost of repair or the reduction in value are not decisive in this respect, but should in any case exceed 5% of the purchase price'; (ii) 'The cultural, or historical significance of the property'; (iii) 'The extent to which the property loss or damage affects the livelihood or quality of life of individuals or communities.'; (iv) 'The permanence of the damage, including whether the property can be restored to its former state.' as well as (v) 'The ripple effects of the damage, such as its impact on surrounding areas or related operations.'[559]

178. With respect to the seriousness of harm to the environment, the draft guidance lists the following parameters: (i) 'the baseline condition of the affected environment'; (ii) 'whether the damage is long-lasting, medium-term or short-term'; (iii) 'the extent of the damage' as well as (iv) 'the reversibility of the damage.' Examples mentioned are the '[c]ontamination of environmental resources' and '[d]isruption of natural ecosystems'.[560]

---

[553] ibid para 26.
[554] Feiler and König (n 527) 77.
[555] Wendehorst (n 525) para 368.
[556] Kirschke-Biller and Füllsack (n 528) para 578.
[557] ISO, 'Environmental Management Systems — Requirements with Guidance for Use' (ISO 2015) ISO 14001:2015 <https://www.iso.org/obp/ui/#iso:std:iso:14001:ed-3:v1:en> accessed 19 May 2026.
[558] Draft Guidance on Article 73 (n 420).
[559] ibid para 27.
[560] ibid paras 29, 30.

2.1.3.2.2.5. Increase or materialisation of systemic risks

179. As mentioned above, there are arguments to assume that the concept of a serious incident needs to be understood more broadly with regard to Article 55(1)(c), compared to Article 3(49).[561] This line of argumentation is built upon the statement in the Commission's GPAI Guidelines that the AI Office considers Article 55(1)(c) to cover 'serious cybersecurity breaches related to the model or its physical infrastructure, including the (self-)exfiltration of model parameters and cyberattacks due to their possible implications for the obligations provided for in Article 55(1), points (b) and (d)'. Additionally, as the GPAI Code of Practice emphasises,[562] the main purpose of the obligations laid down in Article 55 is to assess and mitigate systemic risk. Since not only serious cybersecurity breaches but also other increases and materialisations of systemic risks stemming from the model have implications for the provider's obligations at least under Article 55(1)(b),[563] it makes sense to have the obligation under Article 55(c) encompass not only the outcomes mentioned in Article 3(49) but also material increases and materialisations of the systemic risks stemming from the model identified by the provider pursuant to Commitment 2 of the Safety and Security Chapter of the GPAI Code of Practice.

180. Leaving aside the question of whether the inclusion of 'serious cybersecurity breaches' in the Commission's GPAI Guidelines and the GPAI Code of Practice support a broader reading of the serious incident concept in Article 55(1)(c), it remains to be determined which events precisely qualify as serious cybersecurity breaches. The GPAI Guidelines and the GPAI Code of Practice offer some clarification insofar as both indicate that the term is to be understood as 'including the (self-)exfiltration of model weights and cyberattacks'.[564] Beyond these examples, however, the contours of the concept remain rather unclear. Pending further guidance, the assessment should be informed by whether the systemic risk posed by the model has been materially increased by the breach.[565] This will arguably be the case primarily where model weights have been exfiltrated. By contrast, not every unsophisticated cyberattack targeting the model should trigger the reporting obligation, so as to avoid an information overload at the AI Office. A further question concerns the interplay between the obligation to report serious cybersecurity breaches and the exemption in Commitment 6 of the Safety and Security Chapter of the GPAI Code of Practice. This states that a model is exempt from Commitment 6 where a more capable model's parameters are available for download. Following this logic – an exemption from security measures where a more capable model is freely available – one might argue that a cybersecurity breach cannot be considered 'serious' where the model concerned already falls within the exemption under Commitment 6. Overall, further regulatory guidance on the precise definition of the term appears desirable.

181. If one were to follow the broader reading of Article 55(1)(c), which includes not only serious cybersecurity breaches as reportable incidents, but also other instances of material increases in systemic risk stemming from the model,[566] it remains open as to when such an increase triggers the reporting obligation. A parallel exists in the obligation of signatories to update their Safety and Security Model Reports pursuant to Measure 7.6 of the Safety and Security Chapter. According to

[561] See Section 2.1.3.2.1.3.
[562] Code of Practice, Safety and Security Chapter (n 9) recital (i).
[563] See Section 2.1.2.1.3.
[564] Commission Guidelines (n 16) para 100 and Code of Practice, Safety and Security Chapter (n 9) Measure 9.3(2).
[565] Also see similarly Chatzipanagiotis (n 447) 8 speaking of cybersecurity breaches significantly affecting cybersecurity risks mentioned in recital 115.
[566] See Section 2.1.3.2.1.3.

that measure, signatories must update their Safety and Security Model Report 'if they have reasonable grounds to believe that the justification for why the systemic risks stemming from the model are acceptable ... has been materially undermined'.[567] Under this broader reading, both signatories and non-signatories should be able to orientate themselves by reference to this criterion.[568]

182. It must be acknowledged, however, that this reading sits in some tension with the earlier conclusion that near misses likely fall outside Article 55(1)(c)'s scope.[569] A near miss may, in a given case, be precisely the kind of event that evidences a material increase in systemic risk. One could argue, however, that this overlap is narrower than it first appears. That is because not every near miss will indicate that the model's systemic risk has materially increased: A near miss in which the provider's risk mitigations worked as intended – for example a successfully repelled cyberattack – may confirm the risk-acceptability justification rather than indicating that is has been undermined.[570] The aforementioned tension can therefore be mostly resolved if only those near misses that reveal risk mitigations to be weaker than assumed or surface capabilities that were not previously accounted for qualify as material increases in the systemic risk posed by the model.

183. Additionally, and against this approach, however, one might ask what difference would then remain between the obligation to update and provide the Safety and Security Model Report to the AI Office and the reporting of material increases in systemic risk as serious incidents under Article 55(1)(c). An important difference lies in the fact that the update of a Safety and Security Model Report must be completed within a 'reasonable amount of time' after a signatory has identified the necessity of an update, followed by a further period of five business days within which the updated report must be sent to the AI Office.[571] Given the extensive information required in such a report – see Measures 7.1 to 7.5 of the Code of Practice – it will accordingly take some time before the AI Office receives the relevant information. Treating material increases in systemic risk stemming from the model as serious incidents under Article 55(1)(c) would therefore have the advantage that information could reach the AI Office earlier, without requiring a full Safety and Security Model Report update, thereby enabling the AI Office to react more promptly where necessary. Depending on how the terms 'incident' and 'malfunctioning' are interpreted, it must further be noted that the obligation to provide an updated Safety and Security Model Report also encompasses deliberate changes made to the model by the provider,[572] whereas the obligation under Article 55(1)(c) may not extend to such changes.[573]

2.1.3.2.3. Relevant information

184. It is important to note that providers of GPAI models with systemic risk are required to report, keep track of, and document not the serious incident as such, but 'relevant information' about it.

---

[567] Code of Practice, Safety and Security Chapter (n 9) Measure 7.6(1)–(5) enumerates as examples the materialisation of a condition listed under Measure 7.2(2); a material change in the model's capabilities, propensities or affordances; a material change in the model's use or integration into AI systems; the occurrence of serious incidents or near misses involving the model or a comparable model; and developments that materially undermine the external validity of prior evaluations or otherwise indicate that the systemic risk assessment is inaccurate.

[568] See Section 2.3.; See the commentary on Article 56 in this work.

[569] See Section 2.1.3.2.1.4.

[570] This does not necessarily mean that they can not impact the risk management under article 55(1)(a) and (b).

[571] Code of Practice, Safety and Security Chapter (n 9) Measures 7.6 and 7.7.

[572] ibid Measure 7.6.

[573] See Section 2.1.3.2.1.5.

The Commission has published a reporting template for serious incidents involving GPAI models with systemic risk in November 2025, which is intended to 'serve as a means to demonstrate compliance with Article 55(1), point (c), of the AI Act as part of Commitment 9 of the Safety and Security Chapter of the General-Purpose AI Code of Practice'.[574]

185. The reporting template and the GPAI Code of Practice provide a helpful overview of the types of information that can be considered relevant in this context. Reports must cover: (i) the dates of the incident or best approximations thereof; (ii) the resulting harm and those affected; (iii) the chain of events leading to the incident; (iv) the model involved; (v) available material documenting the model's involvement; (vi) the signatory's response or intended response; (vii) the signatory's recommendations to the AI Office and, where applicable, national competent authorities; (viii) a root cause analysis, including the model's relevant outputs, contributing factors, and any failures or circumventions of systemic risk mitigations; and (ix) any patterns from post-market monitoring reasonably connected to the incident, including data on near misses.[575]

186. As noted above, near misses themselves arguably do not qualify as a serious incident in the sense of Article 55(1)(c) and therefore may not trigger the reporting obligation.[576] Since they are part of the relevant information that providers of GPAI models with systemic risk need to keep track of and document, however, the AI Office will have the possibility to request this information pursuant to Article 91(1). In principle, providers of GPAI models with systemic risk are therefore likely not required to report near misses proactively. An exception may arise where a near miss constitutes a 'reasonable ground' for updating the provider's Safety and Security Framework.[577] Any such update must be reported to the AI Office within five business days.[578]

2.1.3.3. Keeping track of relevant information

187. To be able to document and report serious incidents, providers of GPAI models with systemic risk will need to keep track of relevant information about serious incidents. The AI Act does not specify how providers of GPAI models with systemic risk should fulfil their obligation to keep track of relevant information. The Safety and Security Chapter of the GPAI Code of Practice, however, lists a number of exemplary measures that providers of GPAI models with systemic risk can adopt in order to comply with the obligation.[579]

188. Signatories may comply with this obligation through a range of measures. On the input side, these measures include collecting end-user feedback, providing reporting channels (including anonymous ones) and incident reporting forms, and offering bug bounties. Broader engagement mechanisms encompass community-driven model evaluations and public leaderboards; frequent dialogues with affected stakeholders; and collaboration with academia, civil society, regulators, and independent researchers in support of the scientific study of the model's capabilities, propensities, affordances, and effects. On the monitoring side, signatories may monitor software repositories, known malware,

[574] European Commission, 'AI Act: Commission Publishes a Reporting Template for Serious Incidents Involving General-Purpose AI Models with Systemic Risk' (*European Commission*, 4 November 2025) <https://digital-strategy.ec.europa.eu/en/library/ai-act-commission-publishes-reporting-template-serious-incidents-involving-general-purpose-ai> accessed 12 February 2026 ("Reporting Template").
[575] Code of Practice, Safety and Security Chapter (n 9) Measure 9.2.
[576] See Section 2.1.3.2.1.4.
[577] See Code of Practice, Safety and Security Chapter (n 9) Measure 1.3.
[578] ibid Measure 1.4.
[579] See ibid Measure 9.1 pointing to the exemplary methods in Measure 3.5.

public forums, and/or social media for patterns of use; implement privacy-preserving logging and metadata analysis of the model's inputs and outputs using, for example, watermarks, metadata, and/or state-of-the-art provenance techniques; collect relevant information about breaches of the model's use restrictions and any subsequent incidents; and monitor aspects of the model that are relevant to assessing and mitigating systemic risk but are not transparent to third parties, such as hidden chains-of-thought in models whose parameters are not publicly available for download.[580]

189. According to Measure 9.1 of the Safety and Security Chapter of the GPAI Code of Practice, signatories are required to additionally 'review other sources of information (such as police and media reports, posts on social media, research papers, and incident databases)' and also 'facilitate the reporting of relevant information about serious incidents by downstream modifiers, downstream providers, users and other third parties' by informing them of direct reporting channels.

190. Lastly, under Measure 8.3 of the Safety and Security Chapter of the GPAI Code of Practice, signatories commit to promote a healthy risk culture – one indicator for that being that internal reporting channels are actively used and reports are acted upon appropriately.

2.1.3.4. Documenting relevant information

191. The AI Act does not specify how providers of GPAI models with systemic risk must document the relevant information. Providers should at least document the information mentioned in Measure 9.2 of the Safety and Security Chapter of the GPAI Code of Practice. It is not clear whether providers should document more than what they share in their report to the AI Office. Since Measure 9.2 sets out only a minimum standard ('at least the following information'), it may be advisable for providers to document more than the required minimum to the extent that additional relevant information concerning serious incidents could also inform the assessment of systemic risk stemming from the model.[581] Whether the documentation of additional information is appropriate will have to be determined case by case.[582]

192. The GPAI Code of Practice recommends that signatories keep their documentation for five years after the date of documentation or the date of the serious incident, whichever is later.[583]

2.1.3.5. Possible corrective measures

193. Providers of GPAI models with systemic risk must keep track of, document and report not only relevant information on serious incidents, but also possible corrective measures. The term is not further defined. The GPAI Code of Practice indicates a broad understanding encompassing all measures aimed to rectify the harm.[584] Drawing on Article 2(67) MDR, it can be understood to mean any possible action that can be taken to eliminate the cause of the serious accident as well as the incident itself and its consequences. Possible corrective measures will often include changes to the model itself as well as notices to downstream AI system providers.[585] Additionally, the information gathered on a serious incident as well as the corrective measures taken will inform the provider's systemic risk estimation, which in turn influences what (further) safety mitigations pursuant to Article

[580] ibid.
[581] ibid Measure 1.2(1)(c).
[582] Beurskens (n 20) para 7.
[583] Code of Practice, Safety and Security Chapter (n 9) Measure 9.4.
[584] See ibid Glossary definition of 'resolved'.
[585] Beurksens (n 20) para 7.

55(1)(b) should be introduced.[586] It can therefore be the case that corrective measures adopted *ex post* in response to a serious incident become safety mitigations *ex ante* for compliance with Article 55(1)(b).

2.1.3.6. Reporting relevant information and possible corrective measures

194. Providers of GPAI models with systemic risk must 'report, without undue delay, to the AI Office and, as appropriate, the national competent authorities'.[587] This section will deal with each element of the definition in turn.

2.1.3.6.1. Without undue delay

195. Unlike Article 73, Article 55(1)(c) contains no further specifications or differentiated timelines for reporting different types of serious incidents. Instead, the provision refers only to the indeterminate standard of 'without undue delay'. Accordingly, an individual assessment must, in principle, be made in each specific case.

196. Some authors suggest that guidance can be drawn from the timelines mentioned in Article 73, as both reporting obligations pursue similar objectives.[588] According to those timelines, serious incidents should, in principle, be reported 'immediately after the provider has established a causal link between the AI system and the serious incident or the reasonable likelihood of such a link, and, in any event, not later than 15 days after the provider or, where applicable, the deployer, becomes aware of the serious incident',[589] while also taking into account the severity of the incident.[590] The latter point is further specified in the subsequent paragraphs of Article 73. So-called widespread infringements – as defined in Article 3(61) – as well as serious incidents as defined in Article 3(49)(b) (that is, serious and irreversible disruptions of the management or operation of critical infrastructure) must be reported immediately, and not later than two days after the provider (or deployer) becomes aware of the incident).[591] In the event of the death of a person, the report must be provided immediately, and no later than ten days after the date on which the provider (or deployer) becomes aware of the incident.[592]

197. Although this approach offers the advantage of giving providers clear timelines to adhere to, it nonetheless appears unconvincing at first glance to blindly transfer Article 73's timelines to Article 55(1)(c).[593] Although the reporting obligations in Article 73 and Article 55(1)(c) are similar, and both rely on the same or at least a similar concept of 'serious incident' as a key element, the Union legislature would likely have incorporated the Article 73 timelines directly or by explicit cross-reference if they had wanted to apply them to Article 55(1)(c).

198. Nevertheless, the GPAI Code of Practice works with similar graduated timelines and also further distinguishes between different types of reports – an initial report, an intermediate report and a final

[586] See Code of Practice, Safety and Security Chapter (n 9) Measure 3.4.
[587] AI Act, art 55(1)(c).
[588] Bernsteiner and Schmitt, 'Art 55' (n 24) para 18.
[589] AI Act, art 73(2), first subparagraph.
[590] AI Act, art 73(2), second subparagraph.
[591] AI Act, art 73(3).
[592] AI Act, art 73(4).
[593] See the forthcoming chapter on Interpreting the AI Act through Systematic Analogies in this commentary.

report.[594] This structure closely resembles the distinction found in the NIS2 Directive (Article 23(4)) but is not explicitly found in Article 55(1)(c). According to the GPAI Code of Practice, an initial report[595] is to be provided by the signatories at the following times 'if the involvement of their model (directly or indirectly) led to:'[596]

'*a serious and irreversible disruption of the management or operation of critical infrastructure, or if the Signatories establish or suspect with reasonable likelihood such a causal relationship between their model and the disruption, not later than two days after the Signatories become aware of the involvement of their model in the incident*;'[597]

'*a serious cybersecurity breach, including the (self-)exfiltration of model weights and cyberattacks, or if the Signatories establish or suspect with reasonable likelihood such a causal relationship between their model and the breach, not later than five days after the Signatories become aware of the involvement of their model in the incident*;'[598]

'*a death of a person, or if the Signatories establish or suspect with reasonable likelihood such a causal relationship between their model and the death, not later than 10 days after the Signatories become aware of the involvement of their model in the incident*;'[599]

'*serious harm to a person's health (mental and/or physical), an infringement of obligations under Union law intended to protect fundamental rights, and/or serious harm to property or the environment, or if the Signatories establish or suspect with reasonable likelihood such a causal relationship between their model and the harms or infringements, not later than 15 days after the Signatories become aware of the involvement of their model in the incident.*'[600]

199. According to the GPAI Code of Practice, signatories are then required – in cases of unresolved serious incidents – to update the information provided and add the further information required[601] in an intermediate report.[602] Such an intermediate report shall be provided every four weeks after the initial report.[603] The final report must then be provided not later than 60 days after the serious incident has been resolved.[604] This final report should then cover all the information discussed above.[605]

200. Even though, as noted above, one should not blindly transfer the timelines given in Article 73, there are compelling reasons in favour of a graduated approach as in the GPAI Code of Practice. Given the purpose of the obligation – to enable a coordinated response to serious incidents by the AI Office and providers of GPAI models with systemic risk and thereby secure the situation after an

[594] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.
[595] Containing the information listed in points (1) to (7) of Code of Practice, Safety and Security Chapter (n 9) Measure 9.2.
[596] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.
[597] ibid.
[598] ibid.
[599] ibid.
[600] ibid.
[601] That is, the information listed in Code of Practice, Safety and Security Chapter (n 9) Measure 9.2(8) and (9).
[602] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.
[603] ibid.
[604] ibid.
[605] See Section 2.1.3.3.2.

incident, restore capacity to act, and prevent further harm – it indeed makes sense to report certain pieces of information to the AI Office at different points in time as soon as they become available, rather than waiting until all relevant information has been gathered. The AI Office can already work with the information contained in the initial report and, where appropriate, take first measures in response to the incident.[606] Moreover, the wording of the provision can be read accordingly such that 'without undue delay' refers not the reports as such but to the (respective) relevant information. It can thus reasonably be understood to mean that different timelines may be appropriate for different pieces of information. In that case, it would not be problematic that a graduated approach is not expressly set out, as it is in Article 73 AI Act or Article 23(4) NIS2 Directive.

201. A remaining difficulty concerns the point in time from which the assessment must begin as to whether a report has been made 'without undue delay'. Possible reference points include, on the one hand, the moment of first suspicion, or the point at which the provider is sufficiently certain (depending on the interpretation followed above),[607] that the involvement or an incident or malfunction of its model (directly or indirectly) caused one of the outcomes specified in Article 3(49) or a material increase or the materialisation of another systemic risk. On the other hand, it could be argued that the provider must have actually established the causal relationship. The latter view is supported by the fact that the obligation in Article 55(1)(c) presupposes the existence of a serious incident, which in turn – by its definition – requires the necessary causal relationship.[608] If one were to strictly adhere to this wording, a report would therefore always have to be submitted only after the causal relationship has been established – albeit then 'without undue delay'.

202. This interpretation is also indicated in the GPAI Code of Practice, stating that signatories should report not later than two days after they 'become aware of the involvement of their model in the incident'[609] – also presupposing an actual incident. In light of the purpose of the reporting obligation, however, namely to enable a prompt and coordinated response to serious incidents, it seems more appropriate to assume a reporting duty as soon as the provider suspects a causal relationship. It could be countered that the Union legislature explicitly allows such suspicion to suffice only in Article 73(4), which might suggest, *a contrario*, that suspicion should not suffice elsewhere. Yet, it appears more likely that – in both Articles 73 and 55(1)(c) – the amendment of the Article 3(49) definition from 'leads, might have led or might lead' to 'leads' in the final version was simply not taken into account.[610] Under the earlier definition, the suspected causal relationship would already have been included in the definition of a serious incident. Providers of GPAI models with systemic risk should therefore submit a report 'without undue delay' from the time they suspect a causal connection, not only once they have positively established it.

[606] This also serves the main purpose of article 55(1) to better assess and mitigate systemic risks, see Code of Practice, Safety and Security Chapter (n 9) recital (i).
[607] See Section 2.1.3.2.1.6.
[608] Also see Bernsteiner and Schmitt, 'Art 55' (n 24) para 17; Hartmann (n 417) para 11.
[609] Code of Practice, Safety and Security Chapter (n 9) Measure 9.3.
[610] Hartmann (n 417) para 11.

2.1.3.6.2. AI Office

203. Providers of GPAI models with systemic risk must report 'to the AI Office and, as appropriate, to national competent authorities'.[611] According to Article 55(3), the information or documentation obtained pursuant to Article 55(1)(c) must be treated confidential in the sense of Article 78.[612]

204. The Commission has published a reporting template for serious incidents involving general-purpose AI models with systemic risk.[613] Providers should use the EU SEND platform to fulfil their reporting obligation under Article 55(1)(c). The Commission has published technical guidance for submitting documents via EU SEND on its website.[614]

205. Article 55(1)(c) does not clarify whether – and, if so, which – follow-up obligations of the AI Office exist.[615] By contrast, the NIS2 Directive expressly provides that the 'CSIRT or the competent authority shall provide, without undue delay and where possible within 24 hours of receiving the early warning referred to in paragraph 4, point (a), a response to the notifying entity, including initial feedback on the significant incident and, upon request of the entity, guidance or operational advice on the implementation of possible mitigation measures'.[616] Accordingly there seems to be no institutional duty for the AI Office to provide reporting providers with guidance or operational advice. Nor is it expressly regulated whether (potentially) affected persons and/or the public must be informed of the serious incident. Comparable provisions on informing the public can be found, for example, in Article 23(7) NIS2 Directive or Article 17(2) CRA.

2.1.3.6.3. National competent authorities

206. Article 55 does not clarify when exactly it is appropriate to report the serious incident to national competent authorities. It is likely that this will be the case when their jurisdiction is triggered, in particular due to an incident that has effects within the Member State.[617] This could, for instance, be the case where citizens are killed or injured, where the state's critical infrastructure is being disrupted, or where property within the Member State is affected. Similarly, this could be the case where a cyber threat originates from within the Member State's territory.

2.1.3.6.4. Reporting does not entail admission of wrongdoing

207. Recital (j) of the Code of Practice states that '[t]he Signatories recognise that the reporting of a serious incident is not an admission of wrongdoing'.[618] As a preliminary matter, the normative weight of this formulation must be assessed with care. On its plain wording, recital (j) addresses only the *evidentiary status* of the provider's report: it shields the act of reporting from being construed as an acknowledgement of fault but says nothing about the underlying conduct giving rise to the incident. The mere fact that reporting does not constitute an admission of wrongdoing does not preclude the

[611] Article 55(1)(c).
[612] See Section 2.3.
[613] Reporting Template (n 574).
[614] European Commission, 'Guidelines for Providers of General-Purpose AI Models' (*European Commission*, 28 April 2026) <https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers> accessed 19 May 2026.
[615] Chatzipanagiotis (n 447) 39 notes that it would be 'very useful to set up and maintain a European Central Repository for AI incidents, facilitate information sharing among the national supervisory authorities and the AI Office, and provide tailored safety information to parties with a legitimate interest'.
[616] NIS2, art 23(5).
[617] Bernsteiner and Schmitt, 'Art 55' (n 24) para 16.
[618] Code of Practice, Safety and Security Chapter (n 9) recital (j)

existence of wrongdoing itself. Accordingly, the formulation cannot be read as precluding the AI Office from characterising the conduct underlying a reported incident as wrongdoing – let alone as conferring a general liability exemption on reporting providers.

208. This reading is reinforced by comparison with provisions in EU law that unambiguously establish liability exemptions through explicit language. The safe harbour provision in the DSA[619] – Article 6 – states in unequivocal terms that 'the service provider shall not be liable'. Equally explicit is the Regulation on the reporting, analysis and follow-up of occurrences in civil aviation,[620] which provides that '[t]he sole objective of occurrence reporting is the prevention of accidents and incidents and not to attribute blame or liability'.[621] Against this background, the formulation in the Code of Practice falls short of the standard of clarity that EU law seems to demand of a genuine liability exemption.[622]

209. It might be objected that reading a liability exemption into recital (j) could be normatively desirable. Such an exemption would come with distinct policy advantages: fear of liability – alongside broader legal uncertainty – constitutes a disincentive for companies to report incidents,[623] and this concern gains particular force with respect to incidents that are unlikely to be externally detectable – for example incidents that do not result in a materialisation of harm but only materially increase the systemic risk posed by the model.[624] In the absence of robust reporting incentives – other than existing fines – such incidents may go unreported entirely, leaving the AI Office uninformed. While this is a weighty consideration,[625] it cannot override the plain meaning of the text and would be more appropriately addressed through legislative intervention than interpretive extension.

210. It follows that the AI Office retains full authority to initiate investigations in response to a report – including investigations into whether the reporting provider has taken adequate measures to mitigate the systemic risk posed by its model. It might well be the case that the provider has not violated its obligations under Article 55(1)(a) and (b), and yet a serious incident still occurs. That said, a provider's consistent adherence to its reporting obligations may be treated as evidence of its broader commitment to regulatory compliance, a factor the AI Office can take into account when exercising its discretion to impose fines under Article 101,[626] in particular in light of the principle of proportionality. Moreover, one may read into Article 101 the principle underlying Article 99(7)(h),[627] pursuant to which particular account shall be taken of the extent to which the operator

[619] DSA (n 233).

[620] Regulation (EU) No 376/2014 of the European Parliament and of the Council of 3 April 2014 on the reporting, analysis and follow-up of occurrences in civil aviation and amending Regulation (EU) No 996/2010 and repealing Directive 2003/42/EC, Commission Regulation (EC) No 1321/2007 and Commission Regulation (EC) No 1330/2007 [2014] OJ L 122/18, art 1(2)

[621] See in more detail on 'just culture' in aviation safety and on drawing from reporting in aviation to improve the AI Act's reporting system more generally Chatzipanagiotis (n 447) 24–25 as well as 31 ff.

[622] See in the same sense Chatzipanagiotis (n 447) 33, noting that recital (j) of the Code of Practice (only) 'points to this direction'; it is important to note that even if the formulation in the recital (j) exemption with the same level of clarity as, for example, article 6 DSA, it would arguably be incapable of producing a binding legal effect *strictu sensu* considering its legal character, see, in detail, the commentary on Article 56 in this work.

[623] Wei and Heim (n 422) and further references cited therein.

[624] See Section 2.1.3.2.2.5.

[625] See Bommasani and others (n 422) 34.

[626] See Commission Guidelines (n 16) para 93 according to which 'commitments implemented in line with a code of practice that is assessed as adequate' will be taken into account by the Commission when fixing the amount of fines under article 101(1).

[627] Jens Schefzig, 'Art. 101 Geldbußen für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck' in Jens Schefzig and Robert Kilian (eds), *Beck'scher Online-Kommentar KI-Recht* (5th edn, C.H. Beck 2026) para 32.

notified the infringement on its own initiative. Reporting compliance does not therefore operate as a liability exemption – though it is not necessarily without legal consequence either.

2.1.3.7. Location of the incident

211. It is, at first glance, unclear whether the obligation to document, keep track of and report relevant information also covers serious incidents that occur outside the European Union. This question is quite important since many providers of GPAI models with systemic risk are established in third countries. Three questions must be distinguished.[628] The first is whether the EU legislature *can* regulate conduct outside of the Union. The second is whether the EU legislature *has* actually exercised that possibility in Article 55(1)(c), read in light of the provision's wording, its systematic context within the AI Act, and its regulatory purpose.[629] Third, it must be assessed whether any limits or exceptions speak against extending the obligation under Article 55(1)(c) to serious incidents happening outside the European Union.

2.1.3.7.1. Existence of extraterritorial jurisdiction triggers

212. In general, the Court of Justice of the European Union ("CJEU") permits territorial extensions of EU legislation on the basis of one of seven triggers: nationality, presence, conduct, (qualified) effects, anti-evasion, counterparty and property.[630] For the question at hand, the conduct trigger is of particular relevance. Additionally, the qualified effects trigger might support a reading under which Article 55(1)(c) extends to serious incidents happening outside the European Union.

213. The *conduct trigger* 'relates to an activity that at least partly occurs in the EU's territory' and requires 'some activity of a foreign entity [...] to connect to the EU's territory'.[631] A variant of the conduct trigger – market access – appears most relevant for the purposes of the present analysis.[632] The notion that the EU legislature may condition access to its market on compliance with EU standards has also been recognised by the CJEU. In *United Airlines*,[633] the Court held that 'the EU legislature may in principle choose to permit a commercial activity [...] to be carried out in the territory of the European Union only on condition that operators comply with the criteria that have been

[628] Also see Lena Hornkohl, 'The Extraterritorial Application of Statutes and Regulations in EU Law' (Max Planck Institute Luxembourg for Procedural Law 2022) Research Paper 2022(1) <https://doi.org/10.2139/ssrn.4036688> accessed 19 May 2026, 9 ['the extraterritoriality of EU law has to be identified on a case-by-case basis with regard to the objective of the EU instrument'] and 13 with regard to extraterritorial jurisdiction triggers ['It is worth mentioning that these factors can trigger the extraterritorial application of EU law but do not have to be applicable across the board. They are each only applicable in specific contexts for specific subject matters defined by the EU law itself or interpretation following the rationale and aim of the respective EU legal instrument'] as well as 24 on limits and exceptions.

[629] It is important to note that this section only addresses the EU's so-called prescriptive jurisdiction, i.e. the 'authority of a state to adopt legislation providing norms of conduct which govern persons, property or conduct.' The enforcement jurisdiction of the Union – that is 'the authority of a State to ensure compliance with its law' – is not discussed here, cf. ILC, 'Extraterritorial Jurisdiction' in 'Report of the International Law Commission on the Work of its Fifty-Eighth Session' (1 May–9 June and 3 July–11 August 2006) UN Doc A/61/10, annex E, para 5 <https://docs.un.org/a/61/10> accessed 19 May 2026.

[630] Hornkohl (n 628) 13.

[631] ibid 17.

[632] In greater detail ibid 18.

[633] Case C-561/20 *Q and Others v United Airlines, Inc.* [2022] ECLI:EU:C:2022:266.

established by the European Union and are designed to fulfil the [...] objectives which it has set for itself'.[634]

214. With regard to the question at hand, the conduct trigger is satisfied where a GPAI model provider places its model on the EU market. By doing so, GPAI model providers subject themselves to the rules of the AI Act, which, under Article 2(1)(a), expressly cover GPAI model providers 'irrespective of whether those providers are established or located within the Union or in a third country'.[635]

215. The *qualified effects trigger* – which the CJEU has recently held 'may, on its own, serve as the basis for the Commission's jurisdiction'[636] – is another basis for the extraterritorial application of EU law that may be able to offer additional support for a reading under which Article 55(1)(c) also extends to serious incidents happening outside the European Union. The qualified effects doctrine, developed in competition law and merger control,[637] captures conduct outside the Union that produces foreseeable, immediate and substantial effects on the internal (Union) market.[638] While it falls to the Commission to demonstrate that conduct has foreseeable, immediate and substantial effects in the European Union,[639] it seems possible to argue that it is foreseeable placing models that present systemic risk on the Union market will have immediate and substantial effects on the internal Union market.

2.1.3.7.2. Use of extraterritorial jurisdiction triggers

216. As stated above, the mere possibility of extraterritorial regulation does not, in itself, imply that the EU legislature intended a particular provision to apply extraterritorially. Rather, determining whether a provision carries extraterritorial reach requires a comprehensive interpretation of both the provision at issue and its systematic context.[640] The CJEU implicitly adopted this approach in *Google v CNIL*,[641] contending that although the purpose and objectives of an EU instrument may in principle justify its extraterritorial application,[642] it must also be apparent from the provisions at issue that the EU legislature 'ha[d] chosen to confer a scope on the rights enshrined in those provisions which would go beyond the territory of the Member States'.[643] Following this approach, this subsection now turns to whether Article 55(1)(c) should be interpreted as extending to serious incidents occurring outside the Union, having regard to the broader objective of Article 55 to assess and mitigate systemic risk.

217. The obligation under Article 55(1)(c) to keep track of, document and report relevant information about serious incidents should be understood as further specifying the overarching obligation under Article 55(1)(b) to assess and mitigate systemic risk.[644] Systemic risk is defined in Article 3(65) as a

[634] ibid para 58.
[635] Also see the forthcoming commentary on Article 2 in this work.
[636] Case C-367/22 *Air Canada v European Commission* [2026] ECLI:EU:C:2026:116, para 59.
[637] Bernadette Zelger, 'EU Competition Law and Extraterritorial Jurisdiction – a Critical Analysis of the ECJ's Judgement in Intel' (2020) 16 European Competition Journal 613, 619.
[638] ibid.
[639] *Air Canada v European Commission* (n 636) para 98.
[640] Hornkohl (n 628) 9.
[641] Case C-507-17 *Google LLC, successor in law to Google Inc. v Commission nationale de l'informatique et des libertés (CNIL)* [2019] ECLI:EU:C:2019:772.
[642] ibid para 58.
[643] ibid para 62.
[644] See para 12.

risk 'having a significant impact on the Union market [...] due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain'. Thus, the ability to propagate negative effects at scale is identified as an essential characteristic of systemic risk.[645] In particular, propagation occurs across the AI value chain, which in turn is understood broadly to span from GPAI model providers to AI system deployers and end users,[646] including actors who need not be based in the EU, provided the model is placed on the Union market. This understanding is reflected in Recital 110, which describes systemic risk in terms of an event giving rise to 'a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community'.[647]

218. The structure of Article 3(65) further supports this reading. The provision defines systemic risk as 'a risk having a significant impact on the Union market', which may arise by virtue of 'actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole'.[648] This indicates that harm to 'society as a whole' operates as one of the pathways through which a significant impact on the Union market may be established. Therefore, importantly, 'the society as a whole' criterion could be satisfied by negative effects that materialise outside Union territory,[649] provided a significant impact is realised on the Union market as result of those negative societal effects. A GPAI model with systemic risk that produces negative effects outside the EU but causes harm at a societal level is therefore not outside the AI Act's regulatory scope by virtue of its location, provided those negative effects are reasonably foreseeable to propagate across the value chain and impact the Union market.

219. Indeed, Article 55(1)(b) requires that providers assess and mitigate possible systemic risks stemming from the model throughout its development, placing on the market, and use – that is, along its lifecycle.[650] Notably, model development, including training and fine-tuning, may take place outside the EU even for GPAI models that are placed on the EU market, for example if computing infrastructure is located abroad. It may therefore be inferred that the intention of the EU legislature was to ensure that all processes pertaining to risk assessment and mitigation regardless of the geographical location of those processes, and which ultimately determine whether the risks posed by a model are acceptable for the model to be placed on and be used in the Union market, fall within the scope of the obligation under Article 55.[651] This logic reflects traditional EU product safety, which requires providers to ensure that their product complies with EU standards prior to market placement in a manner that effectively produces extraterritorial effects.[652]

---

[645] Code of Practice, Safety and Security Chapter (n 9) app 1.2.1(3).

[646] Code of Practice, Safety and Security Chapter (n 9) Measure 3.5, ['...do not use findings to threaten Signatories, users, or other actors in the value chain...'] (emphasis added).

[647] AI Act, recital 110.

[648] AI Act, art 3(65).

[649] See the forthcoming commentary on Article 3(65) in this work.

[650] AI Act, recital 110; Code of Practice, Safety and Security Chapter (n 9) recital (a) and Measure 1.2; also see the forthcoming chapter on Modifications in this work, Section 2.2.1.

[651] Hornkohl (n 628) 7, ['On the other hand, external objectives also prevail for problems the EU wants to address which extend beyond the bloc's territory. In light of the transboundary and worldwide nature of many environmental problems, EU environmental law often also includes measures to protect the environment beyond the EU's territory. [...] Action beyond the EU's borders is often necessary to achieve global environmental protection and tackle climate change.'].

[652] Hornkohl (n 628) 18 ['A large group for the conduct trigger giving rise to extraterritoriality consists of market access and imports to the EU.']; Joanne Scott, 'The New EU "Extraterritoriality"' (2014) 51 Common Market Law Review 1343, 1349.

220. As such, the obligations imposed on providers under Article 55 must be interpreted and understood in light of the EU legislature's recognition that systemic risks may arise at any stage of the GPAI model's lifecycle,[653] including at stages taking place outside the EU, and must nonetheless be assessed and mitigated. More specifically, Article 55(1)(a) and (b) make clear that the obligations to assess and mitigate systemic risks apply across the entire lifecycle of a GPAI model.

221. Such a reading is also functionally necessary in light of the objective of Article 55(1), namely to ensure appropriate systemic risk assessment and mitigation. Serious incidents serve as a key trigger for the reassessment of the risk acceptance determination.[654] Excluding serious incidents on the basis of their location would undermine this function and contradict the requirement that risk assessment and mitigation measures account for the entire model lifecycle.[655] This is also reinforced by an analysis of the serious incident concept specific to Article 55(1)(c). As noted above, the concept of serious incidents in Article 55(1)(c) is likely to be understood more broadly than the definition in Article 3(49) suggests.[656] A territorial restriction that required a serious incident to occur on Union territory would give rise to difficulties on several counts for the application of Article 55(1)(c). Again, as noted above, whilst the GPAI Code of Practice formulates the reporting trigger broadly – seemingly requiring only that the GPAI model be involved in, for example, a serious cybersecurity breach – the Commission's GPAI Guidelines are more specific, clarifying that a breach 'related to' a GPAI model suffices.[657] In other words, it seems like some of the serious incidents covered by Article 55(1)(c) need not originate *from* the GPAI model; it can be sufficient that they happen *to* it.[658]

222. Were one to assume that only serious cybersecurity breaches – and, depending on the interpretation followed above, other events that cause a material increase in systemic risk posed by the model[659] – occurring within the Union are covered, the obligation to report serious incidents under Article 55(1)(c) would largely be deprived of its practical effect. The physical infrastructure and the training and development activities related to GPAI models that pose systemic risk themselves will, in the overwhelming majority of cases, be located outside the Union. Moreover, it will often prove impossible to identify the territorial origin of cyberattacks on the model. To prevent the obligation in Article 55(1)(c) from being rendered ineffective with respect to serious cybersecurity breaches, such breaches occurring outside the Union to a GPAI model placed on the Union market should also fall within its scope.

2.1.3.7.3. Limits and exceptions

223. Lastly, interpreting Article 55(1)(c) as to encompass serious incidents occurring outside the European Union is not unreasonable.[660] From a policy perspective, the extraterritorial application of EU law can be seen as 'unreasonable when a state with weaker interests exercises prescriptive extraterritorial jurisdiction and does not defer to the state with stronger interests'.[661] To explore and

[653] AI Act, recital 110.
[654] Code of Practice, Safety and Security Chapter (n 9) Measure 7.6(4).
[655] Code of Practice, Safety and Security Chapter (n 9) recital (a).
[656] See para 136.
[657] See para 137.
[658] In detail on this and the requirement of causality between model and outcome, see Section 2.1.3.2.1.6.
[659] See Section 2.1.3.2.2.5.
[660] In detail on limits and exceptions see Hornkohl (n 628) 25 ff.
[661] Hornkohl (n 628) 25.

explain this doctrine, one can refer to the *Google Spain*[662] and *Google v CNIL* cases.[663] In both, the CJEU sought to reconcile the extraterritorial application of EU law with other interests, in particular those of third states.[664] Particularly in *Google v CNIL*, the Court recognised that third states hold the higher interest compared to the EU in determining the circumstances under which content should be de-referenced from search engines within their territories, particularly when balancing privacy rights against freedom of information.

224. For Article 55(1)(c), however, those criteria are arguably not fulfilled. What is at stake in Article 55(1)(c) is not an order to de-reference content – that is, to actively interfere with the public availability of content in a third state – but rather the extension of a reporting obligation concerning serious incidents to incidents occurring outside the Union. Whereas de-referencing directly removes content from search results in a third state's web version, thereby displacing that state's own calibration of freedom of information versus privacy rights, the reporting obligation at issue in Article 55(1)(c) amounts to, at worst, a duplication of the reporting obligation: the third state remains free to determine when it requires the provider to report serious incidents to its competent authorities. Parallel reporting thus does not override the third state's regulatory interests in the same manner as an extraterritorial de-referencing order.[665] It therefore strongly follows that the obligation under Article 55(1)(c) should be considered to extend to serious incidents occurring outside the Union as well.

2.1.4. Article 55(1)(d): Cybersecurity protection

225. Article 55(1)(d) requires providers to ensure an adequate level of cybersecurity protection for both the GPAI model with systemic risk and the physical infrastructure of the model.

2.1.4.1. General remarks

226. Article 55(1)(d) serves the purpose of ensuring the cybersecurity of GPAI models with systemic risk and thereby a consistent and high level of protection of public interests throughout the Union.[666] Additionally, the rationale of the provision can be seen in promoting the trustworthiness of AI in general,[667] which in turn is aimed at creating a safe and innovation-friendly environment. The provision can also be read in light of the EU's cybersecurity agenda.[668] The widespread deployment of AI gives rise to new and AI-specific risks which traditional approaches to cybersecurity may be

---

[662] Case C-131/12 *Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* [2014] ECLI:EU:C:2014:317.

[663] *Google LLC v CNIL* (n 641).

[664] Hornkohl (n 628) 26.

[665] As AG Kokott has argued, although double regulation may be burdensome for providers, no prohibition on double regulation can be derived from customary international law, see Case C-366/10 *Air Transport Association of America and Others v Secretary of State for Energy and Climate Change* [2011] ECLI:EU:C:2011:637, Opinion of AG Kokott, paras 156 ff.

[666] Henrik Nolte, Miriam Rateike and Michèle Finck, 'Robustness and Cybersecurity in the EU Artificial Intelligence Act' *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025) <https://doi.org/10.1145/3715275.3732020> accessed 20 May 2026, 3 with regard to AI Act, art 15, pointing to recital 7. These thoughts on the purpose of the cybersecurity provision in article 15 seem to be transferable to article 55(1)(d).

[667] Code of Practice, Safety and Security Chapter (n 9) 2.

[668] Nolte, Rateike & Finck (n 666) 2.

unable to address adequately.[669] Ensuring a sufficient Union-wide level of cybersecurity is described as 'one of the key challenges for the Union' and is essential for strengthening both the Union's economy and democracy.[670]

227. As discussed above, the constituent obligations under Article 55(1) 'complement and feed into each other.'[671] Against this background, the relationship between Article 55(1)(d) – the focus of this section – and Article 55(1)(b), which requires providers of GPAI models with systemic risk to 'assess and mitigate possible systemic risks at Union level', requires further clarification, particularly as regards the notion of cybersecurity risk.[672]

228. Article 55(1)(b) primarily captures possible systemic risks that emanate *from* the GPAI model itself as opposed to risks *to* the model.[673] As such, this relates, in particular, to cyber-offensive capabilities that scale with the model's capabilities and are explicitly identified by the Safety and Security Chapter of the GPAI Code of Practice as a specified category of systemic risk.[674] Article 55(1)(d), however, appears to pursue a broader objective: it seeks to ensure an adequate level of cybersecurity even beyond those cases in which vulnerabilities themselves qualify as possible systemic risks, since cybersecurity vulnerabilities may act as triggers for, or amplifiers of, pre-existing systemic risk factors in high-impact models.[675] For example, even the most robust safety mitigations may prove ineffective if the model is infiltrated or its parameters are illegitimately copied as this could allow a similar systemic risk to present itself elsewhere. Article 55(1)(d) thus serves to ensure that measures taken pursuant to Article 55(1)(a) and (b) are – and remain – effective. It thereby tries to 'limit the scenarios that could lead to materialised systemic risks'.[676] In other words, Article 55(1)(d) mainly focuses on model and infrastructural integrity in general –risks *to* the model – thereby securing effective risk assessment and mitigation, while Article 55(1)(b) focuses on possible systemic risks emanating *from* the model.

229. The cybersecurity duties imposed on GPAI models under Article 55 and high-risk AI systems under Article 15 should not be read in isolation.[677] They operate within a wider framework of EU laws on cybersecurity, including the Cyber Resilience Act[678], the Cybersecurity Act ("CSA")[679] and

---

[669] See similarly Henrik Junklewitz and others, 'Cybersecurity of Artificial Intelligence in the AI Act' (Joint Research Centre 2023) Science for Policy Report JRC134461 <https://doi.org/10.2760/271009> accessed 17 October 2025; for AI systems, see Finck (n 36) para 4.370.

[670] CRA (n 427) recital 1.

[671] See para 15; Commission Opinion (n 37) para 35.

[672] See Section 2.1.2.

[673] See Section 2.1.2.

[674] Code of Practice, Safety and Security Chapter (n 9) app 1.4(3): 'Risks from enabling large-scale sophisticated cyber-attacks, including on critical systems (e.g. critical infrastructure). This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved in offensive cyber operations, e.g. through automated vulnerability discovery, exploit generation, operational use, and attack scaling.'

[675] See similarly Hacker, Kasirzadeh and Edwards (n 225) 35.

[676] Commission Opinion (n 37) para 35.

[677] See similarly Schneider (n 20) para 20; and Bernsteiner and Schmitt, 'Art 55' (n 24) para 15.

[678] CRA (n 427); a more in depth analysis of the interplay between the AI Act and the CRA can be found at Hans Graux and others, 'Interplay between the AI Act and the EU Digital Legislative Framework' (Policy Department for Transformation, Innovation and Health Directorate-General for Economy, Transformation and Industry 2025) PE 778.575 <https://www.europarl.europa.eu/RegData/etudes/STUD/2025/778575/ECTI_STU(2025)778575_EN.pdf> accessed 20 May 2026, 59 ff.

[679] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity

the NIS2 Directive.[680] For non-GPAI models as well as GPAI models without systemic risk and non-high-risk systems, the AI Act does not impose any explicit additional cybersecurity requirements,[681] instead deferring to the general cybersecurity framework set out in the mentioned legislations.

2.1.4.2. Meaning of cybersecurity

230. The AI Act does not provide a definition for *cybersecurity* as used in Article 55(1)(d) (or Article 15).[682] The key interpretive question in this regard is how narrow or broad the term should be understood, as this informs the breadth of the providers' obligations. Three approaches may aid in a better understanding of the meaning of the term cybersecurity in Article 55(1)(d). First, one could draw a comparison to the cybersecurity obligation for providers of high-risk AI systems under Article 15. Second, guidance might be found in looking at definitions of cybersecurity in other EU legal acts the AI Act refers to. Third, the GPAI Code of Practice and, particularly, its security measures may help shed light on the meaning of cybersecurity in Article 55(1)(d). Importantly, these approaches should not be seen as alternatives to each other, rather their respective insights, taken together, aid in developing a workable understanding of the term in Article 55(1)(d).

2.1.4.3. Approaches to define cybersecurity

2.1.4.3.1. Comparison to Article 15

231. Article 15 lays down requirements on the '[a]ccuracy, robustness and cybersecurity' for high-risk AI systems. The understanding of 'cybersecurity' under that provision could help elucidate the meaning of cybersecurity in the context of GPAI models with systemic risk.

232. According to Article 15(1), high-risk systems 'shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle'. Article 15(5) further concretises the cybersecurity requirements. Its first sentence stipulates that high-risk AI systems 'shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities'. While this clearly does not amount to a definition of cybersecurity, it provides interpretive guidance as to the scope attributed to the concept in Article 15.

233. It seems like this understanding in Article 15 is limited in several ways. First, it is noteworthy that Article 15(5) covers resilience only 'against attempts by unauthorised *third parties*' (emphasis added). Similarly, Recital 76 refers to 'malicious third parties exploiting the system's vulnerabilities'. The AI Act does not provide for a more detailed definition of what constitutes a third party.[683] At a minimum, however, the term arguably encompasses all persons outside of the provider's

certification and repealing Regulation (EU) No 526/2013 [2019] OJ L 151/15; a more in depth analysis of the interplay between the AI Act and the CSA can be found at Graux and others (n 678) 56 ff.

[680] NIS2 (n 425).

[681] AI Act, art 50 requires providers of AI systems generating synthetic audio, image, video or text content to ensure that the technical solutions used to comply with that provision are 'effective, interoperable, robust and reliable' – requirements that may carry implicit cybersecurity implications, even if not explicitly framed as such.

[682] Finck (n 36) para 4.370; Mario Martini, 'Art. 15 Genauigkeit, Robustheit und Cybersicherheit' in Mario Martini and Christiane Wendehorst (eds), *KI-VO: Verordnung über Künstliche Intelligenz* (2nd edn, C.H. Beck, 2026) para 60.

[683] Rhian L M Moritz 'Art. 15 Genauigkeit, Robustheit und Cybersicherheit' in David Bomhard, Fritz-Ulli Pieper & Susanne Wende (eds), *KI-VO: Verordnung über künstliche Intelligenz* (Deutscher Fachverlag, 2025) para 61.

organisation who are not authorised to access or use the system.[684] Conversely, individuals acting under the provider's authority – such as employees – do not appear to qualify as third parties.[685] As a result, Article 15(5) does not address insider threats – that is, threats arising from individuals from within the provider's organisation who misuse their authorised access in order to harm the provider.[686] Moreover, Article 15(5) first sentence covers only attempts 'to alter [the high-risk AI systems] *use, outputs or performance*' (emphasis added). By implication, malicious attempts that are not directed at one of these three outcomes fall outside the scope of Article 15(5) – for example, the unlawful extraction of data through other channels than the system's output.[687] Finally, Article 15(5)'s first sentence refers exclusively to the exploitation of *'system vulnerabilities*' (emphasis added).[688] This, again, leaves insider threats unaddressed. Where an insider exploits authorised access to a system that might otherwise be free of vulnerabilities, such conduct would not fall within the scope of Article 15(5), first sentence.[689]

234. Article 15(5), third sentence, then turns to 'AI specific vulnerabilities'. This sentence seems to be understood as a concretisation of the first sentence, with AI-specific vulnerabilities treated as a subcategory of system vulnerabilities.[690] It provides that the technical solutions adopted to address those risks shall, where appropriate, include 'measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws'. Notably, Article 15(5)'s third sentence introduces requirements that directly target the underlying model – an atypical approach for a provision in Chapter III.[691] Where the provider of the high-risk AI system is also the provider of the underlying model, compliance with these requirements is unlikely to raise structural difficulties. This situation may differ, however, where the high-risk AI system builds upon a third-party model. This raises the question of the interaction with Article 55(1)(d), in particular as regards the allocation of cybersecurity-related obligations between high-risk AI system providers and providers of GPAI models with systemic risk. In this regard, interfaces controlled by the system provider will generally fall within that actor's realm of responsibility.[692] Where interfaces are instead controlled by the model provider, the model provider is required to draw up and make available to the system provider the information and documentation necessary to enable the latter to comply with its obligations, in accordance with Article 53(1)(b).

235. Finally, Article 15(4) suggests that cybersecurity, within the meaning of Article 15, is primarily concerned with *intentional* causes, while unintentional causes fall more naturally under the concept

---

[684] ibid.

[685] A definition of 'third party' can be found in data protection law in GDPR, art 4(10), which defines the term as 'a natural or legal person, public authority, agency or body other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data'.

[686] For the notions of *insider* and *insider threat*, see Cybersecurity and Infrastructure Security Agency, 'Insider Threat Mitigation Guide' (CISA 2021) <https://www.cisa.gov/sites/default/files/2022-11/Insider%20Threat%20Mitigation%20Guide_Final_508.pdf> accessed 6 March 2026, 9 ff.

[687] Moritz (n 683) para 63.

[688] ibid para 64.

[689] ibid.

[690] Henrik Nolte and Zeynep Schreitmüller, 'Cybersicherheit KI-basierter Medizinprodukte im Lichte der MDR und KI-VO' (2024) Medizinprodukterecht 28; see similarly Benedikt Buchner, 'Artikel 15 Genauigkeit, Robustheit und Cybersicherheit' in Jens Schefzig and Robert Kilian (eds), *Beck'scher Online-Kommentar KI-Recht* (5th edn, C.H. Beck 2026) para 59.

[691] Finck (n 36) para 4.387 ['All other obligations need to be complied with at the level of the AI system'].

[692] See Beurskens (n 20) para 8.

of robustness in Article 15(4).[693] It should be noted, however, that the two concepts do not lend themselves to a fully clean distinction. The choice to place both within the same provision of the AI Act appears to be a deliberate one, indicating their close relationship. Particularly in light of Article 15(5) last sentence, also addressing confidentiality attacks, cybersecurity may implicitly encompass certain robustness dimensions. Indeed, from a technical standpoint, some authors take the view that adversarial robustness is but one aspect of cybersecurity.[694]

236. In summary, the concept of cybersecurity underlying Article 15 appears to be comparatively[695] narrow. This is because it is confined, first, to attempts by unauthorised third parties, thereby excluding insider threats. Furthermore, Article 15 primarily seems to cover only those attempts aiming at altering the use, output or performance of a system with other malicious inferences falling outside the scope. Finally, Article 15(5) mainly addresses intentional causes, whereas unintentional causes are dealt with under the notion of robustness in Article 15(4).

2.1.4.3.2. References to other instruments defining cybersecurity

237. The AI Act also refers to other legal instruments – namely the CSA and the CRA – with regard to its cybersecurity requirements. These interconnections could further help guide the interpretation of the concept in the AI Act in general and Article 55(1)(d) respectively.

238. According to Article 15, read in conjunction with Article 42(2), high-risk AI systems 'that have been certified or for which a statement of conformity has been issued under a cybersecurity scheme pursuant to [the CSA] and the references of which have been published in the Official Journal of the European Union shall be presumed to comply with the cybersecurity requirements set out in Article 15 of this Regulation in so far as the cybersecurity certificate or statement of conformity or parts thereof cover those requirements'. Some authors argue that this indicates that the definition of cybersecurity found in the CSA also applies to the AI Act.[696]

239. That definition found in the CSA is broad, encompassing all 'activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyberthreats'. Article 2(1) CSA defines cyberthreats as 'any potential circumstance, event or action that could damage, disrupt or otherwise adversely impact network and information systems, the users of such systems and other persons'. This formulation – and thus the CSA's definition of cybersecurity – contrasts the characteristics distilled from the analysis of Article 15 AI Act in some points.[697] First, the broad definition contained in the CSA is not limited to attacks by third parties and therefore also encompasses insider threats. Second, it is not confined to attempts aiming at altering the use, output or performance of a system, but instead covers any adverse impact on networks and information systems, as well as users and other persons. Third, this definition is not restricted to intentional causes; it also includes unintentional causes, which, under Article 15 AI Act, are treated as matters of robustness rather than cybersecurity.

240. One argument against adopting the broad understanding of the term cybersecurity as defined in the CSA to Article 55(1)(d) is that no provision equivalent to Article 42(2) AI Act exists for GPAI models with systemic risk. Even if one were to argue that Article 42(2) read in conjunction with the

[693] Finck (n 36) para 4.370.
[694] Nolte, Rateike and Finck (n 666) 2.
[695] See para 242.
[696] Nolte, Rateike and Finck (n 666) 4; see similarly Martini, 'Art 15' (n 682) para 60.
[697] See similarly Buchner (n 690) para 37.

CSA provides the basis for a broad understanding of the term cybersecurity in Article 15, this reasoning would not automatically extend to GPAI models with systemic risk given the absence of an explicit corresponding cross-reference. Moreover, Article 42(2) expressly refers to Article 15 only. Had the EU legislature intended to rely on the CSA's definition throughout the Act, it would have made more sense to refer to that definition expressly. Additionally, even if one were to argue that the same must apply to GPAI models with systemic risk, it seems like there would be no big incentive for providers to obtain such a certification because providers of GPAI models with systemic risk will, in practice, most likely rather rely on the Code of Practice to demonstrate compliance with Article 55(1)(d).[698]

241. Another relevant EU legal act addressing cybersecurity is the CRA, which the AI Act refers to in its Recital 77.[699] High-risk AI systems falling within the scope of the CRA 'may demonstrate compliance with the cybersecurity requirements of [the AI Act] by fulfilling the essential cybersecurity requirements set out in [the CRA]'. The counterpart to the AI Act's Recital 77 can be found in Article 12 CRA on high-risk AI systems.[700] This sets out the exact requirements for when compliance with the CRA results in high-risk AI systems being 'deemed to comply with the cybersecurity requirements set out in Article 15 [of the AI Act]'. Since the CRA refers to the CSA for the definition of cybersecurity,[701] Article 12 CRA also appears to be based on the broad understanding of the term cybersecurity described above.[702] Once again, however, an argument against extending this broad concept of cybersecurity to Article 55(1)(d) is that legal concepts can be understood relatively – even within a single piece of legislation – and that, in the absence of a provision identical to the Article 12 CRA provision for GPAI models (with systemic risk),[703] it appears that the legislature did not intend harmonisation of the concept within the scope of Article 55(1)(d).[704]

242. In summary, the cross-references within the AI Act suggest an endorsement of a broad conception of cybersecurity in Article 55(1)(d), even if this sits uneasily with the comparatively narrow understanding reflected in Article 15(5). Although these provisions are expressly confined to high-risk AI systems, and there are arguments against a direct adoption of the concept to Article 55(1)(d), they might still be of relevance for the interpretation of Article 55(1)(d), since they seem to offer the most developed picture of what cybersecurity means within the AI Act. This cannot determine, but it at least informs, the interpretation of the concept's meaning in Article 55(1)(d).

2.1.4.3.3. Code of Practice

243. Further guidance for the understanding of the concept of cybersecurity in Article 55(1)(d) can be drawn from the Safety and Security Chapter of the GPAI Code of Practice,[705] as its security

[698] Graux and others (n 678) 58.

[699] It seems remarkable that this reference is only made in a recital and not the main text of the AI Act. However, since the interplay is specifically addressed in the CRA as the more recent instrument, this does not seem to be a problem; also see Graux and others (n 678) 64.

[700] In detail see ibid.

[701] CRA, art 3(3).

[702] In detail on the interplay with the AI Act, see Moritz (n 683) para 13 ff.

[703] Critical in this regard Claudio Novelli and others, 'Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' (2024) 55 Computer Law & Security Review 106066, 12.

[704] It may further be argued that the wording of recital 77, which distinguishes between the cybersecurity requirements of the CRA and those of the AI Act, does not in fact point towards a uniform understanding. Rather, recital 77 may be read as acknowledging the divergent approaches of the two instruments whilst nonetheless proceeding on the assumption that equivalent levels of security are achieved.

[705] See, more extensively, the commentary on Article 56 in this work.

commitments help indicate the scope of 'cybersecurity' in Article 55(1)(d). Although the Code itself is not legally binding on non-signatories,[706] it can offer interpretive value in the form of 'expert-crafted guidance'.[707] At the moment, the GPAI Code of Practice is therefore a key point of guidance in informing all providers' breadth of obligations under Article 55(1)(d). That is why the following paragraphs will examine the Code's requirements in detail, with a view to distil insights on the scope and meaning of cybersecurity in Article 55(1)(d).

244. As the Commission's adequacy assessment rightly finds, 'Commitment 6 specifies how providers of general-purpose AI models with systemic risk may ensure an adequate level of cybersecurity protection ... pursuant to Article 55(1), point (d).'[708] To this end, the commitment requires signatories to define a Security Goal identifying the threat actors their mitigations are designed to address – including non-state external threats, insider threats, and other anticipated threat actors. The aforementioned Security Goal can be met by signatories by implementing appropriate security mitigations – staged appropriately in line with the increase in model capabilities.[709] Appendix 4 to the Code of Practice specifies these. Signatories deviating from the mitigations mentioned in Appendix 4 will need to implement alternative adequate mitigations, which the European Commission will likely assess in reference to the ones listed.[710] The following paragraphs will deal with different components to the cybersecurity requirements in Article 55(1)(d), which are evident from the categories under the Code of Practice, in turn.

2.1.4.3.3.1. General security mitigations

245. A first component of the requirements under Article 55(1)(d) – as reflected in Appendix 4.1 of the Code – concerns general security mitigation measures. These measures aim to prevent unauthorised network access and reduce the risk of social engineering, malware infection and malicious use of portable devices, and vulnerability exploitation and malicious code execution.[711] All of these mitigation measures are designed to ensure that only authorised persons have access to the model and other sensitive information and that the integrity of the model and its infrastructure is safeguarded. This illustrates the focus of Article 55(1)(d) on security for the model itself, thereby preventing or at least minimising possible systemic risks emanating from it.

246. Unauthorised network access is to be prevented by requiring providers to implement 'strong identity and access management practices, including restrictions on device and account sharing, multi-factor authentication, strong password enforcement, strong access management tools, 802.1x authentication,[712] zero trust architecture,[713] protection of wireless networks to the same standard as

[706] See the commentary on Article 56 in this work.
[707] Hacker, Kasirzadeh and Edwards (n 225)16; in detail on the interpretive value of the Code of Practice, see Section 2.1.
[708] Commission Opinion (n 37) para 34.
[709] Code of Practice, Safety and Security Chapter (n 9) Measure 6.1.
[710] Code of Practice, Safety and Security Chapter (n 9) Measure 6.2. Cf. para 307.
[711] Code of Practice, Safety and Security Chapter (n 9) app 4.1.
[712] The purpose of this standard is to provide 'compatible authentication, authorization, and cryptographic key agreement mechanisms to support secure communication between devices'. For further details see IEEE, 'IEEE Standard for Local and Metropolitan Area Networks–Port-Based Network Access Control' (IEEE 2020) 802.1X-2020 <https://doi.org/10.1109/IEEESTD.2020.9018454> accessed 20 May 2026.
[713] 'Zero trust (ZT) is the term for an evolving set of cybersecurity paradigms that move defenses from static, network-based perimeters to focus on users, assets, and resources. A zero trust architecture (ZTA) uses zero trust principles to plan industrial and enterprise infrastructure and workflows. Zero trust assumes there is no implicit trust granted to assets or user accounts based solely on their physical or network location (i.e. local area networks versus the

wired networks,[714] and the separation of any guest networks from the work network'.[715] The GPAI Code of Practice does not, however, prescribe in detail how exactly these measures are to be implemented. Multi-factor authentication ("MFA"), for example, may be realised in various forms. It may rely on knowledge factors (something an individual knows, such as a security question), possession factors (something an individual possesses, such as a physical token), inherent factors (something an individual is or has, such as physical characteristics) or location factors.[716] Not all combinations of these factors offer an equivalent level of security,[717] so providers should still ensure that they do not adopt the measures mentioned in the GPAI Code of Practice in a purely formal manner but instead assess their suitability and security in light of their specific model. Providers might find useful guidance in the Commission's implementing regulation for the application of the NIS2 Directive[718] as well as in ENISA's accompanying technical implementation guidance, especially with regard to MFA.[719]

247. Providers are further required to reduce the risk of social engineering[720] by implementing 'email filtering for suspicious attachments, links and other phishing attempts'.[721] It appears reductive to confine mitigation efforts with regard to social engineering to the filtering of emails while excluding other communication channels like Slack or Teams.[722] It is, however, doubtful whether the GPAI Code of Practice can be interpreted to the effect that email filtering is intended to serve merely as an example. In its third draft, the Code of Practice referred more broadly to 'strong protections

internet) or based on asset ownership (enterprise or personally owned).' For further details see Scott Rose and others, 'Zero Trust Architecture' (National Institute of Standards and Technology 2020) NIST Special Publication 800–207 <https://doi.org/10.6028/NIST.SP.800-207> accessed 6 March 2026.

[714] While the requirement seems to treat wired and wireless networks as equivalent, this equivalence is not as clear from a technical standpoint. This is because, as noted by Murugiah Souppaya and Karen Scarfone, 'Guidelines for Securing Wireless Local Area Networks (WLANs)' (National Institute of Standards and Technology 2012) NIST Special Publication 800–153 <https://doi.org/10.6028/NIST.SP.800-153> accessed 6 March 2026, 'WLANs are typically less secure than their wired counterparts for several reasons'.

[715] Code of Practice, Safety and Security Chapter (n 9) app 4.1(1).

[716] Ang Kok Wee, Eyasu Getahun Chekole and Jianying Zhou, 'Excavating Vulnerabilities Lurking in Multi-Factor Authentication Protocols: A Systematic Security Analysis' (arXiv, 29 July 2024) <https://doi.org/10.48550/arXiv.2407.20459> accessed 20 May 2026, also providing a comprehensive taxonomy of possible authentication factors. See also Commission Implementing Regulation (EU) 2024/2690 (n 500) additionally mentioning 'access from an unusual location, from an unusual device or at an unusual time' as possible factors to take into consideration.

[717] Wee, Chekole and Zhou (n 716). See also Konstantinos Moulinos and Marianthi Theocharidou, 'Technical Implementation Guidance on Commission Implementing Regulation (EU) 2024/2690 of 17 October 2024 Laying down Rules for the Application of NIS2 Directive as Regards Technical and Methodological Requirements of Cybersecurity Risk-Management Measures' (ENISA 2025) <https://doi.org/10.2824/2702548> accessed 20 May 2026.

[718] Commission Implementing Regulation (EU) 2024/2690 (n 500).

[719] Moulinos and Theocharidou (n 717) 144 ff.

[720] Also see Sella Nevo and others, 'Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models' (RAND 2024) Research Report RRA2849-1 <https://www.rand.org/pubs/research_reports/RRA2849-1.html> accessed 20 May 2026 on social engineering and prominent examples.

[721] Code of Practice, Safety and Security Chapter (n 9) app 4.1(1).

[722] Jamila Boutemeur and others, 'ENISA Threat Landscape 2025' (ENISA 2025) <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025> accessed 20 May 2026, 12 on phishing in general; see also CISA, 'Phishing Guidance: Stopping the Attack Cycle at Phase One' (CISA 2025) <https://www.cisa.gov/resources-tools/resources/phishing-guidance-stopping-attack-cycle-phase-one> accessed 20 May 2026, 4.

against social engineering'[723] – its adopted version, however, only refers to the 'reduction of risk of social engineering' and expressly only mentions email filtering.[724]

248. Additionally, providers are required to reduce the risk of malware infection and of malicious use of portable devices through 'policies regarding the use of removable media'.[725] This can include, for example, the prohibition of connecting removable media without an organisational reason for use or scanning the media for malicious code before use.[726] Attackers might even place portable devices in or close to target facilities – for example in a parking lot[727] – hoping that employees would plug them into their computer.[728] Again, providers might find useful guidance in the Commission's implementing regulation for the application of the NIS2 Directive[729] as well as ENISA's accompanying technical implementation guidance.[730]

249. Lastly, providers are mandated to reduce the risk of vulnerability exploitation and malicious code execution through 'regular software updates and patch management'.[731] This can include ensuring that patches come from trusted sources, are tested before being applied and are applied within a reasonable time.[732]

2.1.4.3.3.2. Protection of unreleased model parameters

250. A second component of the security requirements under Article 55(1)(d) – as reflected in Appendix 4.2 of the Safety and Security Chapter of the GPAI Code of Practice – concerns security mitigations that aim to protect unreleased model parameters. Securing unreleased model parameters is of high importance, since access to them by unauthorised and/or malicious actors may result in the model being deployed without the necessary systemic risk assessments and mitigations, thereby making the materialisation of systemic risks significantly more likely. The obligation to secure the model's unreleased parameters thus contributes to ensuring the effectiveness of the measures adopted pursuant to Article 55(1)(a) and (b).[733] That is because, even if the provider's mitigations under those provisions are robust, once an attacker gains access to the models weights they can misuse them to operate the model without any restrictions or monitoring[734] – thereby giving rise to systemic risks. Materials such as the RAND Securing AI Model Weights report (hereinafter "RAND report") will be particularly useful for providers with regard to Appendix 4.2, not least because the Code of Practice explicitly lists the report as an example of 'relevant guidance'.[735]

---

[723] European Commission, 'Third Draft of the General-Purpose AI Code of Practice: Commitments by Providers of General-Purpose AI Models with Systemic Risk – Safety and Security Section' (European Commission 2025) <https://ec.europa.eu/newsroom/dae/redirection/document/113608> accessed 20 May 2026 ("Third Draft"), Measure II.7.1(2).

[724] Code of Practice, Safety and Security Chapter (n 9) app 4.1(2).

[725] ibid app 4.1(3).

[726] Commission Implementing Regulation (EU) 2024/2690 (n 500) annex 12.2.3(b), and Moulinos and Theocharidou (n 717) 151 ff.

[727] Nevo and others (n 720) 60.

[728] ibid.

[729] Commission Implementing Regulation (EU) 2024/2690 (n 500).

[730] Moulinos and Theocharidou (n 717).

[731] Code of Practice, Safety and Security Chapter (n 9) app 4.1(4)

[732] Commission Implementing Regulation (EU) 2024/2690 (n 500) annex 6.6, and Moulinos and Theocharidou (n 717) 92 ff.

[733] See Sections 2.1.1. and 2.1.2.

[734] Nevo and others (n 720) 2.

[735] Code of Practice, Safety and Security Chapter (n 9) Measure 7.3(3).

251. First, providers are required to achieve 'accountability over all copies of stored model parameters across all devices and locations' through 'a secure internal registry of all devices and locations where model parameters are stored'.[736] That makes sense because the securing of (copies of) model parameters will largely be ineffective if it is not clear where and how many copies of the model parameters exist in the first place.[737] After all, a single unprotected copy is sufficient to circumvent the security measures adopted by the provider.[738]

252. Further, providers are required to prevent unauthorised copying of model parameters to unmanaged devices through 'access management on all devices storing model parameters, with alerts in case of copying to unmanaged devices'.[739] This measure is closely linked to the previously mentioned requirement to ensure accountability over all copies of the model parameters. Effective access management requires providers to ensure that all devices storing the model parameters are known. External guidance may offer examples of how such access controls can be implemented. For example, the RAND report recommends – in its Security Level ("SL") 3[740] – which has been referenced by earlier preparatory drafts of the GPAI Code of Practice[741] – storing weights in a small number of centrally managed locations so that employees and researchers 'cannot simply make an additional copy'.[742] On this level, the report further suggests to protect all 'sensitive interactions (including access to the weights themselves rather than using them for inference, and any editing of the code of the weights interface system)'.[743] According to the RAND report, this can be facilitated, *inter alia*, by restricting the ability to make copies of the weights to 20 people, not granting third-party services access to the weights, not granting anyone persistent access, requiring multi-party authorisation and requiring security review for sensitive interactions.[744]

253. Additionally, providers are mandated to prevent unauthorised access to model parameters during transport and at rest through 'ensuring model parameters are always encrypted during transportation and storage as appropriate, including encryption with at least 256-bit security and with encryption keys stored securely on a Trusted Platform Module (TPM)'.[745] Providers therefore need to at least make sure that they never use public or unencrypted channels for (plaintext) weight transport and that they secure their parameters at rest accordingly.[746]

254. Providers should also aim to prevent unauthorised access to model parameters during temporary storage through 'ensuring model parameters are only decrypted for legitimate use to non-persistent memory'.[747] Model parameters need to be decrypted, generally said, when the parameters are needed for use – for example during training or fine-tuning – or when the model is being evaluated.[748]

---

[736] ibid.
[737] See Nevo and others (n 720) 51.
[738] Similarly, ibid 33.
[739] Code of Practice, Safety and Security Chapter (n 9) app 4.2(2).
[740] In detail on the Security Levels in the RAND Report, Nevo and others (n 720).
[741] Third Draft (n 723) Commitment II.7.
[742] Nevo and others (n 720) 78.
[743] ibid 80.
[744] ibid 80.
[745] Code of Practice, Safety and Security Chapter (n 9) app 4.2(3).
[746] As suggested by RAND SL2, see Nevo and others (n 720) 95.
[747] Code of Practice, Safety and Security Chapter (n 9) app 4.2(4).
[748] See Nevo and others (n 720).

What constitutes legitimate use can be informed by all other measures prescribed in the Code of Practice as well as other applicable laws.

255. Closely related to that, providers are required to prevent unauthorised access to model parameters during use through 'implementing confidential computing as appropriate, using hardware-based, and attested trusted execution environments'.[749] A trusted execution environment ("TEE") 'provides an isolated environment [...] that safeguards processed data by encrypting the incoming and outgoing data'[750] and 'protects the data and computation against any potentially malicious entity residing in the system'.[751] This measure seems to be in line with SL4 in the RAND report,[752] which further concretises and suggests to ensure that the TEE includes protection against physical attacks, model weights are only encrypted by a key that is generated and stored in the TEE, and the TEE will only run pre-specified and audited signed code.[753]

256. Lastly, providers are required to prevent unauthorised physical access to systems that host model parameters through 'restricting physical access to data centres and other sensitive working environments to required personnel only, along with regular inspections of such sites for unauthorised personnel or devices'.[754] As physical access to systems that host model parameters will often be equivalent to access to parameters on the system,[755] providers will have to operate with the same caution for physical access to systems storing parameters as they operate for the parameters themselves. Additionally, even if robust encryption is in place, physical access to systems can be a first step in enabling further attacks.[756] Similarly, in SL3, for example, the RAND report suggests that the physical security entail that data centres are guarded and locked at all times and that premises are swept for intruders frequently and for unauthorised devices routinely.[757]

2.1.4.3.3.3. Hardening interface access to unreleased model parameters

257. A third component of security requirements under Article 55(1)(d) – as reflected in Appendix 4.3 – deals with hardening interface access to unreleased model parameters while in use. These measures aim to protect the model's parameters in use, because at that time they are specifically vulnerable – especially to be illegitimately copied – since the parameters are decrypted then.[758] As already stated above, it is essential to ensure that there are no less-secure copies of the parameters in existence, as this would risk undermining the whole systemic risk assessment and mitigation process pursuant to Article 55(1)(a) and (b).

258. Providers are required to prevent unnecessary interface access to models through 'explicitly authorising only required software and persons for access to model parameters, enforced through multi-factor authentication mechanisms, and checked on a regular basis of at least every six

[749] Code of Practice, Safety and Security Chapter (n 9) app 4.2(5).
[750] Xiaoguo Li and others, 'A Survey of Secure Computation Using Trusted Execution Environments' (arXiv, 23 February 2023) <https://arxiv.org/abs/2302.12150v1> accessed 18 February 2026.
[751] ibid.
[752] See more extensively on the Security Levels (SLs) Nevo and others (n 720).
[753] Nevo and others (n 720) 86.
[754] Code of Practice, Safety and Security Chapter (n 9) app 4.2(6).
[755] See Nevo and others (n 720) 59.
[756] ibid.
[757] ibid 78.
[758] See ibid.

months'.[759] This measure is based on the observation that, in many organisations, a significant number of individuals have access to models.[760] By restricting both the range of authorised software and the number of authorised users, the risk of illegitimate copying is correspondingly reduced.[761]

259. Additionally, providers must reduce the risk of vulnerability exploitation or data leakage through 'thorough review of any software interfaces with access to model parameters by a security team to identify vulnerabilities or data leakage, and/or automated security reviews of any software interface code at least to the same standard as the highest level of automated security review used for other sensitive code'.[762]

260. Further, providers are required to reduce the risk of model parameter exfiltration through 'hardening interfaces with access to model parameters, using methods such as output rate limiting'.[763] Output rate limiting is an effective means of defending against parameter exfiltration because it ensures that the exfiltration of a significant portion of the weights would take too long to be practical.[764] On SL4, the RAND report even suggests hardware-enforced limits on output rates.[765]

261. Lastly, providers must reduce the risk of insider threats or compromised accounts through 'limiting the number of people who have non-hardened interface-access to model parameters.'[766] As noted above, it is essential to limit the number of people who have access to the model parameters through non-hardened interfaces to the necessary level.

2.1.4.3.3.4. Insider threats

262. Another component of security requirements under Article 55(1)(d) – as reflected in the in the Safety and Security Chapter of the GPAI Code of Practice's Appendix 4.4 – addresses protection against insider threats 'including in the form of (self-)exfiltration or sabotage carried out by models'. GPAI models with systemic risk constitute highly attractive targets, making insider threats a realistic and significant risk vector.[767] Security that only faces external threats is therefore not sufficient for most, if not all providers of GPAI models with systemic risk. Providers might find guidance with regard to the following measures in the CISA's Insider Threat Mitigation Guide[768] and the National Insider Threat Task Force's Insider Program Maturity Framework[769], both of which are referenced in the RAND report.[770]

263. First, providers are required to protect the model's parameters from insider threats attempting to gain work-related access through 'background checks on employees and contractors that have or might reasonably obtain read or write access to unreleased model parameters or systems that

[759] Code of Practice, Safety and Security Chapter (n 9) app 4.3(1).
[760] Nevo and others (n 720).
[761] Similar ibid.
[762] Code of Practice, Safety and Security Chapter (n 9) app 4.3(2).
[763] ibid app 4.3(3).
[764] ibid app 4.3(4).
[765] Nevo and others (n 720) 86.
[766] Code of Practice, Safety and Security Chapter (n 9) app 4.3(4).
[767] See also Nevo and others (n 720) from SL3 onwards.
[768] CISA, 'Insider Threat Mitigation Guide' (n 686).
[769] National Insider Threat Task Force, 'Insider Threat Program Maturity Framework' (Director of National Intelligence 2018) <https://www.dni.gov/files/NCSC/documents/features/NITTF_MaturityFramework_web.pdf> accessed 20 May 2026.
[770] Nevo and others (n 720) 83.

manage the access to such parameters'.[771] The RAND report suggests having employees with parameter access to go 'through extensive screening every six months'.[772] Further guidance on which indicators can be relevant in background checks more generally can be found in the CISA guide.[773]

264. Additionally, providers are required to raise awareness of the risk of insider threats through 'the provision of training on recognising and reporting insider threats'.[774] Oftentimes, insider threats will only be detectable through the cooperation of employees who interact with their colleagues daily.[775] A study found that in nearly 40% of employee data-exfiltration cases suspicious behaviour had been observed in advance by co-workers.[776] This suggests that employee awareness forms a central component of any effective insider-threat mitigation strategy.[777] The RAND report therefore recommends providing employees with guidance on what constitutes suspicious behaviour and on the appropriate reporting and response mechanisms.[778] The CISA guide likewise offers practical orientation on the effective design and implementation of employee training and awareness programmes.[779]

265. Further, providers must reduce the risk of model self-exfiltration through 'sandboxes around models, such as virtual machines and code execution isolation'.[780] Providers must therefore not only have measures against internal human threats in place, but also against those threats arising from the model execution environment itself becoming a factor in parameter or data exfiltration.

266. Lastly, providers are required to reduce the risk of sabotage to model training and use 'through checking training data for indications of tampering'. Research indicates that data-poisoning attacks could be more practicable than previously assumed – finding that injecting as few as 250 malicious documents into pre-training data can suffice to introduce vulnerabilities to backdoor attacks across models of varying sizes.[781]

2.1.4.3.3.5. Security assurance

267. Under another component of the security requirement in Article 55(1)(d) – as reflected in Appendix 4.5 – providers will 'obtain assurance that their security mitigations meet the Security Goal by implementing additional security mitigations'.[782]

268. First, if the provider's internal expertise is inadequate, they must achieve external validation of their security mitigation effectiveness through 'regular independent external security reviews as appropriate to mitigate systemic risks'.[783] On SL2, for example, the RAND report requires review

[771] Code of Practice, Safety and Security Chapter (n 9) app 4.4(1).
[772] Nevo and others (n 720) 83.
[773] CISA, 'Insider Threat Mitigation Guide' (n 686) 64 ff.
[774] Code of Practice, Safety and Security Chapter (n 9) app 4.4(2).
[775] See Nevo and others (n 720) 83.
[776] ibid.
[777] ibid.
[778] ibid.
[779] CISA, 'Insider Threat Mitigation Guide' (n 686) 55ff.
[780] Code of Practice, Safety and Security Chapter (n 9) app 4.4(3).
[781] Alexandra Souly and others, 'Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples' (arXiv, 8 October 2025) <https://doi.org/10.48550/arXiv.2510.07192> accessed 16 May 2026.
[782] Code of Practice, Safety and Security Chapter (n 9) app 4.5.
[783] ibid app 4.5(1).

and penetration testing by an 'accredited third-party organization'.[784] On SL3, it is suggested that the security team should perform continuous penetration testing, laying a focus on interfaces to the weights; penetration testing of physical access and '[a]dvanced red-teaming'.[785] This entails having a highly capable external team[786] which receives 'significant funding' and is given access to the system design and code so they can perform whitebox red-teaming.[787] Additionally, those elite external teams should be given employee credentials to be able to test insider threats[788] as well as expanded access in general.[789]

269. Providers' security assurance obligations further encompass the validation of their network and physical access management as well as their security gap identification through 'frequent red-teaming as appropriate to mitigate systemic risks'.[790] Additionally, providers are required to validate their network software integrity through 'competitive bug bounty programs to encourage public participation in security testing of public-facing endpoints as appropriate to mitigate systemic risks'.[791]

270. Providers will also have to validate their insider-threat security mitigations through 'periodic personnel integrity testing'.[792] This resembles the RAND report's SL4, suggesting occasional employee integrity testing,[793] at the same time noting, however, that 'the predictive reliability of different integrity testing approaches is unclear'.[794] Moreover, providers facilitate the reporting of security issues through 'secure communication channels for third parties to report security issues'.[795] To detect suspicious or malicious activity, providers will 'install Endpoint Detection and Response ("EDR") and/or Intrusion Detection System (IDS) tools on all networks and devices'.[796]

271. Lastly, to be able to respond timely and effectively to malicious activity, providers will need to make 'use of a security team to monitor for EDR alerts and conduct security incident handling, response, and recovery for security breaches in a timely and effective manner'.[797]

272. In summary, the GPAI Code of Practice reflects a notably expansive understanding of cybersecurity within the meaning of Article 55(1)(d). Although its primary focus lies on the protection of model parameters, the concept of cybersecurity in Article 55(1)(d) in general must be understood in structurally broader terms – encompassing, for example, insider-threat mitigation, security assurance and general security mitigations. This approach is structurally consistent with the objectives of Article 55 as a whole: even minor weaknesses in the implemented cybersecurity mitigations as well as a narrow understanding of the concept of cybersecurity may ultimately render the measures taken pursuant to Article 55(1)(a) and (b) ineffective.

---

[784] Nevo and others (n 720) 77.
[785] ibid 84.
[786] Further defined ibid.
[787] ibid.
[788] ibid.
[789] ibid.
[790] Code of Practice, Safety and Security Chapter (n 9) app 4.5(2).
[791] ibid app 4.5(3).
[792] ibid app 4.5(4).
[793] Nevo and others (n 720) 89.
[794] ibid.
[795] Code of Practice, Safety and Security Chapter (n 9) app 4.5(5).
[796] ibid app 4.5(6).
[797] ibid app 4.5(7).

2.1.4.3.4. Synthesis

273. Overall, it can be observed that Article 55(1)(d)'s scope, as clarified by the GPAI Code of Practice, is considerably broader than that of Article 15 AI Act. In particular, the Code's understanding explicitly encompasses insider threats and addresses them in detail. Moreover, the Code is not limited to system vulnerabilities. Rather, it covers any potential weakness that could – even only multiple steps later – ultimately jeopardise the model's systemic risk mitigations. This broader concept is reflected, for example, in the Code's attention to physical infrastructure and its general security mitigations, including, for example, policies on removable media. Finally, the Code is not confined to attempts aimed at altering the use, output or performance of systems, thereby extending beyond the narrower focus of Article 15.

274. In sum, the broader conceptual approach adopted in the Code of Practice appears aligned with the broad legal definition set out in the CSA. This alignment is coherent: since GPAI models (with systemic risk) 'may form the basis for a range of downstream systems',[798] it is particularly important that robust security standards apply already at the model level. Accordingly, the narrower definition underlying Article 15 should not be transposed to Article 55(1)(d).

2.1.4.4. Objective scope of protection: model and physical infrastructure

275. The cybersecurity protection obligation applies to both 'the general-purpose AI model with systemic risk and the physical infrastructure of the model'.[799]

276. Consistent with the broad understanding of cybersecurity mentioned above, the notion of a 'model' must likewise be interpreted broadly in this context.[800] Rather than pursuing a formal definition of the term, the measures set out in the Code of Practice provide a useful indication of its objective scope. In light of the provision's purpose – namely, ensuring that the measures adopted under Article 55(1)(a) and (b) remain effective and are not undermined – we can draw a few conclusions.

277. First, in any case, the scope encompasses the unreleased parameters of the model, its algorithms, and all copies thereof.[801] This is further supported by Recital 115, which expressly refers to the securing of model weights (and algorithms). Illegitimate copies of parameters may enable the circumvention of systemic risk mitigation measures and allow the model to be deployed without the safeguards required under Article 55(1)(a) and (b). Arguably, the notion of a model must likewise include any software interfaces that provide access to the model's parameters. Effective protection of the model as such necessarily entails securing interfaces through which access to the model can be obtained.[802] Finally, the scope also extends to data used to train the model. Recital 115 refers to 'securing [...] data sets' as one way of facilitating cybersecurity. Moreover, the GPAI Code of Practice also addresses the protection of training data by requiring signatories to mitigate the risk of sabotage during model training and use, for example by 'checking training data for indications of tampering'.[803] This interpretation is also teleologically sound: tampered data might equally

[798] AI Act, recital 101.
[799] AI Act, art 55(1)(d).
[800] Commission Guidelines (n 16) para 22.
[801] See Code of Practice, Safety and Security Chapter (n 9) app 4.2 measures.
[802] See Code of Practice, Safety and Security Chapter (n 9) app 4.3 measures.
[803] ibid app 4.4(4).

undermine the effectiveness of the systemic risk mitigation measures undertaken pursuant to Article 55(1)(a) and (b).

278. The AI Act also does not define exactly what is meant by the 'physical infrastructure' of a GPAI model with systemic risk. A first point of guidance for interpretation is Recital 115, which explicitly mentions 'physical access controls' as possible measures to ensure an adequate level of cybersecurity and refers to 'servers' as protected physical assets. Further guidance can additionally be drawn from the Code of Practice. It emphasises the prevention of unauthorised access to 'systems hosting model parameters', which, in its understanding, include 'data centres and other sensitive working environments'.[804] Beyond this, however, the contours of the model's physical infrastructure remain unclarified. It is therefore necessary to revert to the purpose of Article 55(1)(d), which seeks to ensure effective systemic risk assessment and mitigation. On that basis, the concept of the model's physical infrastructure should encompass all physical assets whose compromise or infiltration could ultimately undermine the systemic risk measures in place. This would include, for example, portable devices, irrespective of whether model parameters are stored on them. As long as such devices could, if infiltrated, constitute a first step towards, for example, parameter exfiltration, and thereby threaten the effectiveness of the systemic risk mitigations in place, such devices should be encompassed by the obligation under 55(1)(d).

279. On the other hand, an interpretation according to which all hardware and facilities that in any way support the operation of the model fall within the scope of Article 55(1)(d) would go too far. As noted above, it makes sense to interpret the notion of the model's physical infrastructure as covering those elements whose insufficient protection can lead to an increase in the systemic risk emanating from the model. For instance, the release of model weights could lead to a loss of control over a powerful model and may enable, for example, cyber-offence risks.[805] An inadequate protection of supporting facilities such as ancillary facilities unconnected to parameter storage, power systems or cooling infrastructure, on the other hand, may certainly be relevant for providers of GPAI models with systemic risk more generally.[806] However, it seems inappropriate to argue that such protection duties fall under Article 55(1)(d), as these risks do not primarily amplify or extend the systemic risks by undermining the systemic risk mitigation of the respective model itself, at least as long as compromise or infiltration of such physical assets can ultimately not lead to increased systemic risks stemming from the model – for example because power systems are completely separated from facilities in which model parameters are stored and the provider additionally ensures that power outages cannot undermine the parameters' security.

2.1.4.5. The providers' obligation to 'ensure' an adequate level of cybersecurity

280. It could be argued that the cybersecurity obligation can only apply to physical infrastructure that is under direct control of the provider.[807] However, this can be countered by the fact that Article 55(1)(d) states that the provider must 'ensure' an adequate level of cybersecurity. This can be

[804] ibid app 4.4(6).

[805] ibid app 1.4(3); also see Alfonso de Gregorio, 'Mitigating Cyber Risk in the Age of Open-Weight LLMs: Policy Gaps and Technical Realities' (arXiv, 21 May 2025) <https://doi.org/10.48550/arXiv.2505.17109> accessed 17 October 2025.

[806] Also see Erich Grunewald and Asher Brass Gershovich, 'Accelerating AI Data Center Security' (Institute for AI Policy and Strategy 2025) <https://www.iaps.ai/research/accelerating-ai-data-center-security> accessed 29 September 2025, 18.

[807] Similarly Bernsteiner and Schmitt, 'Art 55' (n 24) para 12.

understood to mean that the provider must guarantee, for example through contractual agreements including control rights, that sensitive working infrastructure not directly under its direct, factual control also meet an adequate level of security. Additionally, it can be argued that it would be contradictory to Article 55's goals if the obligations of providers were reduced as their control diminishes. Such an approach would create incentives for providers to outsource as much as possible to third-party infrastructure (which they can not control), thereby undermining AI safety and reliability.

2.1.4.6. Adequate level of cybersecurity

281. Providers must ensure an 'adequate' level of cybersecurity.[808] In Recital 115, the AI Act appears to further clarify how the adequacy of cybersecurity is to be assessed. It states that protection should be 'appropriate to the relevant circumstances and the risks involved'.

282. The GPAI Code of Practice provides some further, valuable input in this regard. It states that signatories will 'define a goal that specifies the threat actors that their security mitigations are intended to protect against ("Security Goal"), including non-state external threats, insider threats, and other expected threat actors, taking into account at least the current and expected capabilities of their models'.[809] Additionally, the Code clarifies that the 'implementation of the required security mitigations may be staged appropriately in line with the increase in model capabilities along the entire model lifecycle'.[810] This Security Goal plays an important role in the Safety and Security Model Reports the signatories commit to report to the AI Office.[811] That is because, under Measure 7.3(3) of the Safety and Security Chapter, signatories commit to describing their Security Goal, all security mitigations they implemented, and how those measures meet the Security Goal.

283. Importantly, the latter includes 'the extent to which they align with relevant international standards or other relevant guidance (such as the RAND Securing AI Model Weights report)'.[812] This raises the broader question of how the general requirement of an 'adequate level' of cybersecurity is to be understood – and in particular what role the state of the art plays in this context. In this regard, the Code of Practice appears instructive: whilst Article 55 refers to the state of the art in Article 55(1)(a), it does not define it.[813] The Code of Practice, by contrast, defines it in its Glossary as 'the forefront of relevant research, governance, and technology that goes beyond best practice'.[814] This forefront understanding is further underlined in Recital (a) of the Code, according to which signatories must generally adopt 'at least' the state of the art in order to implement appropriate measures.[815] Signatories are furthermore encouraged to advance the state of the art.[816] This seems to suggest that the requirement of an 'adequate level' of cybersecurity under Article 55 is to be understood in light of this forefront standard: adequacy is thus not satisfied merely by complying with established best practices but requires an orientation towards the forefront of current research. Against this backdrop,[817] the aforementioned RAND Securing AI Model Weights report may nevertheless serve

[808] In detail on this wording and its difference to the wording in article 15, see Nolte, Rateike & Finck (n 666) 8.
[809] Code of Practice, Safety and Security Chapter (n 9) Measure 6.1.
[810] ibid Measure 6.2.
[811] ibid Commitment 7.
[812] ibid Measure 7.3(3).
[813] See Section 2.1.1.1.
[814] Code of Practice, Safety and Security Chapter (n 9) 32.
[815] ibid recital (a).
[816] ibid recital (f).
[817] See, in detail, on the state of the art condition Section 2.1.1.1.

as a good starting point for providers to draw upon – not least because it extends to security levels that are not yet fully achievable and that aim to thwart operations by actors that have 'experience and expertise years ahead of the (public) state of the art',[818] thereby reflecting a standard upon which future forefront research may build.

284. As noted above and since the GPAI Code of Practice explicitly addresses the RAND Securing AI Model Weights report as relevant guidance, providers are likely best advised to rely on it as a primary point of orientation. The RAND report identifies 38 distinct attack vectors, distinguishes between a variety of potential threat actors and their respective capabilities, and, on that basis, proposes five distinct security levels accompanied by preliminary benchmarks.[819] The threat actor modelling envisaged under the Code's Security Goal can, in practice, orientate on and align with the Security Levels set out in the RAND report. These levels are structured according to the type and sophistication of attackers against whom a model must be secured. This approach also seems consistent with Recital 115, which refers, more generally, to the 'risks involved': both the threat landscape and the risks to the model are likely to scale with increases in the model's capabilities. The more capable a model, the greater the likelihood that highly sophisticated actors will seek to compromise it. The SLs in the RAND report range from SL1 ('A system that can likely thwart amateur attempts') to SL5 ('A system that could plausibly be claimed to thwart most top-priority operations by the top cyber capable institutions').[820]

285. While the third draft of the GPAI Code of Practice referred more explicitly to the RAND report in the context of the security mitigations under (then) Commitment II.7,[821] the final adopted version of the Code no longer contains an express reference to the RAND report under Measure 6 or Annex 4. Nevertheless, the final Code still, as mentioned above,[822] continues to emphasise the relevance of the RAND report, such that the guidance contained in the third draft may still serve as a useful interpretive aid. Under the third draft, signatories were expected to 'meet at least RAND SL3 or equivalent'. This seems to be a plausible first point of orientation for providers of GPAI models with systemic risk under the final GPAI Code of Practice and under Article 55(1)(d) as well. This is further supported by the fact that many of the security mitigations outlined above appear to be modelled, at least in part, on the measures associated with the RAND report's SL3 benchmark. At the same time, providers must ensure that their threat modelling under the Security Goal adequately considers whether reasonably foreseeable threats may also stem from SL4 or even SL5 adversaries.[823] In this regard, it should be noted that the growing availability of open-weight GPAI models has the potential to reshape the cybersecurity threat landscape, as both the capabilities and the actor profiles – such as those described in the RAND report – may evolve and escalate.[824]

286. Other 'relevant international standards' and 'relevant guidance' providers may build upon to ensure an adequate level in the aforementioned sense could especially include ISO/IEC, NIST or SOC publications.[825] Expressly mentioned by the third draft of the Code were ISO/IEC 27001, NIST 800-53 and SOC 2.[826] With regard to AI-specific risks, orientation might be drawn from ISO/IEC

[818] Nevo and others (n 720) 10.
[819] ibid 2.
[820] ibid 22.
[821] Third Draft (n 723) Commitment II.7.
[822] See para 283.
[823] Third Draft (n 723) Commitment II.7.
[824] de Gregorio (n 805) 3–4.
[825] Third Draft (n 723) Commitment II.7.1.
[826] ibid.

TR 27563:2023[827] or ISO/IEC DIS 27090[828]. Additionally, providers might find guidance in publications by ENISA, given the agency's mandate under the CSA; for example, ENISA's yearly threat landscape report may offer guidance for modelling the provider's Security Goal.

2.1.4.7. Exemption for less capable, publicly available and deleted models

287. Importantly, the Safety and Security Chapter of the GPAI Code of Practice exempts models from Commitment 6 – the security mitigations – in three distinct scenarios. First, models are exempt from the commitment where 'the model's capabilities are inferior to the capabilities of at least one model for which the parameters are publicly available for download' (inferior-capabilities exemption).[829] Second, the Code only requires signatories to implement security mitigations for a model 'until its parameters are made publicly available for download' (open access exemption).[830] Third, in the same sense, models are no longer subject to the exemption as soon as their parameters are 'securely deleted' (deletion exemption).[831]

288. The rationale underlying these exemptions in the Code appears to be that the primary objective of the security mitigations – namely, to protect model weights and thereby prevent the deployment or misuse of a highly capable model in the wrong hands – may no longer be necessary, or at least proportionate, where potential malicious actors can in any event readily download the parameters of an even more capable model, where the model's parameters are made available for download, and – obviously – where the model's parameters have securely been deleted.

289. Two aspects of the inferior-capabilities exemption appear challenging. First, the Code does not define when a model's capabilities are to be considered inferior to those of another model. Second, it remains unclear who is to make this determination and whether, and to what extent, such an assessment is subject to verification. With regard to model capabilities more generally, signatories will likely rely on the aspects listed in Appendix 1.3.1 as relevant indicators. However, it remains uncertain to what extent a reliable comparison between the signatory's own model and a potentially more capable external model is feasible in practice – the risk of misclassification residing with the provider. Second, it remains unclear how this exemption in the GPAI Code of Practice affects non-signatories. One could argue that, since the Code does not apply to providers who have not signed it, they cannot invoke the exemption in the first place. In reviewing whether a provider can demonstrate 'alternative adequate means of compliance' under Article 55(2), the AI Office could, in theory, take the position that it is not bound by an exemption formulated in the Code. On that view, the exemption might simultaneously function as an incentive to adhere to the Code. On the other hand, providers could argue that the same rationale – disproportionality to impose security mitigations in cases where a more capable model is freely available – must also inform the interpretation of 'adequate' in Article 55(1)(d), thereby leading to the same outcome as if an exemption applied. A position taken by the AI Office according to which they are not bound to an exemption in the Code would likely also violate the legitimate expectations of providers (arguably, including non-signatories),[832] since the Commission has expressly stated that providers 'can

[827] See Bernsteiner and Schmitt, 'Art 55' (n 24) para 13.
[828] Currently only available in draft form.
[829] Code of Practice, Safety and Security Chapter (n 9) Commitment 6.
[830] ibid.
[831] ibid.
[832] See the commentary on Article 56, Section 2.7.1.2. in this work.

demonstrate compliance with the obligations in Articles 53(1) and 55(1) AI Act by adhering to a code of practice that is assessed as adequate by the AI Office and the Board.'[833]

290. The two other exemptions – the open access exemption and the deletion exemption – appear less challenging yet are not entirely without difficulties. The former could be seen as creating an incentive for providers to proactively release their model weights in order to escape their obligations under Article 55(1)(d). The deletion exemption raises a distinct concern: the GPAI Code of Practice does not define when parameters are to be considered 'securely deleted', nor does it specify the technical standard against which such deletion is to be assessed. While general guidance on secure deletion exists – most notably NIST SP 800-88,[834] which addresses methods such as cryptographic erasure and data overwriting – this standard was not designed with AI model weights in mind. The question of how the AI Office is to verify compliance with the deletion exemption remains similarly unresolved.

2.1.4.8. Temporal scope

291. Although not apparent from Article 55(1)(d) itself, it follows from Recital 115 that the cybersecurity obligation applies 'along the entire model lifecycle' – but only 'if appropriate'. It remains unclear why this important clarification is included only in the (non-binding) recitals and not in the legislative text. For signatories of the Code of Practice, this additionally follows from Commitment 6 of the Safety and Security Chapter, according to which they 'commit to implementing an adequate level of cybersecurity protection for their models and their physical infrastructure *along the entire model lifecycle*' (emphasis added).[835]

292. The Commission tends to favour a broad interpretation of the term. In its GPAI Guidelines, it acknowledges that it is difficult 'to clearly delineate a model and its lifecycle' because of the 'iterative and interlinked process through which a provider may develop a "model"'.[836] This is why the Commission understands 'the notion of a "Model", and consequently its "lifecycle" in a broad sense'.[837] According to the GPAI Guidelines, the lifecycle of a model begins 'at the start of the large pre-training run'.[838] Importantly, the Commission states that any 'subsequent development of the model downstream of this large pre-training run performed by the provider or on behalf of the provider, whether before or after the model has been placed on the market, forms part of the same model's lifecycle rather than giving rise to new models'.[839]

293. The GPAI Code of Practice does not explicitly address how exactly it defines the lifecycle of a model. Recital (a), the 'Principle of Appropriate Lifestyle Management', does clarify that the model lifecycle includes the 'development that occurs before and after a model has been placed on the

[833] Commission Guidelines (n 16) para 91.

[834] Ramaswamy Chandramouli and Eric Hibbard, 'Guidelines for Media Sanitization' (National Institute of Standards and Technology 2025) NIST Special Publication 800-88r2 <https://doi.org/10.6028/NIST.SP.800-88r2> accessed 20 May 2026.

[835] Code of Practice, Safety and Security Chapter (n 9) Commitment 6.

[836] Commission Guidelines (n 16) para 22.

[837] ibid; also see the detailed discussion on the concept of lifecycle in the forthcoming chapter on Modifications, Section 2.2.1. in this commentary.

[838] ibid fn 5 defines this as 'the foundational training run conducted on a large amount of data to build the model's general capabilities, which may take place after smaller experimental training runs, and which may be followed by fine-tuning for specialisation or other post-training enhancements'.

[839] ibid.

market'. Much like the Commission's GPAI Guidelines, the Code therefore appears to be based on a broad understanding of the model lifecycle.

294. A difficulty that arises not only in this context is that, according to Recital 115, the obligation to ensure an adequate level of cybersecurity applies 'along the entire lifecycle' of the model, whereas Article 2(8) provides an exemption from the scope of the AI Act for research, development and testing activities. This tension is discussed in more detail elsewhere.[840]

2.2. Article 55(2): Compliance pathways

295. Article 55(2) details some of the ways in which providers of GPAI models presenting systemic risk can comply with their obligations under Article 55(1). More specifically, it details three distinct compliance pathways: harmonised standards, codes of practice and alternative adequate means. In choosing a compliance pathway, providers are, to some extent, limited to those pathways that have been formally adopted by the relevant authorities.

296. Regarding the former, it is clear that providers of GPAI models presenting systemic risk cannot rely on a harmonised standard in the absence of such a standard. Article 55(2) expressly acknowledges this in its first sentence, by directing providers to instead rely on a code of practice 'until' a harmonised standard is published. However, the availability of this pathway was itself contingent on the prior creation of such a code of practice. While a code of practice was eventually created,[841] Article 56(9) contemplates the possibility that no such code would exist or that it would be considered inadequate and provides that, in that scenario, the Commission may provide common rules by way of an implementing act.[842]

297. Depending on the availability and adoption of these compliance pathways at a given moment, it is important to assess who may, can or must rely on them and with what legal effects. We will briefly touch on this question of provider discretion throughout the remainder of this assessment, referring the reader elsewhere for a more extensive discussion.[843] On a general level, it is clear that the notion of legitimate expectations plays a key role regarding the question of whether providers *can* rely on these instruments.[844] More specifically, the fact that the Commission has assessed a given code of practice as adequate may preclude the Commission from arguing that a provider who did not sign the relevant code but nevertheless adheres to it regarding some AI Act obligation has violated that obligation.[845] Whether providers *have* to rely on these instruments, or whether the avenue of 'alternative adequate means' remains an option is, in general, more difficult to assess,[846] as will become clear below.

---

[840] See, more extensively, the forthcoming commentary on Article 2 in this work.
[841] Code of Practice, Safety and Security Chapter (n 9).
[842] Also see the commentary on Article 56 in this work.
[843] See commentary on Article 56 in this work.
[844] Also see commentary on Article 56 in this work.
[845] It does, moreover, not seem like the Commission will seek such enforcement, as the Commission Guidelines imply that non-signatories would be expected to demonstrate AI Act compliance through alternative adequate means, that the Commission will nevertheless assess by comparison to the Code of Practice, see Commission Guidelines (n 9) paras 95 and 96; see also, commentary on Article 56, Section 2.7.1. in this work.
[846] Also see the commentary on Article 56 in this work.

2.2.1. Harmonised standards

298. The AI Act positions harmonised standards as the *final* and principal compliance pathway for providers of GPAI models with systemic risk. Article 3(27) clarifies that the notion of 'harmonised standard' is to be understood in the sense of Article 2(1)(c) of Regulation (EU) No 1025/2012,[847] which defines a harmonised standard as 'a European standard adopted on the basis of a request made by the Commission for the application of Union harmonisation legislation'.[848]

299. In the European Union, such standards are developed by three bodies: the European Committee for Standardisation ("CEN"), the European Committee for Electrotechnical Standardisation ("Cenelec") and the European Telecommunications Standards Institute ("ETSI").[849] These standards are normally expected to reflect the state of the art.[850] Given the fast-paced evolutions in the field of AI and its evaluations,[851] and the fact that it typically takes standardisation bodies time to develop a standard (amongst others reasons because of the importance to consultant relevant stakeholders, as discussed below),[852] it is unclear how this reference to the state of the art could reflect 'the forefront of relevant research, governance, and technology that goes beyond best practice',[853] the definition that is applied to the state-of-the-art condition in Article 55(1)(a).[854] Instead, this reference likely[855] refers, more broadly, to 'generally acknowledged state of the art', the definition as it is typically deployed in the context of standardisation, reflecting 'what is currently and generally accepted as good practice.'[856]

300. The time-consuming process required to get standards 'right' implies that not all of the relevant obligations in Article 55(1) can reasonably be 'standardised'.[857] A notable example is thus 55(1)(a)'s reference to the state of the art in the sense of the forefront of relevant research, governance, and technology that goes beyond best practice.[858] This inherently more limited role for some of the Article 55 obligations means that providers of GPAI models with systemic risk will not be able to

[847] See AI Act, recital 121.

[848] Also see Clemens Bernsteiner and Thomas Rainer Schmitt, 'Art. 53 Pflichten für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck' in Mario Martini and Christiane Wendehorst (eds), *KI-VO: Verordnung über Künstliche Intelligenz* (2nd edn, C.H. Beck, 2026) para 61.

[849] Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation [2012] OJ L 316/12, art 10 as well as its annex I.

[850] AI Act, recital 121.

[851] e.g., Robert Kilian, Linda Jäck and Dominik Ebel, 'European AI Standards – Technical Standardisation and Implementation Challenges under the EU AI Act' (2025) 16 European Journal of Risk Regulation 1038, 1052 ['new threats and countermeasures constantly emerge'].

[852] See in the same sense Sebastian Hallensleben, 'Generative AI and International Standardization' (2025) 1 Cambridge Forum on AI: Law and Governance e14, 4 (where the author identifies two challenging components: first, the high pace of technological advancement (which he notes might slow down, however), and second, the fact that most engineers in this field are focused on developing frontier models rather than contributing to standards). Also see Kilian, Jäck and Ebel (n 851) 1043 (discussing how standardisation progress for the AI Act has been 'significantly slower than anticipated by the European Commission'); Marta Cantero Gamito, 'Harmonising Consensus: The (Geo)Political Economy of Standardisation in the AI Act' (SSRN, 9 January 2026) <https://doi.org/10.2139/ssrn.6294878> accessed 9 March 2026, 23 ['slower than anticipated'].

[853] See Code of Practice, Safety and Security Chapter (n 9), Glossary. Also see, on this challenge, Gamito (n 852) 16.

[854] See Section 2.1.1.1.

[855] In the opposite case, the provision would be impossible to implement. An *effet utile* interpretation thus hints at this interpretation. On the principle of effectiveness, see e.g., Case C-928/19 P *European Federation of Public Service Unions (EPSU) v European Commission* [2021] ECLI:EU:C:2021:656, para 38.

[856] See Section 2.1.1.2.

[857] See notes 852 through 855.

[858] See Section 2.1.1.

fully rely on such standards. It is likely that this will, in practice, lead to a continued role for codes of practice alongside some standards if and when those are developed, nuancing Article 55(2)'s textual implication that the role of codes of practice is limited to serve as a placeholder – a 'temporary tool'[859] – until if and when such a standard is published. Nevertheless, to the extent that a harmonised standard does cover an aspect that is equally covered by a code of practice, the standard clearly prevails.[860]

301. The relevant standards are to be created after a standardisation request by the European Commission[861] and should ideally be based on a 'balanced representation of interests involving all relevant stakeholders in the development of standards, in particular SMEs, consumer organisations and environmental and social stakeholders'.[862] This process thus entails similar stakeholders as for the development of codes of practice, as Article 56(3) indicates that '[t]he AI Office may invite all providers of general-purpose AI models, as well as relevant national competent authorities, to participate in the drawing up of codes of practice. Civil society organisations, industry, academia and other relevant stakeholders, such as downstream providers and independent experts, may support the process'.[863]

302. On their surface,[864] harmonised standards hold the key advantage over codes of practice that compliance with the former grants providers a presumption of conformity with Article 55(1), whereas adherence to the latter does not.[865] This presumption of conformity is likely rebuttable.[866] This difference, however, is, at least in part, nuanced for codes of practice that were deemed adequate by the AI Office and the Board,[867] because of the Commission's view that such codes can be used to 'demonstrate compliance with the obligations in Articles 53(1) and 55(1) AI Act'.[868]

303. Recital 117 indicates that the AI Office will assess the suitability of harmonised standards, implying that the presumption of conformity only applies if and when the AI Office has done so.[869] This is not reflected in the text of Articles 55 and 56 but, instead, likely[870] reflects the Article 40(1) and Regulation 1025/2012 requirement that harmonised standards undergo Commission (and standardisation body) assessment[871] before a reference to those standards is published in the Official Journal[872] and thus entail a presumption of conformity.

---

[859] Commission Guidelines (n 9) para 100. Also see Finck (n 36) para 6.84; Bernsteiner and Schmitt, 'Art 53' (n 848) para 62 (referred to in the chapter on art 55).

[860] This is clearly implied by article 55(3)'s wording that reliance on a code of practice is possible 'until a harmonised standard is published'.

[861] See Regulation 1025/2012 (n 849) arts 2(1)(c) and 10; AI Act, recital 121.

[862] AI Act, recital 121.

[863] See the commentary on Article 56 in this work.

[864] For a more elaborate discussion, see the commentary on Article 56 in this work.

[865] See AI Act, art 55(2). For a more elaborate discussion, see the commentary on Article 56, Section 2.7.1.1.1. in this work.

[866] Also see Bernsteiner and Schmitt, 'Art 53' (n 848) para 62.

[867] AI Act, art 56(6).

[868] See Commission Guidelines (n 9) para 94. See, in more detail, the commentary on Article 56 in this work.

[869] AI Act, recital 117 ['Once a harmonised standard is published and assessed as suitable to cover the relevant obligations by the AI Office, compliance with a European harmonised standard should grant providers the presumption of conformity.'].

[870] See AI Act, art 3(47) *in fine* ('references in this Regulation to the AI Office shall be construed as references to the Commission').

[871] Regulation 1025/2012 (n 849) art 10(5).

[872] ibid art 10(6).

2.2.2. Codes of practice

304. Absent a harmonised standard, Article 55(2) indicates that providers of GPAI models with systemic risk can rely on codes of practice to demonstrate compliance with Article 55 first paragraph's obligations until a harmonised standard is published. Article 55(2)'s last sentence clarifies that this necessitates the code to have been 'approved' – by the AI Office and the Board[873] – to serve this compliance function.[874]

305. It should be noted, as discussed earlier, that Article 55(2) does not extend the presumption of conformity that applies to providers adhering to a harmonised standard to apply to providers that rely on an approved code of practice.[875] As such, there is no general presumption that providers that adhere to an approved code of practice are compliant with the AI Act.[876] Nevertheless, this distinction appears to be largely theoretical,[877] as the Commission Guidelines state that providers 'can demonstrate compliance with the obligations in Articles 53(1) and 55(1) AI Act by adhering to a code of practice that is assessed as adequate by the AI Office and the Board'.[878]

306. Lastly, although it is not covered in detail here,[879] Article 56(6) also indicates that an approved code of practice can be given general validity by way of an implementing act.

307. It is clear that approved codes of practice offer important guidance when it comes to the interpretation of the AI Act's provisions, even if a provider did not sign on to that code of practice.[880] The Commission Guidelines indicate as much, by stating that the alternative adequate means adopted by such providers, discussed below, can be shown to result in AI Act compliance, 'for instance by carrying out a gap analysis that compares the measures they have implemented with the measures set out by a code of practice that is assessed as adequate.'[881] The Commission Guidelines also indicate that the 'Commission may take into account commitments implemented in line with a code of practice that is assessed as adequate as a mitigating factor when fixing the amount of fines, depending on the specific circumstances'.[882]

2.2.3. Alternative adequate means

308. Providers of GPAI models that present systemic risk are not *per se* required to adhere to an approved code of practice or harmonised standards.[883] While the situation is less clear for codes of practice that are given general validity by way of implementing act,[884] Article 55(2) recognises that

[873] AI Act, art 56(6).
[874] See more extensively the commentary on Article 56, Section 2.6. in this work.
[875] Also see Bernsteiner and Schmitt, 'Art 53' (n 848) para 62 (referred to in the chapter on Art. 55).
[876] Commission Guidelines (n 9) para 100. Also see commentary on Article 56, Section 2.7. in this work.
[877] Also see commentary on Article 56, Section 2.7. in this work.
[878] Commission Guidelines (n 9) para 94. Also see commentary on Article 56 in this work; Bernsteiner and Schmitt, 'Art 53' (n 848) (referred to in the chapter on Art. 55).
[879] See commentary on Article 56, Section 2.6.2. in this work.
[880] Also see Section 2.1.
[881] Commission Guidelines (n 9) para 95.
[882] Commission Guidelines (n 9) para 96. Also see Bernsteiner and Schmitt, 'Art 53' (n 848) para 62.
[883] See AI Act, art 55(2).
[884] See commentary on Article 56, Section 2.6.2. in this work.

providers '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'.

309. While the AI Act does not directly describe what 'alternative adequate means' entail, this notion clearly refers to a broad category of measures that providers might take to ensure and, perhaps more importantly, demonstrate[885] their compliance with Article 55(1). The latter is also evident from the Commission's position that providers who do not adhere to an approved code of practice (or a harmonised standard) are expected to report to the AI Office how the measures they have implemented ensure compliance with the AI Act,[886] though the AI Act does not contain an explicit obligation in this sense.

310. Despite the inherently broad nature of 'alternative adequate means', the Commission Guidelines do reflect the Commission's expectation that these means would be similar to those found in an approved code of practice.[887] More specifically, when reporting to the AI Office how providers comply with the AI Act through alternative adequate means, the EU Commission expects them to explain *how* their measures ensure compliance with the Act's obligations, listing, as an example, 'by carrying out a gap analysis that compares the measures they have implemented with the measures set out by a code of practice that is assessed as adequate.'[888]

311. The Commission Guidelines (thus) strongly imply that adherence to an approved code of practice – even when that code of practice was not signed on to by a specific provider – constitutes the most straightforward pathway to compliance. In this sense, they indicate that providers who do not adhere to an approved code may expect more requests for information and access,[889] as their choice not to rely on the approved code complicates the AI Office's compliance assessment.[890] Moreover, code of practice commitments that providers *do* adhere to are described as a mitigating factor in case of a potential fine for non-compliance, as discussed above.[891]

2.3. Article 55(3): Confidentiality

312. Article 55(3) requires that the Commission treat the information and documentation it obtains by virtue of Article 55 confidentially, in accordance with the obligations set out in Article 78. In this regard, Article 78(2) is particularly relevant. It requires the Commission to limit their requests to data that is 'strictly necessary' for the exercise of its powers under the AI Act. It also requires them to put in place adequate and effective cybersecurity measures to protect the security and confidentiality of the information and data they obtain, and to delete such data as soon as it is no longer required to assess compliance.

313. Before looking at some of these requirements in more detail, it is interesting to note that a direct consequence of this confidentiality requirement is that the general public will not have access to the

[885] Also see AI Act, art 55(2) ['Providers of general-purpose AI models with systemic risks 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'] (emphasis added).
[886] Commission Guidelines (n 9) para 95.
[887] Commission Guidelines (n 9) para 95 (discussing the potential need for a 'gap analysis'); see also, the commentary on Article 56, Section 2.7.1.2. in this work.
[888] Commission Guidelines (n 9) para 95.
[889] Also see Bernsteiner and Schmitt, 'Art 53' (n 848) para 63.
[890] Commission Guidelines (n 9) para 95.
[891] See Section 2.2.2; see also, the discussion in the commentary on Article 56, Section 2.7.1.2. in this work.

information that providers share with the AI Office. As such, they will, for example, not have access to submitted information about serious incidents.[892]

2.3.1. Strict necessity

314. First of all, Article 78(2) clarifies that the AI Office and Commission shall only request 'data' that is 'strictly necessary' for the exercise of their powers under the AI Act.[893] While some have argued that Article 91(1) contains a more specific applicable provision which would overrule Article 78(2) in the contexts of Articles 53 and 55,[894] that consideration is offset by the latter provisions' express reference to Article 78, strongly implying the applicability of the latter.[895] Article 91(1) empowers the Commission with a broader mandate,[896] allowing it to request all information 'necessary' for compliance assessments, whereas Article 78(2) only permits it to seek information that is 'strictly necessary' for the purposes of exercising the same compliance assessment powers and duties. An argument in favour of sustaining the superlative condition of 'strictly necessary' is the obvious sensitivity of the information the Commission is able to request under this power.[897]

315. The requirement of necessity – and the present requirement of strict necessity, more specifically – imposes a balancing exercise between the Commission's need for information to assess compliance with the AI Act, on the one hand, and various fundamental rights of the model provider, on the other hand. The latter include Article 17(2) of the Charter, which protects intellectual property (including trade secrets[898]), Article 7 of the Charter, which protects the right to privacy and which applies to legal persons as well,[899] and Article 16 of the Charter's protection of the freedom to conduct a business.[900] These fundamental rights imply that any interference within the sphere of private activities of a party is to be proportionate and deliberate.[901]

---

[892] AI Act, art 55(1)(c); see Section 2.1.3.

[893] Also see Bernsteiner and Schmitt, 'Art 55' (n 24) para 29, describing the reference as redundant because, according to them, article 78 would apply even without the reference.

[894] Finck (n 36) para 10.147. See differently (at least implicitly so): Bernsteiner and Schmitt, 'Art 55' (n 24) para 26.

[895] See, in support of this view, at least implicitly: Bernsteiner and Schmitt, 'Art 55' (n 24) para 29 (even describing the reference to article 78 as redundant as it would apply regardless).

[896] Also see Finck (n 36) para 10.147.

[897] Also see Bernsteiner and Schmitt, 'Art 55' (n 24) para 26.

[898] See, in particular, 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 [2016] OJ L 157/1.

[899] Joined Cases 46/87 and 227/88 *Hoechst AG v Commission of the European Communities* [1989] ECLI:EU:C:1989:337, para 19 (discussed as a general principle of Community law); Case C-94/00 *Roquette Frères SA v Directeur général de la concurrence, de la consommation et de la répression des fraudes and Commission of the European Communities* [2002] ECLI:EU:C:2002:603, para 27 (on European Convention for the Protection of Human Rights and Fundamental Freedoms, art 8, though describing this as a general principle of Community law); Case C-450/06 *Varec SA v Belgian State* [2008] ECLI:EU:C:2008:91, para 48; Case C-583/13 P *Deutsche Bahn AG and Others v European Commission* [2015] ECLI:EU:C:2015:404, para 19 (focusing on the inviolability of the home).

[900] On which, e.g., Case C-201/15 *Anonymi Geniki Etairia Tsimenton Iraklis (AGET Iraklis) v Ypourgos Ergasias, Koinonikis Asfalisis kai Koinonikis Allilengyis* [2016] ECLI:EU:C:2016:972 para 66 and 79 ff.

[901] Tobias Lock, 'Article 52 CFR' in Manuel Kellerbauer, Marcus Klamert and Jonathan Tomkin (eds), *The EU Treaties and Charter of Fundamental Rights: A Commentary* (Oxford University Press 2024) 609 ff. For privacy, e.g., Joined Cases 46/87 and 227/88 *Hoechst AG v Commission of the European Communities* [1989] ECLI:EU:C:1989:372 para 19 (discussed as a general principle of Community law); *Frères v Commission* (n 899) para 27 (on Art. 8 European Convention for the Protection of Human Rights and Fundamental Freedoms). For the

316. The term 'strictly necessary' indicates that the proportionality test must be applied with a higher level of scrutiny than when the standard is merely 'necessary'. Where 'necessary' already implies that the information requested is required to pursue the AI Act's aims and that there is no less restrictive equally effective alternative available,[902] 'strictly necessary' goes beyond that and implies that the Commission and AI Office enjoy a narrower margin of discretion. There should thus be some particular circumstance that supports and warrants the Commission and AI Office's request – extending beyond the general idea that more information might help the Commission and AI Office better exercise their mandate.[903]

317. Resultantly, the phrase 'strictly necessary' implies an important procedural consequence. On this basis, the Commission and AI Office can only make a *reasoned* request to exercise this option.[904] That request should thus indicate the reasons why the relevant information is strictly necessary.[905]

2.3.2. Cybersecurity

318. Article 78(2) requires 'authorities' – the Commission and the AI Offices in particular, in this context – to 'put in place adequate and effective cybersecurity measures to protect the security and confidentiality of the information and data obtained'.[906] The need for cybersecurity is a natural corollary of the need for confidentiality; the latter being a natural implication of the general principle of the protection of business secrets.[907]

319. The AI Act's cybersecurity requirement mimics similar requirements found in other European legislation, such as under Regulation (EU) 2018/1725[908] which governs personal data processing by Union institutions, bodies, offices and agencies. In its Article 36, the latter imposes confidentiality and related requirements. Its Article 32 imposes security requirements that, *inter alia*, relate to the confidentiality and integrity of processing systems and services. It is also noteworthy to signal that the latter is applicable in the context of the AI Act, if personal data is involved.[909] Other relevant instruments are Directive (EU) 2022/2555,[910] and, less directly, Regulation (EU, Euratom)

freedom to conduct a business, e.g. *AGET Iraklis v Ypourgos Ergasias, Koinonikis Asfalisis kai Koinonikis Allilengyis* (n 900) paras 79 ff.

[902] In the context of (the) fundamental rights (concerned): EU Charter, art 52(1); Lock (n 901) 611–612.

[903] See in the same sense: Finck (n 36) para 10.147 (by describing how the request should set out such reasons). See, in the context of GDPR: Case C-205/21 *V.S v Ministerstvo na vatreshnite raboti* [2023] ECLI:EU:C:2023:49 paras 117–118, 125 and 135.

[904] Meaning of 'reasoned'.

[905] Finck (n 36) para 10.147.

[906] Article 78(2) thus refers to 'adequate and effective' measures, whereas article 55(1)(d) only requires 'adequate' protection, see Section 2.1.4.6.

[907] Case 53/85 *AKZO Chemie BV and AKZO Chemie UK Ltd v Commission of the European Communities* [1986] ECR I-1965, para 28; Case C-36/92 P *Samenwerkende Elektriciteits-Produktiebedrijven NV (SEP) v Commission of the European Communities* [1994] ECR I-01911, paras 36–37. E.g. on the relationship with confidentiality, Case C-450/06 *Varec SA v Belgian State* [2008] ECLI:EU:C:2008:91, paras 49–51.

[908] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC [2018] OJ L 295/39.

[909] AI Act, art 2(7).

[910] NIS2 (n 425).

2023/2841,[911] which impose cybersecurity requirements on public and private entities and on Union institutions, bodies, offices and agencies, respectively.

320. The comparison with these instruments – which, while useful, has clear limitations, for example due to the different contexts and objectives of some of them, such as Regulation 2018/1725, when compared to the AI Act[912] – is particularly interesting as some of their more explicit requirements offer more context that can help interpret the similar obligation under the AI Act. Most notably, various of these instruments indicate that the cybersecurity measures at hand should live up to the state of the art[913] – which is to be understood as referring to the 'generally acknowledged state of the art'.[914] In this respect, ISO 27001 is particularly interesting (as well as ISO 27002), though not directly applicable, as this sets out the relevant standards for information security management systems.[915]

[911] Regulation (EU, Euratom) 2023/2841 of the European Parliament and of the Council of 13 December 2023 laying down measures for a high common level of cybersecurity at the institutions, bodies, offices and agencies of the Union [2023] OJ L 2841/1.

[912] Also see the forthcoming chapter on Interpreting the AI Act through Systematic Analogies in this work.

[913] E.g., Regulation (EU) 2018/1725 (n 908) art 33 ['appropriate technical and organisational' measures should take into account the state of the art]; NIS2 (n 425) art 21(1); Regulation 2023/2841 (n 911) art 8(1).

[914] See Section 2.1.1. Also see NIS2 (n 425) art 21(1) (which refers to 'state-of-the-art' and 'relevant European and international standards' alongside each other).

[915] See ISO and IEC, 'Information Security, Cybersecurity and Privacy Protection — Information Security Management Systems — Requirements' (ISO and IEC 2022) ISO/IEC 27001:2022 <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27001:ed-3:v1:en> accessed 20 May 2026.