Defining AI Models and AI Systems: A Framework to Resolve the Boundary Problem

Yuanyuan Sun [1] **Timothy Parker** [2] **Lara Gierschmann** [2] **Sana Shams** [3] **Teo Canmetin** [4] **Mathieu Duteil** [2] **Rokas Gipiškis** [5 6] **Ze Shen Chin** [5 7]

Abstract

Emerging regulations assign distinct obligations to different parties along the AI value chain. For example, The EU AI Act differentiates between general-purpose AI (GPAI) models versus AI systems, and their providers or deployers have distinct obligations. Nevertheless, the term *AI model* is itself undefined in the Act, even though it is used to define *GPAI model*. This gap creates regulatory ambiguities, such as when and how an entity acquires provider obligations by modifying a model. Drawing on a systematic review of 896 papers and a manual review of regulatory documents across major jurisdictions (included in the Appendices), we propose conceptual and operational definitions of the terms *AI model* and *AI system* for contemporary neural network-based AI.

1. Introduction

The EU AI Act assigns distinct responsibilities to providers and deployers of both general-purpose AI (GPAI) models and providers of AI systems (European Parliament and Council of the European Union, 2024). The Act defines *AI system* (Article 3(1)) and *general-purpose AI model* (Article 3(63)), but it does not define *AI model* itself, the concept upon which *GPAI model* is defined. This gap becomes consequential because the allocation of some regulatory obligations by the AI Act hinges on the question whether an entity has *modified* an earlier model into a new model. If an entity modifies the system prompt that a GPAI model was shipped with, but without changing any of the model weights, this can substantially modify the model behavior. But does it count under the AI Act as a modification of the model itself? This ambiguity is not resolved by the *Guidelines on the scope of obligations for providers of GPAI models under the AI Act* (henceforth *GPAI Guidelines*) (European Commission, 2025b), which mainly concerns itself with the case of training runs that modify weights. A principled solution would require first defining exactly *where the boundary lies between an AI model and the surrounding AI system components*, then assessing if any of the AI model components were modified, and then determining whether the modification is significant enough to produce a new model.

We now turn to the question if existing definitions achieve this clarity.

Through a systematic literature review of 896 academic papers and a manual review of regulatory documents across major jurisdictions (methodology in Appendix A), we analysed different conceptual perspectives of *AI model* and *AI system*, and the relationship between the two (Appendix B.1). We also find that most regulatory definitions trace back to the OECD's AI frameworks (OECD, 2019b; 2022; 2024), themselves adapted from Russell & Norvig (2009). For reference, the closest existing OECD characterisation of an AI model is as follows. (The OECD AI system definition is reproduced in Appendix C.3.)

OECD on AI model. "An AI model is a core component of an AI system used to make inferences from inputs to produce outputs. [...] [W]hile the parameters of an AI model change during the build phase, they usually remain fixed after deployment" (OECD, 2024, p. 8).

We found that none of the existing definitions adequately address the model-system boundary problem. Even the OECD, whose frameworks represent perhaps the most consensus-driven effort at developing AI terminologies, uses *AI model* inconsistently across its documents, applying the term to world models of an model-based agent, representations of external environment, components of an AI system, abstract input-output mappings, and policies of a model-free agent; without reconciling these perspectives (OECD, 2024) (Appendix B.3.2).

[1]AI Governance Exchange [2]Independent [3]University of British Columbia [4]Oxford Internet Institute, University of Oxford [5]AI Standards Lab [6]Vilnius University [7]Oxford Martin AI Governance Initiative. Correspondence to: Ze Shen Chin <zeshen@aistandardslab.org>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

2. Proposed Disambiguating Definitions

Drawing inspiration from Tomić & Štimac (2025), we propose both *conceptual* and and *operational* definitions. The conceptual ones are intended to remain stable; the operational ones give more details that disambiguate boundaries in neural network based AI system, but they might need to be updated as architectures evolve. The process for developing these definitions are described in Appendix C.

Our proposed definitions are as follows.

AI model (conceptual). A computational construct that makes inferences from inputs to produce outputs and forms (or is intended to form) the reasoning or knowledge core of an AI system.

AI system (conceptual). A machine-based system that makes inferences from inputs from physical or virtual environments to produce outputs that influence physical or virtual environments.

AI model (operational). Trained parameters and their architecture, understood as the computational specification necessary to use them for inference.

AI system (operational). One or more AI models, an interface for receiving inputs from and delivering outputs to an environment, and the configuration connecting these and any additional components.

Our conceptual definition of *AI model* treats the model as a core reasoning component rather than a "representation" of external structure, reflecting the shift in usage since the dominance of neural network-based systems (Appendix B.2.1). For the operational definition, the OECD characterises *AI model* using two conditions: the parameters change during the build phase, and they usually remain fixed after deployment. We compress these two into a single concise criterion, *trained parameters*, referring to them as mathematical objects that can be changed by the gradient descent process during training (LeCun et al., 2015). Architecture is included in the operational definition of AI model as it specifies the relations between the trained parameters necessary to perform inference. Since our operational definition of AI systems includes AI models, it follows that any element which is part of the model is necessarily part of the system, consistent with the OECD's usage which states "an AI system is typically built by combining one or more models [...]" (OECD, 2024, p. 8). While this operational definition may appear straightforward to machine learning practitioners, the term *model*, even as defined by ISO standards on AI concepts and terminology ISO/IEC (2022b), is based on a representational sense, and other coherent approaches (e.g. behavioral or capability equivalence, functional role) may result in different boundaries. Our choice to anchor in trained parameters was made in part to build on consensus formed through the OECD documents, and in part because it yields a clear boundary test.

Table 1. Classification of components as model elements.

Component	Trained parameters?	Part of model?
Weights & biases	Yes	**Yes**
Architecture	N/A[†]	**Yes**[†]
LoRA adapter	Yes	**Yes**
I/O filters	Depends	**Depends**[◇]
RAG database	No	No
System prompt	No	No
Input prompt	No	No
Activations	No	No
Temperature	No	No

[†]Required for trained parameters to perform inference. [◇]A filter consisting of trained parameters constitutes part of a model; a rule-based one does not.

Table 1 applies this test to aspects of a transformer-based large language model (LLM) deployment (Appendix C.3).

3. Regulatory Implications

Under our proposed operational definitions, modifications that alter trained parameters (e.g. fine-tuning, attaching a Low-Rank Adaptation (LoRA) adapter (Hu et al., 2021)) are model modifications, while changes to non-parameter components (e.g. adding retrieval augmented generation (RAG) (Lewis et al., 2021), changing the system prompt) are system modifications. This resolves the currently unresolved question in existing EU regulations, namely *what* is being modified. Clarifying this lets regulators then address *when* a modification is substantial enough to shift obligations from the upstream provider to the downstream modifier. We recommend that future revisions to the *GPAI Guidelines* explicitly incorporate such definitions as well as disambiguating clarifications as per Table 1.

The relevance of clear definitions extend beyond the EU. Any jurisdiction allocating obligations along an AI value chain must draw clear lines between the upstream providers and downstream deployers. While our operational definitions provide a principled default for doing so, another option is to allow model providers to declare in their documentation what they consider the boundaries of their model to be, as long as clear boundaries exists for each product placed on the market for obligations and responsibilities to be reliably allocated. (For case studies where the lack of clear definitions on model-system boundaries may have led to regulatory uncertainties, see Appendix C.6.)

Impact Statement

This paper proposes terminological definitions intended to reduce ambiguity in regulatory frameworks governing AI. We do not foresee specific negative societal consequences beyond those inherent in any ongoing regulatory discourse.

LLM Usage Statement

Large language models (LLMs) were used to assist with the literature search, for general background research, for editing and language polishing, and for reviewing drafts. All references surfaced by LLMs were independently retrieved, read, and verified before citing.

Acknowledgements

This work was carried out through the Supervised Program for Alignment Research (SPAR), whose program structure and infrastructure support made the research possible. We thank Koen Holtman, Antonio Puertas Gallardo, Yi-Yang Chua, Zifeng Wang, and Adrian Regenfuß for helpful comments on earlier drafts. All errors remain our own.

References

ACLU of Illinois. In Big Win, Settlement Ensures Clearview AI Complies With Groundbreaking Illinois Biometric Privacy Law. ACLU of Illinois Press Release, may 2022. URL https://www.aclu-il.org/press-releases/big-win-settlement-ensures-clearview-ai-complies-groundbreaking-illinois-biometric/. Accessed: 2026-02-21.

American Psychological Association. Operational definition. APA Dictionary of Psychology, 2018. URL https://dictionary.apa.org/operational-definition. Accessed: February 21, 2026.

Anthropic. Model card addendum: Claude 3.5 Haiku and upgraded Claude 3.5 Sonnet. Technical report, Anthropic, oct 2024. URL https://www-cdn.anthropic.com/c7822cdc35ad788ec87e14b3a9d45010f1f86c38.pdf.

Anthropic. Claude 3.7 sonnet system card. Technical report, Anthropic, feb 2025. URL https://www-cdn.anthropic.com/9ff93dfa8f445c932415d335c88852ef47f1201e.pdf.

Bailey, D. and Norbury, L. Clarity in legislation. *Statute Law Review*, 46(3):hmaf035, sep 2025. ISSN 0144-3593, 1464-3863. doi: 10.1093/slr/hmaf035. URL https://doi.org/10.1093/slr/hmaf035.

Birnhack, M. D. Reverse Engineering Informational Privacy Law. *SSRN Electronic Journal*, feb 2012. ISSN 1556-5068. doi: 10.2139/ssrn.2002757. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2002757.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the Opportunities and Risks of Foundation Models, jul 2022. URL http://arxiv.org/abs/2108.07258.

Braun, M. Italy's privacy regulator takes action against ChatGPT. WilmerHale Privacy and Cybersecurity Law Blog, apr 2023. URL https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20230403-italys-privacy-regulator-takes-action-against-chatgpt. Accessed: 2026-02-14.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pp. 1877–1901, Red Hook, NY, USA, dec 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6. URL https://dl.acm.org/doi/10.5555/3495724.3495883.

Burden, J., Pacchiardi, L., Martínez, P. F., and Hernández, O. J. *A Framework for General-Purpose AI Model Cate-*

gorisation. Publications Office of the European Union, oct 2025. ISBN 978-92-68-31431-9. doi: 10.2760/5330387. URL https://publications.jrc.ec.europa.eu/repository/handle/JRC143256. JRC number: JRC143256.

Burgess, A., Cappelen, H., and Plunkett, D. (eds.). *Conceptual Engineering and Conceptual Ethics*. Oxford University Press, Oxford, UK, 2020. ISBN 978-0-19-880185-6.

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., and Lasa, G. How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9: 101895, jan 2022. ISSN 2215-0161. doi: 10.1016/j.mex.2022.101895. URL https://www.sciencedirect.com/science/article/pii/S2215016122002746.

Center for Humane Technology. Ai in society, 2024. URL https://www.humanetech.com/ai-society. Accessed: 2026-02-14.

Chin, Z. S. and Holtman, K. White paper: GPAI model providers and modifiers and their obligations under the AI act. White paper, AI Standards Lab, may 2025. URL https://aistandardslab.org/white-paper-gpai-model-providers-and-modifiers-and-their-obligations-under-the-ai-act/. Accessed: February 21, 2026.

Clearview AI. Clearview ai settles ACLU illinois lawsuit, confirming continuity of business supporting public safety. Clearview AI Press Room, may 2022. URL https://www.clearview.ai/press-room/clearview-ai-settles-aclu-illinois-lawsuit-confirming-continuity-of-business-supporting-public-safety. Accessed: 2026-02-14.

Copi, I. M., Cohen, C., and MacMahon, K. *Introduction to Logic*. Pearson custom library. Pearson Education Limited, Harlow, 14 ed., pearson new international ed edition, 2014. ISBN 978-1-292-02482-0.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, dec 1989. ISSN 0932-4194, 1435-568X. doi: 10.1007/BF02551274. URL http://link.springer.com/10.1007/BF02551274.

DeLancey, C. *A Concise Introduction to Logic*. Open SUNY Textbooks, mar 2017. ISBN 978-1-942341-42-0. URL https://milnepublishing.geneseo.edu/concise-introduction-to-logic/.

Dhar, V. The Paradigm Shifts in Artificial Intelligence, aug 2023. URL http://arxiv.org/abs/2308.02558.

Dickerson, F. R. *Legislative drafting*. Greenwood Press, Westport, Conn, 1977. ISBN 978-0-8371-9688-6.

Dunietz, J., Hasselbalch, G., Orssich, I., Renda, A., Schwartz, R., and Tabassi, E. EU-U.S. Terminology and Taxonomy for Artificial Intelligence. Technical report, European Commission and National Institute of Standards and Technology (NIST), apr 2024. URL https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence-second-edition. EU-U.S. Trade and Technology Council (TTC) Working Group 1: Technology Standards.

Ebers, M. No. 101: Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act. Working paper, Stanford-Vienna Transatlantic Technology Law Forum, 2024. URL https://law.stanford.edu/publications/no-101-truly-risk-based-regulation-of-artificial-intelligence-how-to-implement-the-eus-ai-act/.

Estévez Almenzar, M., Fernández Llorca, D., Gómez Gutiérrez, E., and Martínez Plumed, F. Glossary of human-centric artificial intelligence. JRC Science for Policy Report EUR 31113 EN, Joint Research Centre (European Commission), Luxembourg, 2022. URL https://publications.jrc.ec.europa.eu/repository/handle/JRC129614.

European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (AI Act) and amending certain Union legislative acts. COM/2021/206 final, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. Accessed: 2026-02-13.

European Commission. Guidelines on the scope of obligations of providers of general-purpose AI models under the AI Act. European Commission, jul 2025a. URL https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act. Annex to the communication to the Commission. Accessed: 2026-02-13.

European Commission. Commission Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act), jul 2025b. URL https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application. Communication from the Commission.

European Data Protection Supervisor. Guidance for Risk Management of Artificial Intelligence systems, nov 2025. URL https://www.edps.europa.eu/system/files/2025-11/2025-11-11_ai_risks_management_guidance_en.pdf.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on Artificial Intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), jul 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. Official Journal of the European Union L 2024/1689. Adopted: 2024-06-13. Published: 2024-07-12.

Executive Office of the President. Safe, secure, and trustworthy development and use of artificial intelligence. Federal Register, Vol. 88, No. 210, 75191–75226, nov 2023. URL https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Executive Order 14110.

Fernández-Llorca, D., Gómez, E., Sánchez, I., and Mazzini, G. An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. *Artificial Intelligence and Law*, 33(4):875–888, dec 2025. ISSN 0924-8463, 1572-8382. doi: 10.1007/s10506-024-09412-y. URL https://link.springer.com/10.1007/s10506-024-09412-y.

French, R. M. Moving beyond the Turing test. *Communications of the ACM*, 55(12):74–77, December 2012. ISSN 0001-0782. doi: 10.1145/2380656.2380674. URL https://dl.acm.org/doi/10.1145/2380656.2380674.

Future of Life Institute. General Purpose AI and the AI Act. Technical report, Future of Life Institute, may 2022. URL https://futureoflife.org/document/general-purpose-ai-and-the-ai-act/.

Ghosh, D., Gupta, A., and Levine, S. Learning Actionable Representations with Goal-Conditioned Policies, 2018. URL https://arxiv.org/abs/1811.07819. Version Number: 2.

Google DeepMind. Gemini 3 pro model card. Technical report, Google DeepMind, dec 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Greenberg, B. A. Rethinking technology neutrality. *Minnesota Law Review*, 100:1495–1562, 2016. URL https://www.minnesotalawreview.org/wp-content/uploads/2016/04/Greenberg_ONLINEPDF.pdf.

Gupta, A. and Mackereth, S. Definitions. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition, 2023. URL https://plato.stanford.edu/archives/fall2023/entries/definitions/.

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C., and Franklin, M. A Proposal for a Definition of General Purpose Artificial Intelligence Systems, oct 2022. URL https://papers.ssrn.com/abstract=4238951.

Hacker, P. and Holweg, M. The Regulation of Fine-Tuning: Federated Compliance for Modified General-Purpose AI Models, jun 2025. URL https://papers.ssrn.com/abstract=5289125.

Helle, P., Schamai, W., and Strobel, C. Testing of Autonomous Systems – Challenges and Current State-of-the-Art. *INCOSE International Symposium*, 26(1):571–584, jul 2016. ISSN 2334-5837, 2334-5837. doi: 10.1002/j.2334-5837.2016.00179.x. URL https://incose.onlinelibrary.wiley.com/doi/10.1002/j.2334-5837.2016.00179.x.

High-Level Expert Group on Artificial Intelligence. A Definition of AI: Main Capabilities and Disciplines. Technical report, European Commission, apr 2019a. URL https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. Technical report, European Commission, apr 2019b. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Holtman, K. and Chin, Z. S. An analysis of the GPAI model guidelines published by the European Commission. Technical report, AI Standards Lab, jul 2025. URL https://aistandardslab.org/gpai_guidelines_analysis/. Accessed: 2026-02-21.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, jan 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

House Legislative Services. *Drafting Style and Usage Manual*. Louisiana House of Representatives, Baton Rouge, LA, nov 2015. URL https://documents.ncsl.org/wwwncsl/LegislativeStaff/RELACS/2016LAHouseDraftingManual.pdf.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Hupont, I., Fernández-Llorca, D., Baldassarri, S., and Gómez, E. Use case cards: a use case reporting framework inspired by the European AI Act. *Ethics and Information Technology*, 26(2):19, jun 2024. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-024-09757-7. URL https://link.springer.com/10.1007/s10676-024-09757-7.

ISO. *ISO 704:2022. Terminology work — Principles and methods*. International Organization for Standardization, Geneva, Switzerland, 4 edition, jul 2022. URL https://www.iso.org/standard/79077.html.

ISO/IEC. *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*. International Organization for Standardization, Geneva, Switzerland, 2022a. URL https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en. Standard number: ISO/IEC 22989:2022(en).

ISO/IEC. *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. model*. International Organization for Standardization, Geneva, Switzerland, 2022b. URL https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en:term:3.1.23. Standard number: ISO/IEC 22989:2022(en). Term: 3.1.23.

ISO/IEC. *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. symbolic AI*. International Organization for Standardization, Geneva, Switzerland, 2022c. URL https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en:term:3.1.33. Standard number: ISO/IEC 22989:2022(en). Term: 3.1.33.

ISO/IEC. *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. artificial intelligence system*. International Organization for Standardization, Geneva, Switzerland, 2022d. URL https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en:term:3.1.4. Standard number: ISO/IEC 22989:2022(en). Term: 3.1.4.

ISO/IEC. *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. machine learning*. International Organization for Standardization, Geneva, Switzerland, 2022e. URL https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en:term:3.3.5. Standard number: ISO/IEC 22989:2022(en). Term: 3.3.5.

Jopek-Bosiacka, A. Defining Law Terms: A Cross-Cultural Perspective. *Research in Language*, 9(1):9–29, jun 2011. ISSN 1731-7533. doi: 10.2478/v10015-011-0008-y. URL https://czasopisma.uni.lodz.pl/research/article/view/8122.

Jung, W. K. and Kwon, H. Y. Privacy and data protection regulations for AI using publicly available data: Clearview AI case. In *Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance*, ICEGOV 2024, pp. 48–55, Pretoria, South Africa, oct 2024. ACM. ISBN 9798400717802. doi: 10.1145/3680127.3680200. URL https://doi.org/10.1145/3680127.3680200.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Laher, Z. Openai faces €15 million fine as the italian garante strikes again. Lewis Silkin Insights, jan 2025. URL https://www.lewissilkin.com/insights/2025/01/14/openai-faces-15-million-fine-as-the-italian-garante-strikes-again-102jtqc. Accessed: February 14, 2026.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, may 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Li, Y. and Zhan, J. SAIBench: Benchmarking AI for Science. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2(2):100063, apr 2022. ISSN 27724859. doi: 10.1016/j.tbench.2022.100063. URL https://linkinghub.elsevier.com/retrieve/pii/S2772485922000503.

Margolis, E. and Laurence, S. Concepts. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition, 2023. URL https://plato.stanford.edu/archives/fall2023/entries/concepts/.

McCulloch, W. S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, dec 1943. ISSN 0007-4985, 1522-9602. doi: 10.1007/BF02478259. URL http://link.springer.com/10.1007/BF02478259.

Nahra, K., Evers, A., Jessani, A., Braun, M., Vallery, A., and Benizri, I. The European Parliament adopts the AI Act. WilmerHale Privacy and Cybersecurity Law Blog, mar 2024. URL https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20240314-the-european-parliament-adopts-the-ai-act. Accessed: 2026-02-14.

National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, jan 2023. URL https://www.nist.gov/itl/ai-risk-management-framework.

Nessler, B. and Wendehorst, C. Commission guidelines on the application of the definition of an AI system and the prohibited AI practices established in the AI act: Response of the European Law Institute. Technical Report RC-2024-9, European Law Institute, Vienna, Austria, dec 2024. URL https://www.europeanlawinstitute.eu/news-events/news-contd/news/eli-submits-a-response-to-the-european-commissions-public-consultation-on-the-ai-act/.

OECD. Recommendation of the Council on Artificial Intelligence. Technical Report OECD/LEGAL/0449, Organization for Economic Co-operation and Development, 2019a. URL https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449. Adopted on 22 May 2019; amended on 3 May 2024.

OECD. Scoping the OECD AI principles: Deliberations of the expert group on artificial intelligence at the OECD (AIGO). Technical Report 291, Organisation for Economic Co-operation and Development, nov 2019b. URL https://www.oecd.org/en/publications/scoping-the-oecd-ai-principles_d62f618a-en.html. Reference code: DSTI/CDEP(2019)1/FINAL.

OECD. OECD Framework for the Classification of AI systems. *OECD Digital Economy Papers*, feb 2022. doi: 10.1787/cb6d9eca-en. URL https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html.

OECD. Explanatory memorandum on the updated OECD definition of an AI system. Technical Report 8, OECD Publishing, Paris, mar 2024. URL https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html. Reference code: DSTI/CDEP/AIGO(2023)8/FINAL.

OpenAI. Gpt-3 model card. GitHub Repository, sep 2020. URL https://github.com/openai/gpt-3/blob/master/model-card.md. Technical documentation for the GPT-3 language model.

OpenAI. GPT-4 system card. Technical report, OpenAI, mar 2023. URL https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Oxford English Dictionary. model, n. & adj., n.d. URL https://www.oed.com/dictionary/model_n. Accessed 14 February 2026.

Pacchiardi, L., Burden, J., Martínez-Plumed, F., and Hernández-Orallo, J. *A Framework to Categorise Modified General-Purpose AI Models as New Models Based on Behavioural Changes*. Publications Office of the European Union, oct 2025. ISBN 978-92-68-31545-3. doi: 10.2760/4372557. URL https://publications.jrc.ec.europa.eu/repository/handle/JRC143257. JRC number: JRC143257.

Ramanathan, A., Pullum, L. L., Hussain, F., Chakrabarty, D., and Jha, S. K. Integrating symbolic and statistical methods for testing intelligent systems: Applications to machine learning and computer vision. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 786–791, mar 2016. URL https://ieeexplore.ieee.org/document/7459413. ISSN: 1558-1101.

Reed, C. Taking Sides on Technology Neutrality. *SCRIPTed*, 4(3):263–284, sep 2007. ISSN 1744-2567. doi: 10.2966/scrip.040307.263. URL http://www.law.ed.ac.uk/ahrc/script-ed/vol4-3/reed.asp.

Renda, A. and Engler, A. What's in a Name? Getting the definition of Artificial Intelligence right in the EU's AI Act. Technical Report 2023-02, Centre for European Policy Studies, Brussels, 2023. URL https://www.ceps.eu/ceps-publications/whats-in-a-name/. CEPS Explainer.

Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Pearson, Upper Saddle River, NJ, USA, third edition, dec 2009. ISBN 9780136042594.

Sapkota, R., Roumeliotis, K. I., and Karkee, M. AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges, may 2025. URL https://arxiv.org/abs/2505.10468v5.

Schuett, J. Defining the scope of AI regulations, aug 2021. URL https://papers.ssrn.com/abstract=3453632.

Sharkey, L., Ní Ghuidhir, C., Braun, D., Scheurer, J., Balesni, M., Bushnaq, L., Stix, C., and Hobbhahn, M. A Causal Framework for AI Regulation and Auditing, jan 2024. URL https://www.preprints.org/manuscript/202401.1424/v1.

Tomić, S. and Štimac, V. Pinning down an octopus: towards an operational definition of AI systems in the EU AI Act. *Journal of European Public Policy*, pp. 1–36, aug 2025. ISSN 1350-1763, 1466-4429. doi: 10.1080/13501763.2025.2534648. URL https://www.tandfonline.com/doi/full/10.1080/13501763.2025.2534648.

Turing, A. M. Computing Machinery and Intelligence. *Mind*, 59(236):433–460, oct 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://academic.oup.com/mind/article-abstract/LIX/236/433/986238.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. Technical Report SHS/BIO/PI/2021/1, United Nations Educational, Scientific and Cultural Organization, Paris, France, 2022. URL https://unesdoc.unesco.org/ark:/48223/pf0000381137. Adopted: 2021-11-23.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention Is All You Need, jun 2017. URL http://arxiv.org/abs/1706.03762.

Volkov, D. Badllama 3: removing safety finetuning from Llama 3 in minutes, jul 2024. URL http://arxiv.org/abs/2407.01376.

Zhang, P. Jailed Chinese AI chatbot developers appeal landmark pornography case. *South China Morning Post*, jan 2026. URL https://www.scmp.com/news/china/politics/article/3339908/jailed-chinese-ai-chatbot-developers-appeal-landmark-pornography-case. Accessed: 2026-02-11.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A Survey of Large Language Models, mar 2025. URL http://arxiv.org/abs/2303.18223.

A. Methods

First, the literature review was conducted in two ways: a manual review and a systematic literature review (SLR). Completing a manual review of the literature ensured that important papers in the legislative context were included, while the SLR allowed for the inclusion of academic findings and discussions on existing definitions, capturing both well-established definitions as well as variations across more domain-specific research. Next, we analysed the conceptual perspectives of these definitions, and traced how past sources have influenced later definitions. Finally, we outline the approach taken to propose our own definitions.

A.1. Overview of Methods

A.1.1. Manual Review of Global Regulatory and Standards Documents

Alongside the SLR, we conducted a manual review of how "AI system" and "AI model" are defined across government documents, regulations and policy documents. These include key documents related to the jurisdictions of the EU, United States, and China — published up to October 2025. Chinese-language documents were assessed and translated by a native Mandarin speaker with expertise in policy terminology, reducing ambiguities associated with machine translation. Because relevant materials are decentralised and fragmented across various internet archives, websites, repositories, and legal publishers, we opted for manual search rather than automation when reviewing texts. Keywords such as "definition", "glossary", "system", "model" were used to identify relevant sections across documents. These sections were then screened against our inclusion and exclusion criteria, similar to that being used for the systematic literature review as described in Section A.1.2 (Systematic Literature Review). For example, the use of the word "model" outside of contexts related to AI were excluded. We grouped institutions into four categories: (1) national/state-level government bodies, (2) intergovernmental organisations, (3) standards bodies, and (4) NGOs. Verbatim definitions were extracted and tabulated with source, date, and organisational category.

A.1.2. Systematic Literature Review

To identify how *AI model* and *AI system* are defined in academic literature, we conducted a systematic literature review following Carrera-Rivera et al. (2022). We searched Scopus, Web of Science, and IEEE Xplore for papers from 2012 onwards, marking the beginning of the modern deep learning era with AlexNet (Krizhevsky et al., 2012). To ensure papers actually discussed definitions rather than merely using these terms, we employed proximity operators (NEAR/4 or W/4) requiring definitional terms ("definition", "ontology", "taxonomy", etc.) to appear within four words of "AI model", "AI system", or their synonyms. After excluding papers focused on domain-specific applications (e.g., medical diagnosis, educational contexts) to maintain focus on general conceptual definitions, 896 papers remained after deduplication. Papers were then assessed for definition clarity and completeness, with particular attention to whether definitions were original or cited from other sources. Full details including search strings, PICOC criteria, and complete inclusion/exclusion criteria are provided in Appendix A.2 Overview of the Systematic Literature Review.

A.1.3. Analysis of the Definitions

Using the definitions compiled from the systematic literature review and the manual review, we analysed the different conceptual perspectives of these definitions. We identified three key dimensions along which definitions vary: (i) parent categories (subordinate concept class), (ii) features (production, function, and for systems, mechanism), and (iii) relationships between models and systems. We also traced citation patterns to identify influential sources and analysed how past definitions have influenced recent regulatory frameworks, particularly the EU AI Act.

A.1.4. Development of Proposed Definitions

We developed evaluation criteria for proposed definitions by drawing on principles from ISO 704:2022 (technical concept definition) (ISO, 2022), legal drafting scholarship, and AI governance literature. These sources informed criteria related to structural clarity, technical coherence, and legal operability. Using these criteria, we developed both conceptual definitions (clarifying fundamental nature and relationships) and operational definitions (enabling practical regulatory application). For operational definitions, we specifically addressed contemporary neural network-based machine learning AI, where we systematically analysed aspects of a transformer-based LLM (e.g. weights, prompts, activations, RAG, fine-tuning adapters, filters, etc.) to establish if these aspects are part of a model or a system.

A.2. Overview of the Systematic Literature Review

A systematic literature review was conducted to obtain definitions of the terms *AI model* and *AI system*. The proposed structure by Carrera-Rivera et al. (2022) regarding the planning and conducting of a systematic literature review was followed. Specifically, we adopted the planning-phase steps outlined in their framework, adapted to the present topic, including defining the research objective using the Population, Intervention, Comparison, Outcome, and Context (PICOC) structure (Carrera-Rivera et al., 2022), selecting databases, establishing inclusion and exclusion criteria, conducting quality assessment, and performing data extraction.

Research Objective

Carrera-Rivera et al. (2022) suggest the PICOC structure to define the research objective and formulate search terms. Table 2 provides an overview of the keywords found using the PICOC criteria. Based on this breakdown, the following research question was formulated: How are the terms *AI model* and *AI system* defined?

Table 2. SLR Keywords from PICOC Criteria

Criterion	**Keywords**
Population	AI model and AI system *Synonyms:* Machine learning system, autonomous system, automated system, algorithmic system Machine learning model, foundation model, algorithmic model, predictive model AI application, machine learning application, AI solution, machine learning solution, AI services
Intervention	Definition *Synonyms:* Ontology, conceptual framework, taxonomy, terminology, scope
Comparison	Not applicable
Outcome	Find different definitions, clarity on differences
Context	Computer science, engineering, social sciences, decision sciences, business, management and accounting, arts and humanities, multidisciplinary, psychology

Databases

In order to sufficiently cover the area under investigation, different databases were selected. The following databases were used to obtain an overview of the literature: Scopus, Web of Science, IEEE Xplore. Google Scholar was used as a secondary source.

Inclusion and Exclusion Criteria

The following inclusion and exclusion criteria were chosen for the literature review. In particular, literature from 2012 onwards were included, as 2012 was the beginning of the modern AI era, with the influential AlexNet architecture (Krizhevsky et al., 2012) being released, marking the beginning of machine learning and deep neural networks as the dominant paradigm in AI. Moreover, only literature written in English was chosen, and their relevance to the research question was ensured by only including articles that provide a definition of the terms *AI model*, *AI system* or a concept synonymous to them in the final review. Lastly, only articles, reviews and conference papers were included. An overview of the criteria is provided in Table 3.

Table 3. SLR Inclusion and Exclusion Criteria

Category	**Selection Criteria**
Period	From 2012 to present
Language	English
Relevance to research question	The article provides a definition of *AI model*, *AI system*, or a concept synonymous with these terms
Type of source	Articles, reviews and conference papers

Quality Assessment and Data Extraction

The quality of the articles is mainly assessed by manually evaluating for how clear and complete their definitions are. The relevance of each article's definition to the research question is also evaluated.

Extracted data include the definitions of the relevant terms, as well as their sources. In some cases, the definitions in a source were attributed to another cited source. This was also noted during the manual review of the SLR results.

Search String

Figure 1 shows the final search string. Synonyms of the terms *AI system*, *AI model* and "definition" were chosen as defined by the PICOC criteria. In order to ensure articles were discussing definitions, the proximity operator W/4, or NEAR/4, was used to only include articles where the definitional term is within a distance of four words of the AI term. To further narrow the search, articles with terms such as "defined via" and "defined through" in the abstract were excluded, as this would typically result in definitions of other terms in the context of AI models or systems, instead of the definition of AI models or systems themselves. Furthermore, only articles in the subject areas specified in the context criterion of the PICOC method were included. Lastly, as a first literature search showed high occurrence of article keywords that were not relevant for this search, the following keywords were excluded: diagnosis, female, adult, male, students, diseases, explainable ai, major clinical study, covid-19, fairness, energy.

```
(``AI system'' OR ``artificial intelligence system'' OR ``machine learning system''
OR ``ML system'' OR ``Autonomous system'' OR ``Automated system''
OR ``algorithmic system''

OR ``AI model'' OR ``artificial intelligence model'' OR ``machine learning model''
OR ``ML model'' OR ``foundation model'' OR ``algorithmic model''
OR ``predictive model''

OR ``AI application'' OR ``artificial intelligence application''
OR ``machine learning application'' OR ``ML application''

OR ``AI solution'' OR ``artificial intelligence solution''
OR ``machine learning solution'' OR ``ML solution''

OR ``AI services'' OR ``artificial intelligence services'')

W/4 (``ontology'' OR ``defin*'' OR ``conceptual framework'' OR ``taxonomy''
OR ``terminolog*'' OR ``conceptualis*'' OR ``conceptualiz*'' OR ``scope'')

AND NOT (``defined via'' OR ``defined through'' OR ``well defined''
OR ``user defined'')
```

Figure 1. SLR Search String

This search was performed on October 15, 2025, and resulted in **864 papers** on Scopus, **113** on Web of Science, **16** on IEEE Xplore. After removing duplicates, a total of **896** papers were screened. Details can be found in Figure 2.

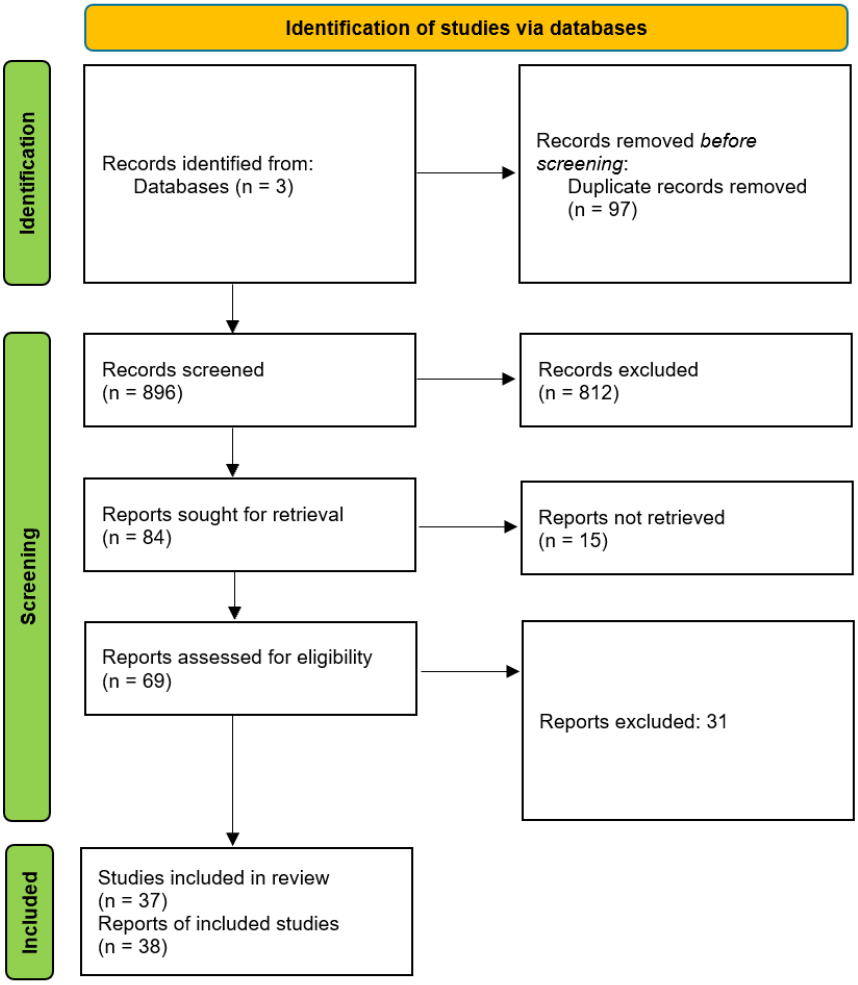


Figure 2. PRISMA Flow Diagram of the SLR

B. Findings and Analysis

This section presents the results of both manual and systematic literature reviews examining how *AI models* and *AI systems* are defined across academic and regulatory sources. These findings provide the empirical foundation for understanding the definitional landscape and the relationships between different formulations of *AI model* and *AI system* definitions. Then, we discuss paradigm shifts that lead to an evolution of the meanings of these terms. Finally, we explore definitional lineages in major academic references and regulatory frameworks, and investigate how changes in definitions in the OECD frameworks have affected later definitions.

B.1. Conceptual Perspectives on Existing Definitions

As described in Section A.1.1 (Manual Review of Global Regulatory and Standards Documents), we manually gathered over 80 definitions from four categories of institutions: (i) intergovernmental organisations, (ii) standards bodies, (iii) national or state-level governmental organisations, and (iv) NGOs. Across documents produced by governments, regulatory agencies, inter-governmental and non-governmental organisations, we observed that AI definitions often remain ambiguous and inconsistent, particularly at the boundary between an AI model and an AI system.

Additionally, based on the methods described in Section A.1.2 (Systematic Literature Review), we conducted an SLR which screened 896 records from the literature resulting in 37 studies included. The process and findings of this SLR is further described in Appendix A.2 (Overview of the Systematic Literature Review).

Collectively, the definitions gathered through both the manual and systematic literature review can be interpreted from several perspectives. While many definitions appear similar, they can nonetheless be considered from different angles that highlight different conceptual emphases. Overall, when including both the papers found through the SLR and the manual review, more definitions were found for *AI system* and related terms (83 instances), than for *AI model* and related terms (43 instances), without excluding double-counting for instances that appear in both the manual review and the SLR. Included in the following conceptual perspectives were the terms *AI model*, *machine learning model*, *AI system*, and *machine learning system*, as well as sources providing clarifications on autonomous systems and intelligent systems.

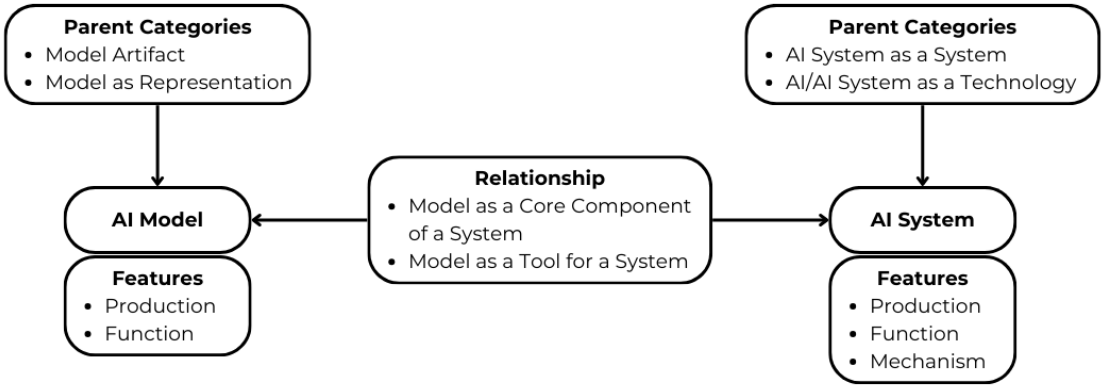


Figure 3. Overview of conceptual perspectives for definitions of *AI model* and *AI system*.

We organise these conceptual perspectives along three dimensions (Figure 3). The first captures what kind of "thing" an AI model or system is understood to be; for example, a representation, a component, or a system. This is about identifying the subordinate concept of a term, which ISO 704:2022 (ISO, 2022) notes as an important component of a definition. We refer to this dimension as "parent categories", as it reflects the broader conceptual class to which the definition assigns the term.

The second dimension addresses the features of the model or system referenced in the definitions. For both models and systems, this addresses their production and function, analysing how models and systems are made, and what they do respectively. Additionally, for AI systems, this dimension also addresses its mechanism in order to define how the system operates.

Third, we also consider the relational dimension, looking at the relationship between AI models and AI systems and how they are distinguished. Having a closer look at this difference provides further clarity on both terms. We refer to this dimension as "relationship between AI model and AI system".

As an example of these dimensions, one source might define an AI model as the result of machine learning algorithms (parent category), used to make predictions (features), and referred to as the reasoning component of an AI system (relationship). Any of these perspectives can overlap within a single definition. For example, several definitions define an AI model as both the core reasoning or knowledge component of an AI system (relationship), and also as a representation of some external structure (parent category).

The following sections elaborates on these different perspectives with several examples.

B.1.1. AI Model

We first consider how the term *AI model* is conceptualised across definitions. The definitions can be roughly divided into two parent categories: model artifact and model as representation. Definitions under model artifact describe models as the

result of building or training, specifying the required elements a model possesses after building (e.g. weights, parameters, or the specific architecture). Definitions describing models as representations emphasise not the way a model has been built, but its purpose of providing a representation of some external structure (e.g. representing laws of the external world or the data it has been trained on).

Parent Categories:

1. **Model artifact (result of building/training)**: Focusing on weights, parameters, or architecture.
   (a) "A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms." (OECD, 2019b, p. 7)
2. **Model as representation**: A model represents some external structure (e.g. the laws of the external world or the distribution of token sequences in a dataset).
   (a) "AI models are mathematical representations that capture, in a set of parameters, the patterns underlying their training personal data" (European Data Protection Supervisor, 2025, p. 9)
   (b) "physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data" (ISO/IEC, 2022b, ISO/IEC 22989:2022 3.1.23)
   (c) "A Model is an actionable representation of all or part of the external environment of an AI system that describes the environment's structure and/or dynamics" (OECD, 2019b, p. 7)

Next, the features of an AI model are split into two perspectives: production and function. In accordance with the model life cycle, we first consider how models are built, and next what they are used for after being built. Accordingly, the production perspective contains all definitions referring to how a model is built or trained, including specific techniques or approaches. For example, a model can be described as being built using machine learning. The function perspective addresses definitions describing what a model is used for after building. This can be its intended use, specifications on its purpose, and describing the capabilities and characteristics it possesses to fulfil its function.

Features:

1. **Production (how it is built/trained)**: Considering the techniques or approach used to develop or create the model (e.g. use machine learning with data or hard-coded expert knowledge).
   (a) "uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs" (Executive Office of the President, 2023, Sec. 3(c))
2. **Function (what it is used for after building)**:
   (a) Emphasising what the model is intended to do (e.g. inference, prediction, pattern recognition).
      i. "used to make inferences from inputs to produce outputs" (OECD, 2024, p. 8)
      ii. "can make predictions/decisions over data" (Estévez Almenzar et al., 2022, p. 43)
   (b) Wide purpose or narrow purpose: Describing its characteristics and capabilities (adaptable across a wide range of tasks or specific to one).
      i. "every AI Model possesses the potential to adapt to other tasks, some (for example, neural networks) even being universal approximator" (Li & Zhan, 2022, p. 1)

These perspectives are non-exclusive and orthogonal, meaning that definitions are not separated into different categories, but can fall into several different perspectives. Furthermore, perspectives on the features of a model only suggest different focal points of a definition. Therefore, different definitions grouped into these features may have other differences, as long as they share the same focal point of either production or function. For example, one definition might say a model is built using machine learning, while another might say it is built using hard-coded expert knowledge; but they both are defined with similar functions. It is also important to note that while two definitions may occupy different points along the first dimension (different parent categories of *AI model*), they can agree on the same definition along the second dimension (the same features are referenced in the definition). For example, one definition might describe a model as the result of machine learning, while the other describes it as a representation of the external world, resulting in different parental categories. However, they might both describe the function of an AI model as making inferences.

B.1.2. AI SYSTEM

In contrast to AI models, AI systems are defined more frequently and comprehensively. As with models, definitions can be viewed from two different dimensions: the parent categories and the features. Definitions of *AI system* can be divided into two different parent categories: AI systems as a type of system or as a type of technology. Many definitions describe AI systems as a type of system, specifying for example an AI system as a kind of machine-based system or an engineered system. Definitions falling into the technology category are focused on a broader category, either describing AI systems as technologies, or defining AI as a system, suggesting that the terms *AI* and *AI system* are often used interchangeably.

Parent Categories:

1. **AI system as a specific system**: AI systems as a specific type of system.
   (a) "An AI system is a machine-based system" (OECD, 2024, p. 4)
   (b) "an engineered system" (ISO/IEC, 2022d, ISO/IEC 22989:2022 3.1.4)
   (c) "Artificial intelligence (AI) systems are software (and possibly also hardware) systems" (High-Level Expert Group on Artificial Intelligence, 2019b, p. 36)
2. **AI/AI system as a technology**: AI systems as a technology and AI as AI-based systems in general.
   (a) "AI systems are information-processing technologies" (UNESCO, 2022, p. 10)
   (b) "Artificial Intelligence (AI) refers to systems" (High-Level Expert Group on Artificial Intelligence, 2019a, p. 7)

The features of an AI system are grouped into similar perspectives as for an AI model. The first angle, considering the production, addresses what a system is made of and which techniques were used to build it. For example, a system can be described as requiring different components, such as an AI model, data pipelines, and an architecture. The second perspective describes the function of an AI system, entailing definitions that describe what an AI system does (e.g. giving predictions) and what kind of action it is conducting (e.g. acting autonomously or affecting the physical world). In addition to the perspectives of production and function, definitions of AI system have also been found to describe its mechanism. This perspective answers how the system does what it is created to do. For example, if a system is built to provide recommendations (function), this perspective may describe it as generating recommendations by inferring an output from the input it receives, or by perceiving, planning and reflecting.

Features:

1. **Production (how it is built)**: What a system is made of and the techniques used to build it.
   (a) "can either use symbolic rules or learn a numeric model" (High-Level Expert Group on Artificial Intelligence, 2019b, p. 36)
   (b) "more 'grown' or 'trained' than they are programmed [...] produced by rewarding an AI for 'good behavior' and punishing it for 'bad behavior'" (Center for Humane Technology, 2024)
   (c) "Bringing one or more AI models to a real-world application is not immediate, as it implies the effort of integrating them in a functional system (i.e., an AI system), including the necessary infrastructure, user interfaces, data pipelines, and other components required for the application to operate effectively in a production environment" (Hupont et al., 2024, p. 3)
2. **Function (what it does)**: What is the system intended to do and what kind of actions it performs (intelligent action, impactful action).
   (a) "generate outputs such as predictions, content, recommendations, or decisions" (European Parliament and Council of the European Union, 2024, Article 3(1))
   (b) "vary in their levels of autonomy" (OECD, 2024, p. 4)
   (c) "influence physical or virtual environments" (European Parliament and Council of the European Union, 2024, Article 3(1))
3. **Mechanism (how it does it)**: How does the system fulfil its function.

(a) "It does so by utilising machine and/or human-based inputs/data" (OECD, 2019b, p. 7)

(b) "[AI systems act] by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal." (High-Level Expert Group on Artificial Intelligence, 2019b, p. 36)

B.1.3. Relationship between AI Models and AI Systems

Another dimension to be considered is the relationship between AI models and AI systems. This relationship has been split into two perspectives. "Model as a core component of a system" encompasses definitions that describe models as the core of an AI system, often performing a system's main function. "Model as a tool for a system" takes a broader view on the relationship of an AI model and an AI system, where an AI model is defined according to its function as a tool for an AI system.

Relationship between AI models and AI systems:

1. **Model as a core component of a system**: A model defined as the core of an AI system, performing functions such as reasoning and knowledge recall (e.g. frontier language models in modern chatbots).
   (a) "A model characterizes an input-output transformation intended to perform a core computational task of the AI system" (Dunietz et al., 2024, p. 12)
2. **Model as a tool for a system**: Focusing on the function of a model as a tool within an AI system.
   (a) "Al systems contain a model which they use to produce predictions [...]" (ISO/IEC, 2022a, ISO/IEC 22989:2022 7.1)
   (b) "AI systems are information-processing technologies that integrate models and algorithms that produce a capacity to learn and to perform cognitive tasks" (UNESCO, 2022, p. 10)

B.2. Paradigm Shifts in the Meaning of AI Models and AI Systems

Having laid out the different conceptual perspectives of *AI models* and *AI systems*, we now describe the observed trends and evolutions of definitions of these terms over time. This section is not intended as a comprehensive history. Its purpose is to show that shifts in the meaning of *AI model* and *AI system* track changes in technical paradigms rather than deliberate definitional refinement, which become inadequate when applied to modern general-purpose models.

B.2.1. AI Models

Conceptual Senses of AI models The Oxford English Dictionary (n.d.) entry distinguishes four senses of *model* as a noun: a representation of structure, an object of imitation, a type or design, and other uses (obsolete historical senses). This corresponds to the sense of "model as representation" that we identified in our literature review. It is also compatible with "model artifact (result of building/training)" as its parent category and "model as a tool for a system" as its relationship with the system. However, this definition of *model* is not clearly compatible with "model as a core component of a system", since reasoning and knowledge recall are not capabilities typically associated with representations.

The notions of "model as representation" and "model as a particular tool" are closely tied to the concept of a "world model" which is a key component of some (typically older) archetypes of AI systems or agents. The idea is that since most AI agents will only have partial observability of the world around them, they need some way to deduce or predict changes in the world that they cannot directly observe, such as food decaying in a closed fridge, or the outcomes of actions that the agent has yet to perform. Thus, an AI agent will require some component that "models" the exterior world and allows them to make better-informed decisions. For example, *Artificial Intelligence: A Modern Approach* (Russell & Norvig, 2009) says that "knowledge about 'how the world works' [...] is called a model of the world" and "an agent that uses such a model is called a model-based agent." Crucially, under this paradigm the "model" is not expected to decide what action the agent should take, merely to predict what the likely effects of different actions would be.

In the field of modern AI, LLMs are typically referred to as *models* despite not seeming to be a model of anything, and are relied upon to perform tasks and make decisions by themselves, rather than being used to inform a decision-making

process. Therefore the use of the term *model* to describe these entities appears to be due to the ancestry of current frontier models, since nearly all such models are (primarily) neural networks using transformer architecture (Zhao et al., 2025). Neural networks were originally proposed as a model of networks of neurons in the human brain (McCulloch & Pitts, 1943), and are known to be capable of approximating (or "modelling") a wide range of mathematical functions (Cybenko, 1989; Hornik et al., 1989). Furthermore, transformers were originally proposed in the field of computational language **modelling** (Vaswani et al., 2017). For these reasons, it is commonplace to refer to any neural network with transformer architecture as a *model*, even if it is not clear what, if anything, it is modelling (as is the case in the neural networks that form the core of modern frontier LLMs).

As transformer-based neural networks have become an increasingly dominant force in AI, particularly in the most capable and general systems, and as more classical agent architectures have become correspondingly less popular, we argue that the usage of the term *AI system* has shifted. This can be observed in the OECD definitions. The 2022 definition states "AI models are actionable representations of all or part of the external context or environment of an AI system (encompassing, for example, processes, objects, ideas, people and/or interactions taking place in context)" (OECD, 2022), which corresponds to the first (representation) sense of *AI model*. The 2024 definition states "An AI model is a core component of an AI system used to make inferences from inputs to produce outputs [. . .] while the parameters of an AI model change during the build phase, they usually remain fixed after deployment once the build phase has concluded" (OECD, 2024), corresponding to a different second sense of a model, which views the model primarily as a component instead of a representation.

While frontier models are not typically used for modelling purposes, they are still built from an architecture (transformers) that was intended to model something (natural language). What, if anything therefore, can frontier models be said to be models of? The pre-training of GPT-3 attempts to maximise its accuracy at modelling its training dataset (Brown et al., 2020), meaning that pre-trained GPT-3 can meaningfully be said to be a model of whatever corpus of text it was trained on. The fine-tuning will, strictly speaking, make it less representative of its training data (for example, it will produce less offensive language than contained in its dataset). However, it will move the model closer to being a useful AI assistant. Therefore, if a fine-tuned GPT-3 can be meaningfully said to be a model of anything, it is an (imprecise) model of the perfect AI assistant.

Historical Shift towards Generality Beyond a shift in the different senses of an AI model, there was also a notable shift from AI models that have specific purposes to being general-purpose during the machine-learning era.

2016-2020: Statistical Methods, Specific Tasks

When specifically searching for definitions of *AI model* in the systematic review, some early definitions characterised AI models predominantly through the lens of machine learning and statistical models, emphasizing their task-specific nature. One 2016 definition described machine learning models as "non-deterministic statistical approximations" bound to function correctly only "a certain portion of the time," emphasizing their probabilistic nature and limited scope (Ramanathan et al., 2016). Following the conventions of this era, ISO/IEC (2022a, ISO/IEC 22989:2022) used the term "AI model" with reference to two separate definitions for "AI" and "model", where model is defined in part as a "mathematical or otherwise logical representation of a system". This pattern in the retrieved literature suggests that the *AI model* terminology was closely associated with traditional machine learning paradigms during this period.

2021-2022: Foundation Models

By 2021, definitions had started to include structured components such as algorithms, training datasets, and performance metrics while remaining fundamentally task-oriented. The concept of a foundation model was introduced in 2021, marking a shift in moving away from machine learning for specific applications to focus on versatility. These definitions focus on models being "pretrained on massive amounts of data" and capable of adaptation "to perform a wide variety of tasks through fine-tuning" (Bommasani et al., 2022). This period marked the transition from AI as specialised tools to AI as general-purpose capabilities. The characteristics of foundation models would later be recognised under the term "general-purpose" after its use in the EU AI Act in 2023.

2023-2024: Formalisation of General-purpose

By 2023, the shift in capability influenced the definition of *general-purpose AI models* within formal regulatory recognition. Article 3(63) of the EU AI Act defined these as models "capable of competently performing a wide range of distinct tasks

regardless of the way the model is placed on the market" (European Parliament and Council of the European Union, 2024). This trend reflects a fundamental reconceptualisation of AI from being application-specific to general-purpose, which played a significant role in expanding the scope of how systems were to be defined.

B.2.2. AI SYSTEMS

From the systematic literature review and manual extraction, several observations emerge regarding the evolution of the definition of AI systems. Unlike the evolution observed in the definitions of AI models, the trends in the definitions of AI systems emerged in parallel, hence a lack of distinct timelines.

The Emergence of Autonomy and Adaptiveness as Core Attributes Initial definitions of autonomy in AI systems and/or autonomous systems focused on independence from human supervision. The 2016 characterisation emphasised systems "capable of performing certain tasks in an unstructured environment without human supervision," with self-awareness defined as the ability to "analyze the situation and do the decision-making on its own" (Helle et al., 2016). The OECD's 2019 definition introduced the idea that AI systems are "designed to operate with varying levels of autonomy," acknowledging autonomy as a spectrum rather than a binary state (OECD, 2019b).

The period from 2019 to 2024 saw increasing sophistication in how autonomy was conceptualised within the definitions of systems. The 2024 EU AI Act definition represented this evolution, noting that the autonomy offers the possibility of continual adaptation, acknowledging systems "designed to operate with varying levels of autonomy [. . .] that may exhibit adaptiveness after deployment" (European Parliament and Council of the European Union, 2024, Article 3(1)).

The concept of adaptiveness introduced an additional temporal dimension: systems could change their behavior after deployment through "self-learning capabilities, allowing the system to change while in use" (European Parliament and Council of the European Union, 2024, Recital 12). This post-deployment evolution distinguished modern AI systems from earlier conceptions of AI as a program for symbolic logic, and traditional software which operates according to fixed programming.

Recognition of Unintended Capabilities Early definitions implicitly assumed that AI systems would perform tasks for which they were explicitly designed and trained. Capabilities were understood as direct results of intentional design choices, training data selection, and optimisation objectives.

Later definitions explicitly recognised that modern AI systems exhibit capabilities beyond those explicitly programmed or trained for. The 2022 definitions of general-purpose AI systems could have "multiple intended and unintended purposes" (Future of Life Institute, 2022). By 2023, definitions acknowledged systems that could "accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained" (Gutierrez et al., 2022).

Overall, the recognition of emergent capabilities signals an acknowledgement that traditional approaches based on comprehensive specification and validation become vulnerable to rapid technological change.

Legal Frameworks and Regulatory Motivations Pre-2021 definitions originated primarily from academic and technical communities, focusing on capabilities, methods, and performance characteristics. These definitions were adequate for descriptive purposes in technical research, but were not written with regulatory needs in mind.

The *OECD AI Principles* (OECD, 2019b), adopted in May 2019, played a foundational role in shaping subsequent regulatory standards. Along with their supplementary documentation that define their scope, they established a common reference point for the EU, ISO, and NIST.

The EU AI Act (European Parliament and Council of the European Union, 2024), developed between 2021 and 2024, uses a risk-based classification framework (unacceptable, high, limited, and minimal risk) rather than attempting comprehensive definitional precision.

The 2023 American Executive Order on "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (later rescinded in 2024) provided an alternative regulatory approach, defining AI systems broadly as "any data system, software, hardware, application, tools, or utility that operates in whole or in part using AI," demonstrating generalised and all-encompassing approach (Executive Office of the President, 2023).

Overall, the emergence of legally operative definitions represents the maturation of the field and its growing significance.

These definitions establish frameworks for governance and compliance. However, as the analysis above demonstrates, approaches to definitional precision vary considerably, reflecting competing priorities between regulatory clarity and future-proofing.

B.2.3. Overall Trends

In general, the trends observed in the definition of AI systems may also apply to definitions of AI models. However, trends identified for AI models were derived primarily from changes in explicit definitional language across the sources retrieved from our search. In contrast, several trends associated with AI systems were observed in their definition and accompanying explanatory context and regulatory framing. Definitions of AI models were typically accompanied by far less contextual elaboration, reflecting the larger volume of system-level definitions retrieved in the review. Moreover, governance discourse mostly focused on the concept of systems over models. As such, these system-level trends should not be interpreted as uniformly applying to AI models, though both systems and models may share broader conceptual paradigms.

B.3. Definitional Lineages and Influence

In this section we consider more broadly how the various definitions from our surveys have influenced each other. Our analysis reveals that many secondary definitions (summarisations, syntheses, and adaptations), and most definitions proposed by major organisations, can be traced back to a particularly concentrated set of foundational sources, including OECD frameworks. We also find that these OECD frameworks have evolved in ways that integrate previously distinct concepts, blurring the boundary between models as representations and models as components of AI systems. We examine these issues in depth, tracing the chain of influence through various sources, analysing how ambiguities have compounded through successive iterations.

B.3.1. Overall Chain of Influence

As described, much of the definitions used in major regulatory frameworks have their lineages traced back to the OECD definitions, which in turn are heavily based on the work of Stuart Russell and Peter Norvig, as illustrated by (Figure 4).

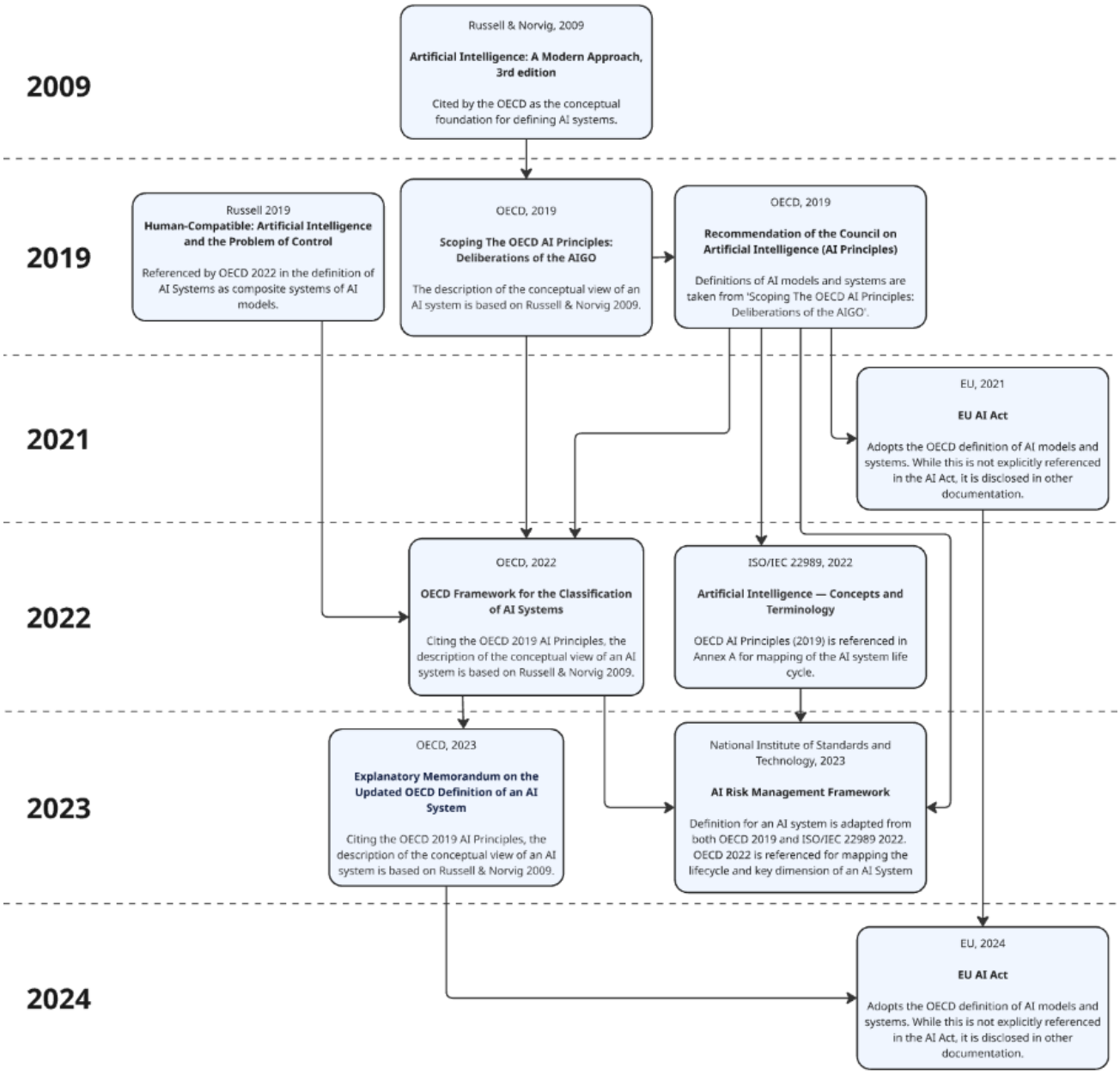


Figure 4. Definitional lineage of terms related to *AI models* and *AI systems*

The lineage begins with Russell and Norvig's *Artificial Intelligence: A Modern Approach* (Russell & Norvig, 2009, 3rd edition), which articulates the agent-based architecture of perception, representation, reasoning, and action. This framework is adopted directly into OECD's *Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)* (OECD, 2019b), which identifies Russell and Norvig as the conceptual basis for understanding AI systems as composite systems of interacting models. AIGO deliberations informed the *OECD Recommendation on Artificial Intelligence* (OECD, 2019a, AI Principles). OECD (2019b) established key definitional elements including "machine-based system," "varying levels of autonomy," and the capacity to "influence the environment," which would propagate through subsequent regulatory instruments. Notably, while the 2019 OECD documents (OECD, 2019b;a) incorporated the concept of *model* as a component of AI systems, they did not provide a formal standalone definition of *AI model* at this stage.

In 2022, the OECD published the *Framework for the Classification of AI Systems* (OECD, 2022), which built upon OECD

(2019b) by embedding the same conceptual architecture into a more granular classification taxonomy. This document formally introduced and defined the term *AI model* as a distinct concept, establishing a clearer conceptual boundary between models and systems.

Parallel to this, ISO/IEC 22989:2022 *Artificial Intelligence—Concepts and Terminology* (ISO/IEC, 2022a) adopted and detailed AI terminology for international technical standardisation. While ISO drew on its own pre-existing general definition of *model* from other standards, the document explicitly references OECD (2019b) in Annex A when mapping the AI system lifecycle.

By 2023, the U.S. National Institute of Standards and Technology (NIST) published its *AI Risk Management Framework (AI RMF 1.0)* (National Institute of Standards and Technology, 2023). NIST explicitly notes that its definition of an AI system is "adapted from: OECD Recommendation on AI: 2019; ISO/IEC 22989:2022," and it uses the OECD (2022) to structure its lifecycle and dimensional analysis.

Also in 2024, the OECD published the *Explanatory Memorandum on the Updated OECD Definition of an AI System*(OECD, 2024), revising the 2019 definition. The memorandum states that the updated definition was shaped to "support broad alignment of the definitions of AI systems in several ongoing processes in the European Union" as well as other emerging regulatory frameworks.

Subsequently, the European Commission (2021) adapted and expanded upon the OECD (2019b). Although the OECD is not explicitly cited in the legislative text itself, the influence is acknowledged in recitals and supplementary documentation. The final EU AI Act (European Parliament and Council of the European Union, 2024, Article 3(1)) demonstrates convergence with the updated 2024 OECD definition, incorporating OECD's new additions: "machine-based system," "explicit or implicit objectives," "varying levels of autonomy," "infers from input," and "influence physical or virtual environments."

This represents a bidirectional feedback loop: OECD (2024) notes that its revised definition was developed with awareness of the evolving EU AI Act, while the EU AI Act's final text converged toward the OECD formulation.

B.3.2. Analysis of OECD Frameworks

The OECD's AI taxonomy has become one of the most influential frameworks for AI governance globally. Its definitions have been adopted or adapted by numerous jurisdictions, In particular, definitions originating in the OECD framework have influenced ISO standards, the NIST AI Risk Management Framework, and most significantly the EU AI Act.

This section assesses whether successive OECD revisions preserve a stable conceptual boundary between *AI model* and *AI system*. We observe that across the 2019, 2022, and 2024 iterations, the framework progressively expands the meaning of *AI model* without sufficient conceptual clarification. As a result, the distinction between model and system becomes increasingly flexible, which may create challenges when used for regulatory allocation of obligations.

OECD (2019b): The Original Conceptual Framework OECD (2019b) noted its adaptation of Russell & Norvig (2009) for its definitions, a connection made clear by comparing the two conceptual diagrams, as shown in Figure 5.

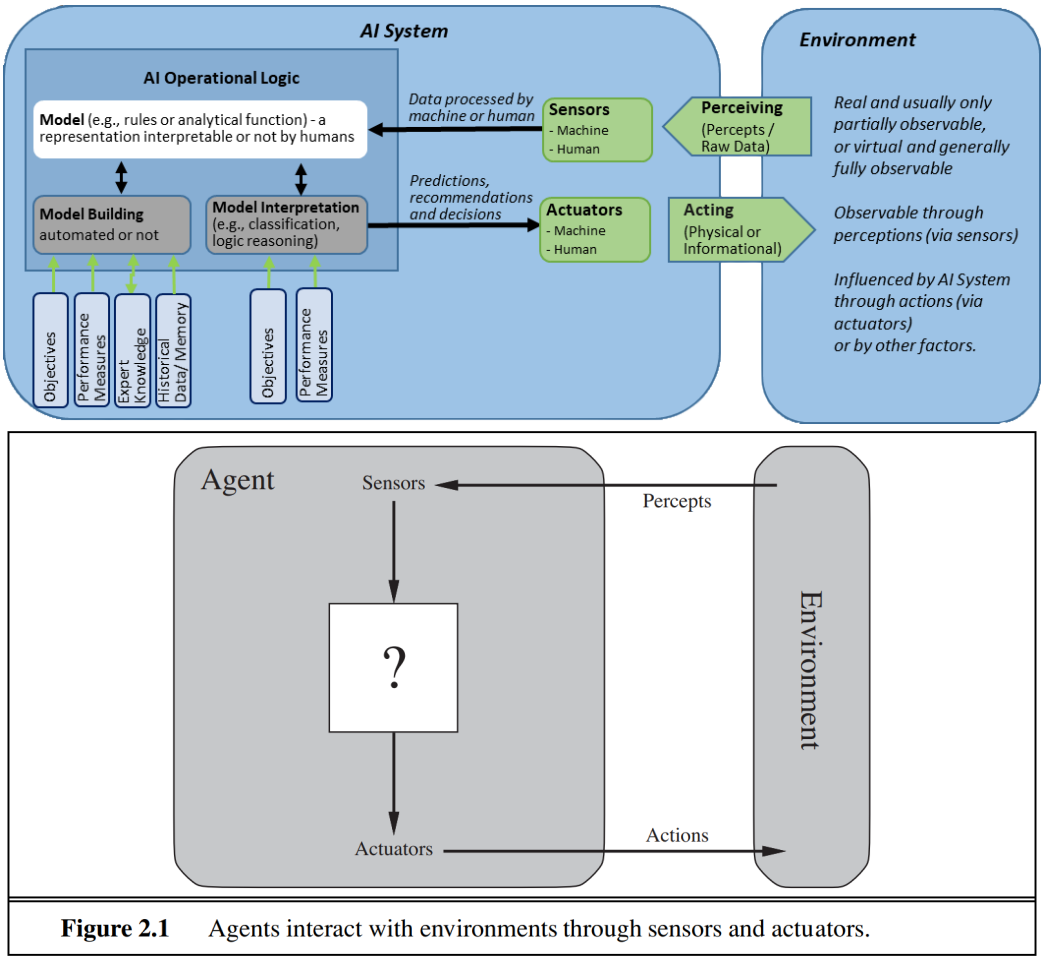


Figure 5. OECD (2019b) (top) versus Russell & Norvig (2009) (bottom)

The OECD's 2019 report *Recommendation of the Council on Artificial Intelligence* (OECD, 2019a) established a comprehensive taxonomy with three core concepts. These definitions appear to draw heavily from Russell and Norvig's agent-based framework, while introducing several additional terms and conceptual reinterpretations (Russell & Norvig, 2009). The main relevant definitions are AI system (or intelligent agent), AI operational logic, and model, as summarised in Table 4.

Table 4. Key definitions related to AI models and AI systems from OECD (2019b)

AI system / Intelligent Agent	"An AI system is a machine-based system that is capable of influencing the Environment by making recommendations, predictions or decisions for a given set of Objectives. It does so by utilising machine and/or human-based inputs/data to: i) perceive real and/or virtual environments; ii) abstract such perceptions into models manually or automatically; and iii) use Model Interpretations to formulate options for outcomes." (OECD, 2019b, p. 7)
AI Operational Logic (p. 6)	"The key power of an AI system resides in its Operational Logic, which, for a given set of objectives and based on input data from Sensors, provides output for the Actuators — as recommendations, predictions or decisions — that are capable of influencing the state of the Environment." (OECD, 2019b, p. 6)

Model (p. 6-7)	"A Model is an actionable representation of all or part of the external environment of an AI system that describes the environment's structure and/or dynamics. The model represents the core of an AI system. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms. Model Interpretation is the process of deriving an outcome from a model." (OECD, 2019b, p. 7) "Model (a data object constructed by the Model Building process)" (OECD, 2019b, p. 6)

OECD (2019b) defines an AI system as comprising two parts: its function and its mechanisms (including both building and using). The latter two terms are defined by their relationship to the AI system itself: the AI operational logic maps the "input from sensor" to the "output for actuator," while the model represents the "external environment" of the AI system.

This taxonomy presents three conceptual challenges that may affect its application in regulatory contexts:

1. Ambiguous Inclusion or Exclusion of Human Involvement

The report states: "An AI system consists of three main elements: Sensors, Operational Logic and Actuators" (OECD, 2019b, p. 8). Under this strict interpretation, a software program that generates a credit score would not, by itself, constitute an AI system, as it lacks a physical actuator (e.g., the bank staff who ultimately approve or deny the loan). However, the OECD (2019b)'s conceptual diagram introduces flexibility by noting that sensors and actuators can be human. Consequently, a credit scoring system would be considered a hybrid of human and machine, which is an interpretation that may differ from common public understanding.

This ambiguity marks the beginning of a broader, recurring issue in subsequent years, where human involvement is often reconciled by classifying such systems as AI with a "low level of autonomy."

2. The Conflation of Real, Digital, and Abstract Objects

This framework illustrates a conflation of object types. It describes an AI system as being made of Sensors, Actuators, and Operational Logic. While sensors and actuators are tangible, physical components, the operational logic is presented as a pure, abstract object described as a "logical mapping" with no reference to the physical computational substrate (e.g., hardware like GPUs) required to implement it.

Take an autonomous vehicle as an example. Under this description, the self-driving car is said to be "made of" the radar (sensor), the engine (actuator), and the logical mapping between them. The computing hardware that physically instantiates this logic is not explicitly referenced in the definition, which leaves the relationship between abstract function and real-world implementation less clearly specified.

Russell & Norvig (2009, p. 35) maintain strict separation: "The agent function is an abstract mathematical description; the agent program is a concrete implementation, running within some physical system."

They formalise this as:

$$Agent = Architecture + Program$$
$$Architecture = Sensor + Actuator + Computing\ Device$$

Therefore:

$$Agent = (Sensor + Actuator + Computing\ Device) + Program$$

The OECD framework does not explicitly articulate this distinction, instead presenting the system as encompassing both tangible components and abstract functional elements.

3. The Ambiguity of "Actionable Representation"

The term "actionable representation" is not formally defined within OECD (2019b)'s taxonomy. Nevertheless, there are academic sources which define it. For example, Ghosh et al. (2018, p. 1) defines it as the following:

1. Is "aware of the dynamics of the environment."
2. Captures "only the elements of the observation that are necessary for decision making rather than all factors of variation, eliminating the need for explicit reconstruction."
3. Is typically "obtained from a goal-conditioned policy—a policy that knows how to reach arbitrary goal states from a given state."

This narrower technical interpretation requires a "goal-conditioned policy" which may not align with all AI systems encompassed by the OECD framework, potentially creating ambiguity between terminology and intended scope. The OECD appears to use "actionable" colloquially to mean "can inform actions," but this divergence from technical usage creates ambiguity about what exactly a "model" is supposed to be.

Terminological Relationships

OECD (2019b)'s taxonomy introduces the non-standard concept of AI operational logic, which is rarely found in other literature. Its definition, describing the mapping of outputs for actuators from inputs received from sensors, is conceptually almost identical to Russell and Norvig's agent function, which is defined as a description that "maps any given percept sequence to an action" (Russell & Norvig, 2009, p.35).

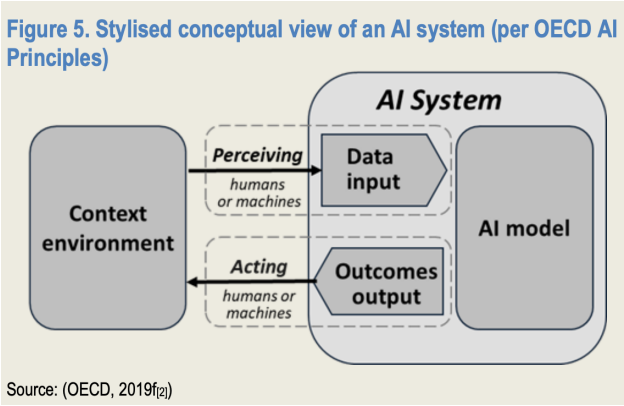
The relationship and differences between the core concepts in the two frameworks are summarised in Table 5:

Table 5. Relationship and differences between Russell & Norvig (2009) term and OECD (2019b)

Russell & Norvig's Term	**OECD (2019b) Equivalent**	**Key Difference in Taxonomy**	**Primary Object Type**
Agent / Rational Agent / Intelligent Agent	**AI System / Intelligent Agent**	The definitions are almost identical.	OECD: A mixture of real, digital, and abstract objects.
Agent Function	**AI Operational Logic**	Near-identical definitions (mapping inputs to outputs)	Abstract
Agent Program	**(Most analogous to) Model**	Agent program is concrete software; OECD model is "data object"	Digital
World Model (of a model-based agent)	A specific type of Model	Russell & Norvig (2009): specific to subset of agents (model-based); OECD: implied for all systems	Digital or Abstract

OECD (2022): Compounding the Conceptual Confusion As an update to the previous version, OECD (2022) did not fully resolve the conceptual questions identified in 2019; instead, it consolidated previously distinct elements into a unified framework. The framework merged the already problematic AI operational logic and model into a single, ill-defined term: "AI model". This move effectively formalises the earlier conflation, redefining the model as the component that dictates the system's operational logic while simultaneously treating it as a representation of the environment. The new conceptual diagrams are shown in Figure 6.

The original rationale provided was a desire for simplification (OECD, 2022, p. 23):



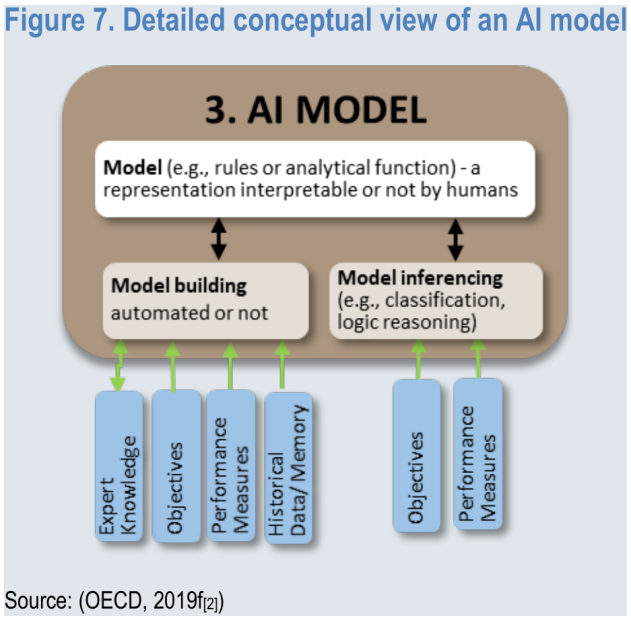


Figure 6. (OECD, 2022)'s conceptualisation of the term *AI model*

> "It should be noted that in the present classification framework, the 'Perceiving' (data collection) and 'data/input' elements – that were separate elements in the original OECD work presented in Figure 5 – have been combined in an effort to simplify the framework (see dotted lines in Figure 4). The 'Outcomes' and 'Acting' elements have also been combined."

This simplification creates significant ambiguity. Combining "Perceiving" with "Data/Input" blurs the boundary between the AI model and the AI system. Inputs now flow directly from the environment into the AI model, bypassing the conceptual layer of the system and its sensors that would typically collect and pre-process data. The AI Model is not only expanded to encompass what was previously called the AI's operational logic, but the operational logic itself is broadened relative to the system, where its input becomes the system's input. This, in turn, narrows the definition of the AI System itself, potentially excluding essential components like data pre-processing modules.

For this taxonomy to remain coherent, there is a logical pressure to treat the **AI model as an abstraction of the entire AI system** — a tendency that becomes explicit in the 2024 documents, which we discuss further in Section B.3.2 (OECD (2024): Expansion and Functional Broadening). If both the AI model and the AI system are conceived as concrete entities, the conceptual boundary between them becomes less clearly defined. This interpretation may not fully capture the structure of certain real-world deployments. For example, a recommendation model abstracts user preferences for items, not the entire recommendation system, which requires additional components to collect live user data and integrate business rules (e.g., paid promotions) that ultimately determine how the model's outputs are applied.

Internal Inconsistencies: Diagram versus Definition

The conceptual diagram places the label "AI model" precisely where "AI Operational Logic" was previously located. Yet, the textual definitions of an "AI model" remain closer to the OECD (2019b) notion of a "model" (a representation of the environment) rather than to operational logic. A key inconsistency is that the diagram depicts system inputs and outputs flowing directly to and from the AI model, while the definition states a model may represent "all or part" of the external environment. This implies the model's inputs/outputs need not align with the system's.

Table 6. Differences and similarities between terminologies in OECD (2019b) and OECD (2022)

Aspect	OECD (2019b)	OECD (2022)	Consequences

Core Definition	An actionable representation of all or part of the **external environment** of an AI system.	An actionable representation of all or part of the **external context or environment** of an AI system.	Slight expansion from "environment" to "context."
Relationship with Environment	The model [...] **describes** the environment's structure and/or dynamics.	The model **represents, describes, and interacts with** real or virtual environments.	Introduces the incorrect idea that a model can "interact."
Relation to System	Clearly a component *within* the system, representing its environment.	Blurred; diagrammatically equated with the system's core logic, definitionally remains a representation.	Fundamental tension between the visual framework and its textual definitions.

The OECD (2022) claim that a model can "interact with" environments is conceptually flawed and contradicts common usage of the term "actionable representation," where "actionable" should mean the representation's outputs can inform actions taken by *another component* (e.g. an actuator), not that the model itself acts. This misinterpretation is evident in the report's own examples. It states that actions are taken "based on the outcomes of an AI system, e.g. by autonomous vehicles," confirming the system provides an output that is actualised by the vehicle. Similarly, an AI gamer's policy model generates action probabilities; the system architecture executes the actions. The capacity to act resides in the system, not in the model.

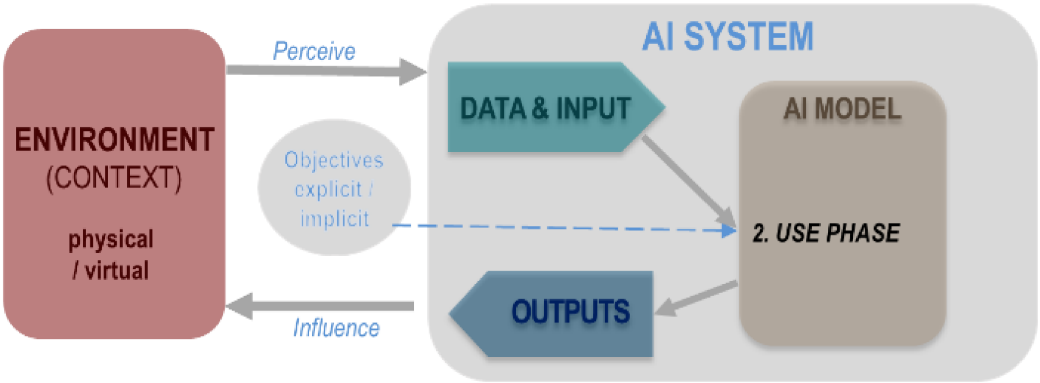


Figure 7. OECD (2024)'s conceptual diagram

OECD (2024): Expansion and Functional Broadening Subsequently, OECD (2024) illustrates the extent to which the taxonomy has evolved in scope and application. In its latest conceptual diagram as seen in Figure 7, the term "influence" replaces "acting," a change that further abstracts the system's relationship with its environment without resolving core ambiguities.

Critically, OECD (2024) appears to have recognised a foundational flaw, where many modern AI systems, such as those using model-free reinforcement learning, operate without an explicit "model" as previously defined. Instead of revising the earlier conceptual structure, the organisation expanded the definition of *model* to retroactively include these systems, thereby altering the term's meaning and its relationship to the AI system.

This has resulted in a set of scattered, often conflicting definitions for *AI model* within the same document:

1. **A system component**: "[. . .] an AI model is a core component of an AI system used to make inferences from inputs to produce outputs." (OECD, 2024, p. 8)

2. **A representation (per ISO)**: "A model is defined as a 'physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data' [. . .]" (OECD, 2024, p. 8)
3. **An input-output mapping**: "AI models include, among others, various kinds of input-output functions (such as decision trees and neural networks)." (OECD, 2024, p. 8)
4. **A world model**: "An AI model can represent the transition dynamics of the environment, allowing an AI system to select actions by examining their possible consequences using the model." (OECD, 2024, p. 8)
5. **A policy or Q-function**: "In such cases ['model-free reinforcement learning'], the policy or Q-function would be an AI model in the more general sense of an input-output function." (OECD, 2024, p. 11)

These definitions create a high degree of contextual and internal conflict. For instance, in a model-based agent, the AI system uses the world model to determine actions. In a model-free agent, the AI model (the policy/Q-function) is now said to determine the action itself. This contradiction highlights an underlying shift: the model is inconsistently framed as a component within the system, a functional abstraction of the system, and the system's driving logic.

This new, expansive definition finally aligns the term *AI model* with what was originally termed *AI operational logic* or Russell & Norvig (2009)'s *agent function*. The taxonomy now reflects a substantially broadened conception of *AI model* across contexts, as summarised in Table 7, which tracks the term's shifting applicability across frameworks.

Table 7. Conceptions of *AI model* across different frameworks

Concept	**Russell & Norvig (2009)**	**OECD (2019b)**	**OECD (2022) (Text)**	**OECD (2022) (Image)**	**OECD (2024)**
World model of a model-based agent	Yes	Yes	Yes	Sometimes	Sometimes
Representation of (part of) the external environment	Yes	Yes	Yes	Sometimes	Sometimes
Core component of an AI system	Sometimes (model-based only)	Yes	Yes	Yes	Yes
Abstract mapping between system input and output	No (Agent Function)	No (AI Operational Logic)	Sometimes	Yes	Yes
Policy/Q-function of a model-free agent	No	No	No	Yes	Yes

Regulatory Implications

The OECD itself acknowledges the indeterminacy surrounding core concepts, noting that "different interpretations of the word 'model' exist." While such open-textured definitions are appropriate within a non-binding, standards-setting context, they acquire regulatory significance when subsequently referenced in binding legal instruments.

In the same document, the OECD explicitly disclaims any engagement with liability allocation, stating that its definition of an AI system "intentionally does not address the issue of liability and responsibility for AI systems," which it leaves to human actors and jurisdiction-specific regulatory choices.

This creates a structural challenge when OECD-derived taxonomies are repurposed within the EU AI Act, which must operationalise distinctions between *AI models* and *AI systems* in order to assign legal obligations. The result is a potential misalignment when a deliberately non-normative technical taxonomy is referenced within a normative legal framework.

Overall, the progression from 2019 to 2024 demonstrates how well-intentioned simplification can increase conceptual ambiguity. OECD's acknowledgment of "different interpretations" may sit in tension with the precise boundaries that legal instruments required to allocate liability, determine compliance, and enable enforcement.

This analysis motivates our proposed framework in Section C (Developing Definitions for AI Models and AI Systems), which provides technically accurate and legally robust definitions capable of withstanding regulatory scrutiny while reflecting the architectural realities of modern AI systems.

C. Developing Definitions for AI Models and AI Systems

We now turn to the constructive task of developing improved definitions for *AI model* and *AI system*. Section B (Findings and Analysis), particularly Section B.3.2 (Analysis of OECD Frameworks) identified several deficiencies in existing definitions: the conflation of distinct conceptual categories (representation versus operational component), internal inconsistencies between textual definitions and accompanying diagrams, the progressive expansion of terms to accommodate technological developments without corresponding conceptual clarification, and the absence of clear criteria for distinguishing where a model ends and a system begins.

In the case of both *AI model* and *AI system* we argue that these problems cannot be solved with a single, all-encompassing definition, and instead propose two separate but complementary definitions. First, we provide conceptual definitions that capture what AI models and systems fundamentally are, providing clarity about the nature of these entities and their relationship to one another. Second, we provide operational definitions that enable consistent, practical application in the real world, allowing non-specialists to determine whether a given artifact falls within the scope of a particular definition. This two-level approach draws inspiration from Tomić & Štimac (2025), who distinguish between conceptual and operational definitions in the context of AI regulation. However, their analysis is primarily concerned with operationalizing the concept of an AI system itself by identifying its invariant features, whereas our focus is to clarify the boundary between AI models and AI systems.

The following sections develop both types of definition, beginning with an examination of the criteria that good definitions should satisfy, proceeding to our proposed conceptual definitions, followed by our operational definitions specifically tailored to the dominant paradigm of neural network-based machine learning systems, and concluding by discussing regulatory implications and case studies.

C.1. Criteria for Technically and Legally Robust Definitions

Definitions exist for a variety of purposes. They can be meant for coordinating usage and ensuring that interlocutors are referring to the same phenomenon (DeLancey, 2017; Copi et al., 2014). They can be used to stipulate a term's meaning for the purposes of a specific inquiry or to refine an imprecise expression into a form suitable for analytical or regulatory reasoning, as discussed in analytic philosophy and legislative drafting scholarship (Gupta & Mackereth, 2023). Definitions can also be used to create or stabilise a concept by naming it, turning a diffuse or loosely recognised phenomenon into an object of inquiry (Burgess et al., 2020). They can clarify a concept's content by identifying the features that constitute it and distinguishing it from related concepts (Margolis & Laurence, 2023). Finally, in law and regulation, definitions serve a normative and operational function: they simplify language, secure consistent interpretation across a text, and determine the boundaries of obligations, rights, and responsibilities (Jopek-Bosiacka, 2011; Bailey & Norbury, 2025). Because AI-related terms sit at the intersection of these different functions (descriptive, conceptual, and normative) this section focuses on the latter purposes and on how they can best contribute to clearer regulation.

ISO 704:2022 (ISO, 2022) outlines principles for defining technical concepts in a precise, systematically structured way. It requires definitions to identify indispensable features that distinguish a concept from related ones, while excluding variable non-essential attributes. Therefore, according to ISO 704:2002, extensional definitions that attempt to specify a concept by listing its instances are inadequate for technical and conceptual clarity. The norm requires that the distinguishing characters of similar or adjacent terms are explicitly listed, clarifying the boundaries within the broader conceptual structural ontology of the domain (what ISO 704:2002 calls a "concept system"). Specifically, definitions should follow a specific pattern where the broader category to which the concept belongs is first identified, then differentiating characteristics from neighboring concepts is further specified.

Definitions can be written for various purposes, such as clarifying research scope or explaining technical concepts. A technical definition primarily captures meaning and facilitates communication within a domain, and might not fit the

criteria we can expect from a legal definition, which must establish clear, enforceable boundaries that support consistent application in practice. Below, we describe properties of a definition suitable for this regulatory function, particularly when distinguishing between AI models and AI systems.

Structural Clarity and Definitional Boundaries

A legal definition should express necessary and sufficient conditions in plain, unambiguous language, avoiding circularity and unstated background assumptions (Dickerson, 1977; House Legislative Services, 2015). It must enable a non-specialist decision-maker to determine inclusion or exclusion from scope using the text alone. At the same time, a definition should establish the outer boundary of application while leaving detailed classifications or risk tiers to secondary instruments. This layered and calibrated structure allows precision to evolve without reopening primary legislation and prevents the definition from becoming overly rigid or granular (Ebers, 2024; Renda & Engler, 2023).

Technical Coherence and Adaptability

Legal definitions in fast-changing domains must both reflect and withstand technological evolution. They should be technology-neutral, describing functions rather than methods. For example, they should focus on what a system does rather than how it operates (Reed, 2007; Greenberg, 2016; Birnhack, 2012). Because underlying technologies develop rapidly, definitions should also include mechanisms for future-proofing: delegated standards, guidance documents, or periodic reviews. These mechanisms should enable reinterpretation without statutory amendment (Ebers, 2024), and remain coherent with the technical object it describes, even as that object changes over time.

Legal Operability and Policy Alignment

The point of a good legal definition is not only to describe phenomena accurately, but also to work as a legal instrument. It should be administrable and enforceable, enabling authorities and courts to apply it consistently without deep technical expertise. Each definitional element should correspond to observable or documentable features (for instance, the presence of a learning process or adaptive behavior), allowing compliance to be verified objectively. Schuett (2021) observes that many AI definitions fail such tests because they rely on vague or colloquial descriptors. It should also be interoperable with existing technical and international vocabularies to minimise conflict between legal and engineering interpretations. The EU AI Act's definition of an AI system, for instance, was intentionally aligned with the OECD wording (Fernández-Llorca et al., 2025; Nahra et al., 2024). Finally, the definition should exhibit policy fit and proportionality, including only the systems the law intends to govern and excluding those that would make enforcement disproportionate (Ebers, 2024; Renda & Engler, 2023). In practice, this means drawing the boundary of AI systems where risk-based obligations attach, not where the term is used informally.

Additionally, where appropriate, providing examples and explicit carve-outs clarifies intent and prevents disputes at the margins. Legislative drafting manuals recommend this practice to ensure shared understanding among implementers (Dickerson, 1977; House Legislative Services, 2015).

These criteria show that the purpose of a legal definition for a technical concept such as *AI model* or *AI system* go beyond describing, as they must also establish enforceable boundaries. A good definition has to draw stable, policy-aligned lines that regulators can administer consistently and that remain intelligible and applicable as technology develops.

C.2. Conceptual Definitions

As described in the Section B (Findings and Analysis), we can see that the conceptual frameworks behind AI models and systems have changed over time. In this section, we attempt to flesh out useful conceptual definitions of AI models and AI systems. Note that these proposed definitions are not operational definitions which we will discuss further in Section C.3 (Operational Definitions for Contemporary AI). Our conceptual definitions are not recommended for technical use, but help to explain the motivation behind our operational definitions, and outline the principles according to which the operational definitions could be further extended if needed. We propose the following conceptual definitions, largely adopted from OECD (2024):

> **AI model**: A computational construct that makes inferences from inputs to produce outputs and forms (or is intended to form) the reasoning or knowledge core of an AI system.

AI system: A machine-based system that makes inferences from inputs from physical or virtual environments to produce outputs that influence physical or virtual environments.

Our definition of *AI model* represents what we see as the conceptual shift in the usage of the term from "model as representation" to "model as core reasoning/knowledge component" as discussed in more detail in Section B.2.1 (Conceptual Senses of *AI model*). In particular, we feel that the current usage of the term in industry (particularly as it relates to frontier models, the kind of models where effective legislation is by far the most important) has shifted almost entirely to the "model as core reasoning/knowledge component" perspective. This is primarily because most of the post-training "fine-tuning" that is done to these entities makes them less accurate representations of the datasets on which they were originally trained (for example, they produce far less offensive text than randomly sampling from the dataset). The fact that this does not appear to be considered as making the entity "less of a model" strongly indicates that the sense of "model" being used in frontier AI development now has very little to do with the notion of "representation".

In principle we allow that a strict subset of an AI model may be an AI model, and a strict subset of an AI system may itself be an AI system. Whether real-world techniques such as mixture-of-experts (MoE) architectures should be considered as nested models or not will be discussed in the Section C.3 (Operational Definitions for Contemporary AI). In addition, we allow for the possibility that an AI system may contain multiple AI models. In principle, we consider it possible that an AI system may not clearly contain a distinct component that can be called the AI model, however for all important contemporary systems, such a component clearly exists.

We now explain the reasoning behind each element of these definitions and their implications for the model-system boundary. In practice, given the current paradigm of large neural networks based on transformer architecture, most AI models will start out as representations of some dataset. However, in order for our definition to function as intended, it is important that an entity continues to be an AI model even if it is not used for the purpose of representation, and even if it is modified in ways that make it a less accurate representation.

According to our definition, AI models are core components of AI systems. However, the key difference is that AI systems receive input from and influence external (real or virtual) environments, which AI models cannot do. Therefore, according to our definition, an AI model can never be a system by itself, since it lacks the necessary properties. This puts us in line with the EU's view on models and systems (European Parliament and Council of the European Union, 2024, Recital 97).

Our definition of AI system is a simplified version of the definition found in OECD (2024). We used OECD as a basis because their definition of AI system already effectively captured the core concept behind the term, and that it is sensible to re-use existing definitions where possible, to aid compatibility with existing work. We simplified the definition slightly (mostly by removing examples) to make it easier to read. We are not concerned by the loss of technical precision or applicability, since this is the role of the operational definition, which will be presented later

We recognise that the boundary between AI systems and systems that are not considered AI can be fuzzy. Since the Dartmouth Workshop in 1956, which is said to be the inception of AI (Russell & Norvig, 2009), there have been several paradigms and trends of AI development, some of which we have described in Section B.2 (Paradigm Shifts in the Meaning of *AI model* and *AI system*). Notably, an earlier paradigm known as Good Old-Fashioned AI (GOFAI) mainly uses symbol manipulation explicitly defined by rules, and AIs developed in this manner are referred to as symbolic AI ISO/IEC (2022c, ISO/IEC 22989:2022 3.1.33). After the 21st century, especially after the creation of AlexNet in 2012, the dominant paradigm has been machine learning, where model parameters are optimised through computational techniques (ISO/IEC, 2022e, ISO/IEC 22989:2022 3.3.5). We elaborate on these paradigms below.

These two main paradigms of AI do not have a clean boundary between them, as they have evolved slowly over decades. In addition, modern hybrid systems (e.g., neurosymbolic AI) increasingly combine learned statistical models with symbolic reasoning mechanisms, further blurring the boundaries between these approaches. Our intention is that our conceptual definition of *AI system* covers both approaches, making it as widely applicable and future-proof as possible. Our operational definition (given in the next section) focuses more heavily on the current paradigm of AI.

Given the historical context of the term AI, which is to create intelligence artificially (i.e. non-human intelligence), and with the different paradigms that the field of AI went through, we find it challenging to come up with a clean conceptual definition of AI systems and AI models.

Nevertheless, for AI systems, we believe it is important that, in line with cybernetic principles which preceded the field of AI, AI systems should carry the concept of producing outputs from inputs, where it would interact with its environment.

C.3. Operational Definitions for Contemporary AI

Having a conceptual definition helps us understand what AI systems and AI models are, but it does not necessarily draw clean boundaries in a way that makes it clear what counts as AI systems or AI models. In this section, we attempt to develop an operational definition such that the boundary between an AI model and an AI system is clear.

According to the American Psychological Association (2018), an operational definition is "a description of something in terms of the operations (procedures, actions, or processes) by which it could be observed and measured". For example, Alan Turing's proposed "imitation game" (Turing, 1950), now widely known as the Turing test, is considered to be an operational definition of machine intelligence (French, 2012). In other words, while a conceptual definition helps us understand the ideas behind a concept, an operational definition allows us to reliably identify instances of that concept in the real world.

As discussed previously, there are many types of AIs, including symbolic AI, machine learning, and neurosymbolic systems. The question of where the boundary between the model and the system lies is thus applicable across the different types of AI. However, because the dominant paradigm in current high-impact and general-purpose AI deployments is neural network-based machine learning AI such as LLMs, the operational definitions developed here are explicitly scoped to this class of systems, since at present these are the systems most in need of effective regulation. We therefore focus on contemporary general-purpose AI systems such as transformer-based LLMs. These operational definitions are not intended to exhaustively characterise model-system boundaries for other types of AIs. This means that our operational definition is much narrower and less future-proof than our conceptual definition, while being much easier and clearer to apply in practice. This is why we provide both a conceptual and operational definition, as the conceptual definition can be used to guide future updates to the operational definition.

We propose our operational definition in this section, summarised as follows:

> **AI model** (for contemporary AI): Trained parameters and their architecture, understood as the computational specification necessary to use them for inference.
>
> **AI system** (for contemporary AI): One or more AI models, an interface for receiving inputs from and delivering outputs to an environment, and the configuration connecting these and any additional components.

In the rest of this section, we explain the rationale behind establishing these operational definitions.

Based on both the systematic literature review and the manual review as well as the observed definitional developments in Section B (Findings and Analysis), we have found the OECD reports to have been the most informative and influential in developing such definitions. Hence, here we attempt to use definitions from the OECD reports as a starting point, and further operationalise it to clarify the boundary between AI systems and AI models.

OECD (2024) defines AI system as follows:

> "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." (p. 4)

Although the report does not explicitly provide a definition of an AI model, they describe it in several paragraphs, including the following:

> "An AI model is a core component of an AI system used to make inferences from inputs to produce outputs. It is important to note that while the parameters of an AI model change during the build phase, they usually remain fixed after deployment once the build phase has concluded" (p. 8)

Several points follow from the OECD description. First, an AI system is described as inferring "from the input it receives," implying that inputs originate outside the system and are not constitutive parts of it. Second, an AI model is characterised as "a core component of an AI system" which means it is considered a subset of the system rather than a representation of it. This means all parts of the model are necessarily parts of the system, but not vice-versa. Third, the model is the component

responsible for performing inference. While the OECD's phrasing is somewhat ambiguous (it is unclear if the AI model or the AI system is the one making inference), we take this to mean that the model is the component responsible for performing inference, and that the system possesses this capability by virtue of containing the model.

Nevertheless, we can draw relationships between an AI system, AI model, and the inputs and outputs, as shown in Figure 8. This would be similar to the diagrams that OECD uses.

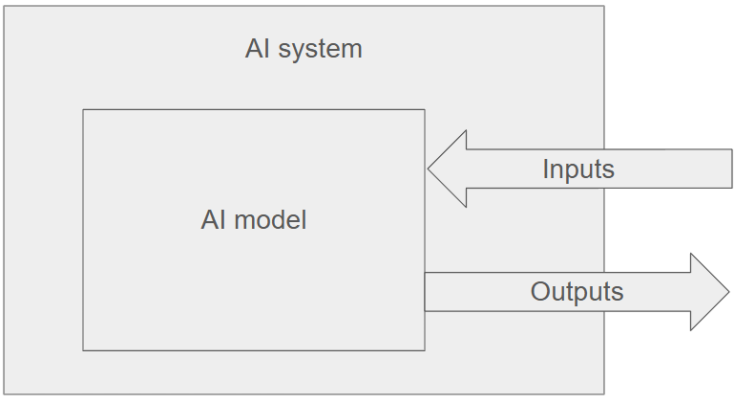


Figure 8. AI system containing an AI model

In most cases, however, the AI system would contain other components that are not part of the AI model, which we elaborate further below. This can be illustrated in Figure 9.

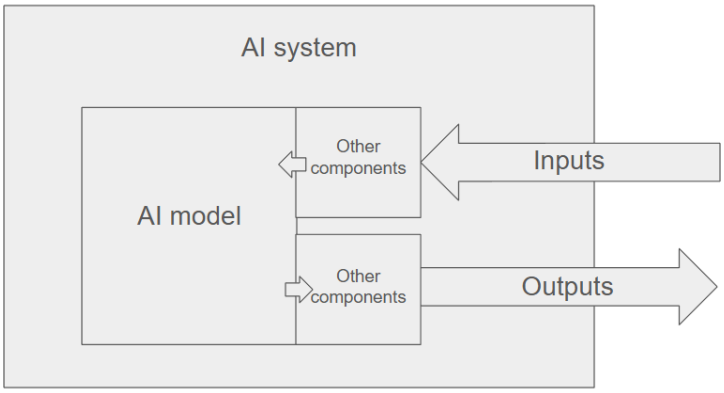


Figure 9. AI system containing an AI model with other non-model components

The description of an AI model being "a core component of an AI system" alone does not disambiguate what parts of a system are a part of the model. We must then rely on the additional elaboration, where the OECD notes that "while the parameters of an AI model change during the build phase, they usually remain fixed after deployment once the build phase has concluded". This suggests two additional properties that can be operationalised as criteria for AI models: its parameters change during the build phase, and that these parameters usually remain fixed after deployment.

OECD does not specify an operational test for "core component." In applying the definition, we interpret "core" as being necessary for the AI system to perform inference, and "component" as that it is part of the system's configuration rather than a transient computational state.

Also referenced above are the "parameters" and the "build phase." While the "build phase" is not defined in the same OECD (2024) framework, it is described in OECD (2022) which references (OECD, 2019b), which states "the model-building process is often called "training" or "optimisation" (OECD, 2022, p. 46).

In the context of machine learning, the process of training updates the model parameters to achieve the training objectives. Parameters can also be defined as the mathematical objects that can be changed by the gradient descent process during training (LeCun et al., 2015).

Interpreting the OECD description operationally yields three candidate criteria for identifying what counts as part of an AI model:

1. Core component of AI system to make inferences
2. Parameters change during the build phase, i.e. training phase
3. Usually remain fixed after deployment

The concepts of "build phase," "parameters," and "after deployment" are compatible with machine-learning-based AI. We therefore use the three criteria above as operational criteria for determining whether a given element belongs to the AI model or not. We apply these rules below to elements of a transformer-based LLM system, including weights, input prompts, activations, system prompts, retrieval-augmented generation (RAG), temperature settings, input or output filters, and fine-tuning.

1. **Weights (and biases, where available)**: In the transformer architecture, these include embeddings and unembeddings weights, attention weights, feed-forward network weights, and layer normalisation weights. They are certainly a core component of AI systems, where these parameters change during the build phase, and usually remain fixed after deployment (unless the model undergoes continuous learning post-deployment). As such, weights are certainly part of the AI model, and they are often referred to as "model weights".
2. **Input prompt**: Inputs are processed by the system but are not part of the system's configuration. They are runtime data rather than system components and are not made of parameters. Thus, they are not part of the AI model.
3. **Activations**: These are intermediate computational states produced during forward passes as a function of inputs and parameters. Although they are necessary for performing inference, they are not themselves trained parameters and do not persist as part of the system after computation. While they are usually referred to as "model activations", they are not part of the system's configuration, as they are results of intermediate calculations performed during model inference.
4. **System prompt**: As inputs to the AI model (and correspondingly, the AI system), system prompts can be important components of an AI system as it greatly influences how it makes inferences. A system prompt stored as part of deployment configuration can be a core component of the system, but it is not made of parameters and thus not part of the AI model.
5. **Retrieval-augmented generation (RAG)**: As a technique that takes in additional sources beyond the input or system prompts to perform model inference, it does not consist of parameters and therefore is not part of the AI model. The associated knowledge base or retrieval database may constitute a component of the AI system if it is integrated into deployment, but it contains data rather than trained parameters. Hence, RAG databases are treated as system components but not as parts of the AI model.
6. **Temperature**: As a setting that determines which tokens is being selected as output, this is not a component but a model setting. They do not contain parameters, and can be changed after deployment.
7. **Input or output filters**: These can be said to be components that modify inputs before it is being processed by the model, or modify model results before it is being output by the system. These filters can be rule-based or produced using a training process. In the latter case, they would then contain parameters that are updated during training.
8. **Fine-tuning**: As a training process that modifies model parameters, it is not itself a component. However, certain parameter-efficient fine-tuning (PEFT) methods such as low-rank adaptation (LoRA), produce trained parameters in an adapter that is then "attached" to a frozen base model. These adapters can be argued as a core component of an AI system, which contains parameters that are trained and usually do not change after deployment.

These are summarised in Table 8.

Table 8. Assessment of whether an element is part of an AI model according to the three candidate criteria from OECD (2024)

	Core component of AI system to make inferences	**Parameters change during the build phase**	**Usually remain fixed after deployment**	**Part of AI model?**
Model weights	Yes	Yes	Yes	Yes
Input prompt	Yes	Not parameters	Not parameters	No
Activations	Not a component - they are dynamic intermediate computation results	Not parameters	Not parameters	No
System prompt	Not a component - it's an input of the model	Not parameters	Not parameters	No
RAG	Not a component - it's a technique; but the associated knowledge base used by RAG can be a core component of the system	Not parameters	Not parameters	No
Temperature setting	Not a component - it's a setting	Not parameters	Not parameters	No
Input / output filters	Component, but ambiguous if core	It depends on whether it has parameters	Yes if they are parameters	Maybe, depending on whether it has parameters
Parameter-efficient fine-tuning (PEFT), e.g. LoRA	Yes	Yes, even though it may not be built together with the rest of the model	Yes	Yes

We can now see a fairly principled way of determining if something is part of an AI model or not. Nevertheless, we believe these conditions can be simplified. With regards to the point on "parameters change during the build phase", where the "build phase" refers to model training, we can simplify the criteria of being considered part of the AI model as being "trained parameters". In short, the definition of an AI model would contain two criteria: being a component of the AI system, and being trained parameters.

Table 9. Assessment of whether an element is part of an AI model according to our candidate criteria

	Component of AI system	**Trained parameters**	**Part of AI model?**

Model weights	Yes	Yes	Yes
Input prompt	No	No	No
Activations	Not really	No	No
System prompt	No	No	No
RAG	No for the technique itself; yes for the associated knowledge base, if implemented as part of the system	No	No
Temperature setting	No	No	No
Input / output filters	Yes	Yes in some cases	Yes in some cases
Dumb classifier	Yes	No	No
Parameter-efficient fine-tuning (PEFT), e.g. LoRA	Yes, if it is a component e.g. adapter	Yes, if it is a component e.g. adapter	Yes, if it is a component e.g. adapter

A notable omission from the list above is model architecture. Architecture is not itself a set of trained parameters, and in its broad sense (e.g., "transformer" or "convolutional neural network") it denotes only a general design family rather than a specific model. However, in a narrower technical sense, architecture refers to the full computational graph specification, including the precise topology, connections, and operations that determine how trained parameters transform inputs into outputs. It is this narrower sense that is relevant here.

Under our conceptual definition, an AI model is something capable of performing inference. Trained parameters alone cannot perform inference without a computational specification that determines how they are applied. Accordingly, model architecture in this technical sense must be treated as part of the AI model, as it provides the necessary computational structure through which trained parameters are used to generate outputs.

This definition provides a principled way of determining whether something is part of an AI model. It also helps clarify when multiple trained parts belong to the same model. When several trained parts work together to produce a single inference and are not usable on their own to perform that inference, they are treated as parts of one AI model rather than as separate models.

Under this approach, PEFT adapters such as LoRA are considered to be part of the AI model, since they adjust the model's parameters and are not used on their own. Likewise, instances of mixture-of-experts (MoE) architectures (including weights of all the experts) can be treated as a single AI model, because the experts and routing mechanisms work together to produce one inference and are not designed to operate separately.

A further implication of treating trained parameters as constitutive of the model is that the underlying data used to produce those parameters (i.e., the training data) should be considered as falling within the scope of the model, since it directly shapes the model's trained parameters. By contrast, data that is merely connected to or retrieved by the model at runtime as part of the wider system (e.g., via RAG) should be considered as falling within the scope of the wider system rather than the model itself.

Stepping back to the system level, the distinction between models and systems becomes clearer. An AI model consists of trained parameters and their architecture and is capable of performing inference. An AI system, by contrast, includes not only one or more such models but also the surrounding components that receive inputs and produce outputs in relation to an environment. This includes interfaces, configuration choices, and any additional components that are part of the system's overall functioning.

This distinction becomes especially clear in the case of AI agents, which are typically designed to interact with an environment, maintain state or memory across steps, and select actions in pursuit of goals. Such systems generally include one or more AI models alongside additional components such as memory stores, planning modules, or tool-use mechanisms. Under the definitions proposed here, these assemblies clearly qualify as AI systems, as they combine models with interfaces and other components that enable interaction with an environment.

C.4. Comparison Against Existing Definitions

As per the previous sections, our proposed definitions and clarifications on the terminologies are relevant to some of the conceptual perspectives and not others. We summarise our proposed definitions in Table 10.

Table 10. Our proposed conceptual and summarised operational definitions of *AI model* and *AI system*

	Conceptual definition	**Summarised operational definition for contemporary AI**
AI model	A computational construct that makes inferences from inputs to produce outputs and forms (or is intended to form) the reasoning or knowledge core of an AI system.	Trained parameters and their architecture, understood as the computational specification necessary to use them for inference.
AI system	A machine-based system that makes inferences from inputs from physical or virtual environments to produce outputs that influence physical or virtual environments.	One or more AI models, an interface for receiving inputs from and delivering outputs to an environment, and the configuration connecting these and any additional components.

In defining an AI model, we use the conceptual perspective of treating the model as an artifact resulting from training (i.e. trained parameters) instead of treating the model as a representation of some external structure. As we are mainly concerned with contemporary general-purpose AI which are predominantly ML models, we operationally define a model as the result of the training process, which may be a definition that may not be appropriate for AI models that are not built through a training process.

Additionally, we do not define an AI model in terms of its task-level function (e.g., prediction, recommendation, classification), since these functions are often properties of the system as a whole. While specifying particular functions may be suitable for defining specific AI systems, such as how Article 3(1) of the EU AI Act defines an AI system as being able to generate outputs "such as predictions, content, recommendations, or decisions", this specificity may not be helpful for defining AI models. This is because task-level functions of a system may arise from the interaction between the model and other system components and may not arise solely from the model itself. Hence, we define a model only by the general function of being able to perform inference.

For the relationship between the AI model and the AI system, we use one as a subset of the other, where the model is the subset of the system, and the system is the one that deals with the environment by taking in inputs and spitting out outputs. This is similar to but in contrast with Sharkey et al. (2024) who defines an AI model as a component of the AI system, but states that AI models can also be AI systems that have no other system parameters. We posit a model, considered in isolation, does not constitute a perception-action loop with an environment, as environmental interaction arises only at the system level. This is in line with Recital 97 of the EU AI Act which says "although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for example a user interface, to become AI systems" (European Parliament and Council of the European Union, 2024).

C.5. Regulatory Implications and Recommendations

In this section, we discuss regulatory implications and our recommendations both generally as well as specifically in the EU-context.

C.5.1. General Implications and Recommendations

Regulatory frameworks implicitly assume a supply-chain structure. The distinction between AI models and AI systems is valuable insofar as it corresponds to separable stages of design and creation of components, integration into products, and deployment into consumer markets that are often handled by different actors with different capacities to foresee and mitigate risk. Its legal relevance therefore lies not in the terminology itself, but in how it helps map obligations onto real divisions of responsibility in the AI value chain.

While some jurisdictions explicitly rely on a distinction between AI models and AI systems, notably the EU, other jurisdictions may frame similar separations albeit using different categories such as components, services, or applications. This issue becomes more salient with the rise of AI agents and agentic AI, which are often implemented by layering tools, memory, and other components onto existing or augmented models or systems (Sapkota et al., 2025), often by different actors in a modular manner. Across these variations in terminology and architecture, the key regulatory question is whether an artefact constitutes a separable or additional unit associated with a distinct actor and a distinct locus of control.

Ambiguity in the boundary between AI models and AI systems or similar terminologies thus creates predictable regulatory failure modes, including misallocated obligations, duplicative compliance burdens, and enforcement gaps. As such, definitions should have sufficient clarity to result in regulation that clearly spells out what is regulated, which actor is responsible, and how obligations are triggered when specific modifications are performed.

Our proposed operational definition of an AI model is deliberately narrow, identifying only the trained parameters and their architecture. In practice, however, commercial providers often place on the market a product that bundles the bare model with pre-specified components such as safety filters, content moderation layers, and system prompts, and refer to the whole package as a "model." This terminological variation is reflected even in provider documentation which have evolved over time. OpenAI published a "Model Card" for GPT-3 (OpenAI, 2020) but shifted to "System Card" from GPT-4 onward (OpenAI, 2023); Anthropic similarly used "Model Card" through Claude 3.5 (Anthropic, 2024) before adopting "System Card" from Claude 3.7 (Anthropic, 2025). The terminologies for these documentations are updated as their products appear to expand from bare models to include other system components. On the other hand, Google DeepMind's Gemini 3 Pro "Model Card" explicitly describes "product-level mitigations such as safety filtering," acknowledging that what is being documented extends beyond the model while still labelling the document as a Model Card (Google DeepMind, 2025).

More broadly, different parties along the value chain can and should draw their own contractual lines to allocate responsibilities. This is acceptable as long as some identifiable entity is designated as bearing each relevant obligation, and the precise boundary between a model and a system is only instrumental to ensuring no responsibilities or obligations fall between the cracks. Definitions should not be treated as ends in themselves; they serve a regulatory purpose, and if technological practice shifts enough that a given boundary no longer maps onto meaningful differences in control or responsibility, the definition should be revised or abandoned.

Finally, we offer three principles for how regulatory definitions should relate to technical terminology. First, regulatory definitions should not deviate from technical consensus. Legal categories ultimately map onto technical artefacts in the real world, and a definition that is internally coherent from a drafting perspective may fail its purpose if it does not correspond to how real-world systems are actually built. We approached our proposed definitions with this consideration as a priority. Second, where a technically grounded definition is too narrow for a regulation's objectives, the appropriate response is to expand the regulatory scope explicitly through carve-ins, special provisions, or adapted definitions, rather than to distort the underlying technical term. For example, if regulation targets systems based on autonomy or societal impact, some recommendation systems may fall within scope even without a clearly identifiable AI model. This should be handled through explicit provisions, not by stretching what *model* means. Third, where no existing technical term cleanly captures the intended regulatory object, it may be better to coin a new or more specific term purpose-built for the regulatory context rather than forcing an ill-fitting label into service.

C.5.2. Implications specific to the EU

The EU AI Act (European Parliament and Council of the European Union, 2024) imposes obligations for providers of both high-risk AI systems (Article 16) and general-purpose AI models (Article 53). Providers are not limited to entities who first developed an AI system or an AI model. For high-risk AI systems, Article 25(1)(b) states that entities who make substantial modifications to a high-risk AI system shall be considered to be a provider of a high-risk system and become subject to provider obligations. For general-purpose AI models, a similar implication follows from two provisions. Recital 97 states

that general-purpose AI models may be further modified or fine-tuned into new models, and Article 3(3) defines a provider to include entities who develop a general-purpose AI model and place it on the market. Together, these provisions support the view that placing a modified model on the market may confer provider status.

However, ambiguity remains as to when a modifying entity qualifies as a provider. This stems from two related uncertainties: what counts as an AI model or AI system, and what degree of change counts as a substantial modification or as creating a new GPAI model. As such, the European Commission issued guidelines for providers of GPAI models to clarify the scope of these obligations (European Commission, 2025b, Article 96), alongside criteria for identifying AI models as GPAI models (Section 2.1) and for determining when an entity is considered a GPAI model provider (Section 3). However, because they do not clearly define what constitutes an AI model, uncertainty about modification and provider status persists, creating ambiguities on whether certain modified models are unregulated (Holtman & Chin, 2025).

Several authors have attempted to address the ambiguity surrounding when a modification yields a new model. For example, building on work by Burden et al. (2025), Pacchiardi et al. (2025) proposes a framework for treating modified GPAI models as new ones by assessing behavioral changes, either by direct measurement or by using proxy metrics. On the other hand, Hacker & Holweg (2025) proposes categorizing modifications into insubstantial and substantial modifications, where insubstantial modifications include RAG, fine-tuning with minor adjustments, system prompts; while substantial modifications include fine-tuning with substantial adjustments, reinforcement learning with human or AI feedback (RLHF/RLAIF), and jailbreaking and related interventions. Unfortunately, categorizing modifications by substantiveness merely shifts the ambiguity to the question of when certain modifications are considered substantial. In both proposals, whether a change is a modification of a model is evaluated through proxies concerning intervention or behavior, without first specifying what constitutes the model whose modification is at issue. Our definitions address this prior question of what the model is, even though we do not purport to resolve when a modification is substantial enough to create a new model that comes with the relevant provider obligations.

Our proposed operational definitions aim to help address these ambiguities and support more consistent allocation of responsibilities across the AI value chain. Current definitional gaps create uncertainty about what constitutes an AI model versus general-purpose software, where system boundaries lie when multiple components are integrated, and which entities bear provider obligations in complex deployment scenarios. For instance, when an entity combines a retrieval system with a generative model to create a retrieval-augmented generation application, our operational definition clarifies that the retrieval component alone (which merely retrieves data) does not constitute an AI model, whereas the generative component (which consists of trained parameters) does, thereby clarifying which aspects of modification ought to trigger relevant provider obligations. Similarly, the definitions help allocate responsibilities when models are accessed remotely via APIs and integrated into downstream applications. In such cases, even though a model in itself may have passed through rigorous safety assessments, the system into which it is integrated could behave in ways that are inconsistent with the model's intended performance once additional components are added to the system. Hence, the obligations of the provider of this system can thus be considered separately from those of the underlying model provider.

A related source of ambiguity concerns safety-enhancing components bundled with a model prior to market placement. Article 55(1)(b) of the EU AI Act requires providers of GPAI models with systemic risk to "assess and mitigate possible systemic risks", and Measure 5.1 of the GPAI Code of Practice explicitly lists system-level interventions among appropriate safety mitigations, including input/output filtering, staged API access, and tools for downstream actors (European Commission, 2025a). This makes clear that it is acceptable and even expected for model providers to deploy system-level components to mitigate model-level risks. Under our operational definition, such additions (e.g., output filters, guardrails) are system-level components rather than parts of the model itself. However, the provider may package and market the entire bundle as a single "model." This does not, in our view, change the underlying technical composition, any more than packaging changes the nature of the packaged product. For regulatory purposes, this bundling should not affect the provider's status: if the product being offered clearly contains a model, the provider remains subject to model-provider obligations regardless of what additional components are included. Our recommendation is in-line with that of Chin & Holtman (2025), who propose as a general principle that the provider of a new model should unilaterally and pre-emptively define in their documentation exactly what they consider the boundaries of their model to be.

Notably, our conceptual and operational definition of *AI model* and *AI system* is built upon the definition of OECD 2024, which also served as the basis of the definition in the EU AI Act (Nessler & Wendehorst, 2024). As a result, these definitions can easily be incorporated into existing regulations simply by providing additional clarification rather than requiring amendments. In the EU context, the most practical mechanism for incorporating these refined definitions would

be through updated Commission guidelines issued under Article 96 of the AI Act, such as the guidelines for providers of general-purpose AI models (European Commission, 2025b) issued in July 2025. We believe additional clarity will be helpful for various actors in the value chain, such as providers and deployers for AI systems and GPAI models.

C.6. Case Studies

The following case studies illustrate practical consequences of ambiguities in distinguishing between models and systems, and demonstrate how our proposed operational definitions may contribute to resolving these ambiguities.

C.6.1. AlienChat Dispute

In September 2025, the Xuhui District People's Court in Shanghai convicted the two developers of the AlienChat app of "producing obscene materials for profit" under Chinese criminal law (Zhang, 2026). AlienChat integrated a third-party large language model to generate conversational responses and monetised access through subscriptions. The court found that the chatbot routinely generated sexually explicit content at scale.

At the core of the dispute was not whether the underlying model was capable of producing explicit language, but why such content was regularly produced at scale in deployment. The defendants argued that the ability to generate sexual content was an inherent property of the large language model they had integrated, and that the application merely exposed this pre-existing capability. The prosecution and the court, by contrast, focused on system-level design and deployment choices, including modifications to system prompts, the configuration of safety constraints, and the overall interaction design of the application. These were treated as evidence that the app developers had intentionally shaped the chatbot's behaviour to elicit and sustain explicit interactions, rather than merely passing through the default neutral model outputs.

Our proposed definitions would arrive at the same conclusion, where the addition of non-model components of the system had changed the behavior of the system, even though modifications were not performed on the model. Nevertheless, in this case, the primary contribution of clearer definitions is analytic clarity and evidentiary burden. They help distinguish between harms arising from deployment choices and those arising from latent model capabilities or propensities, which can structure legal and factual analysis with regards to the allocation of liability.

C.6.2. ChatGPT Geofencing in Italy

In 2023, the Italian Data Protection Authority (*Garante per la Protezione dei Dati Personali, Garante*) issued a temporary limitation order on ChatGPT, which operated on OpenAI's GPT-3.5 model, citing concerns about compliance with multiple provisions of the GDPR. Garante found that OpenAI lacked an appropriate legal basis for the collection and processing of personal data used to train the system's underlying algorithms, raising potential infringement issues under the GDPR (Braun, 2023). One issue highlighted in public debate around the case was the difficulty of applying data subject rights, where the relationship between model training data and ongoing service-level processing became conceptually contested.

OpenAI was subsequently fined a total of €15 million in relation to GDPR violations connected to ChatGPT's data processing (Laher, 2025). Although the service was reinstated after OpenAI implemented additional transparency and user-rights mechanisms, the company did not publicly clarify whether personal data used in prior training had been removed or technically isolated. This uncertainty highlights a structural tension in data protection law: regulatory restrictions were applied to the deployed service, while some of the underlying concerns related to historical training practices that shape model behavior.

The order targeted the system level (ChatGPT service in Italy), but the contested data processing was reflected in the model parameters through training. While in this particular case, the provider of both the AI model and the AI system was the same entity i.e. OpenAI, a lack of clarity between the model and system may prove legally challenging had they been provided by different entities. Under our proposed definitions, disputes about personal data used to train the model would concern the model itself, whereas disputes about personal data used as part of the system's operation or accessed by the system at runtime (e.g., via RAG databases) would concern the system. This distinction does not by itself resolve how regulators should respond when personal data are already incorporated into model parameters and cannot be selectively removed. However, it clarifies whether a regulatory measure is directed at data used in model training or system-level data use, which in turn affects which actors are in a position to comply.

C.6.3. Clearview AI Litigation

Clearview AI, a facial recognition company based in the United States, developed a search engine built on an AI system trained using data scraped from publicly available websites, including social media platforms. By 2024, the company's database contained over 50 billion facial images, which were converted into biometric identifiers and stored as feature vectors, which are numerical embeddings that allow the system to match uploaded images against the database (Jung & Kwon, 2024). Regulators across Europe determined that this end-to-end pipeline, from mass data collection to the embedding of facial features, violated privacy rights under GDPR. Similar legal scrutiny followed in Canada, Australia, and the United States. Under the terms of the *ACLU v. Clearview AI* settlement, the restrictions applied to access to the faceprint database, while the facial recognition algorithm itself was not subject to the injunction (ACLU of Illinois, 2022). Although banned from operating in the EU, Australia, and Canada, and barred under the ACLU settlement from providing database access to most private entities nationwide and to any entity in Illinois for five years, Clearview retained the ability to sell its facial recognition algorithm without the accompanying embeddings database to public and private entities elsewhere in the US (ACLU of Illinois, 2022). In the *ACLU v. Clearview AI* settlement, the company admitted no liability (Clearview AI, 2022).

This divergence in regulatory outcomes illustrates that jurisdictions do not consistently regulate both the AI model and the broader system in which it operates. The US settlement left the algorithm untouched, restricting only access to the faceprint database, effectively treating the model and database as separable for remedial purposes. However, authorities in the EU, Canada, and Australia concluded that the embeddings themselves constitute biometric data that remain linked to identifiable individuals, and therefore subject to data protection law.

Our proposed operational definitions do not determine whether embeddings qualify as biometric data, which is a separate legal question. They instead clarify what is being referred to when actors and regulators speak of the *model* versus the *system*, which can matter in disputes involving modular architectures or multiple responsible parties. In this case, regulators differed in whether they treated the database at the system level or the trained model as the object of regulation. Making these layers explicit helps describe such differences in regulatory practice more precisely.

D. Limitations

In this section, we discuss the limitations of the manual and systematic literature review process, as well as limitations of our proposed conceptual and operational definitions.

D.1. Methodological Limitations

Several limitations emerged throughout the literature review process during the different stages of the review that affect the completeness of the collected dataset and the interpretive certainty of the findings. This section discusses the challenges and limitations of both the manual review and the systematic literature review.

D.1.1. Manual Review: Limitations in Screening and Inclusion

Incorrect References

In some cases, different versions of the same regulatory or standards documents were inadvertently referenced, leading to inconsistencies in citations and uncertainty about which definition was operative at the time of publication. This issue is exacerbated by the frequent revision cycles of regulatory drafts and newly updated standards. Relatedly, definitions were occasionally misattributed, due to publications reusing earlier definitions without direct credit. These incorrect references may have impacted the analysis and comparison of different definitions.

Ambiguity in Definitional Boundaries

In many cases, there was a lack of explicit definitional sections in the documents. Several sources discuss AI models and systems descriptively but do not formally define them, and descriptions were often dispersed throughout the text. This would require interpretive judgments about whether a passage constituted a definition or was merely contextual commentary. This may have introduced subjectivity into inclusion decisions and may have led to the exclusion of borderline cases.

Identifying Original Contributions

During the manual extraction of definitions, there was some difficulty in tracing the chain of influence across definitions. Many definitions that appear as standalone entries in different sources are in fact replications or paraphrases of earlier definitions. This limited the ability to assess definitional originality and may have artificially inflated counts of "unique" definitions.

D.1.2. Systematic Literature Review: Limitations in Search and Corpus Construction

Language Constraints and Jurisdictional Representation

The scope of the corpus was constrained by the inclusion and exclusion criteria, which limited the search to materials written in English. While this ensured terminological consistency given the study's focus on the specific phrases *AI model* and *AI system*, it necessarily excluded regulations, standards, and academic literature from jurisdictions where English is not the primary language of publication. Consequently, the final corpus exhibits a predominantly Eurocentric representation, reflecting jurisdictions such as the EU and OECD member states, whose materials are both accessible in English and where these terms are most explicitly formalised. It is important to acknowledge, however, that non-English jurisdictions may use alternative terminology or conceptual framings that could contribute meaningfully to future analyses of how these definitions evolve.

Source Fragmentation

It was observed that relevant regulatory and standards documents were distributed across agency websites, standards repositories, and institutional portals, many of which lacked unified indexing or persistent identifiers. This decentralisation made it difficult to retrieve materials through database-driven searches and often required supplementary manual search and retrieval, leading to reduced reproducibility. As a consequence, results may be skewed towards more accessible or well-indexed documents.

Access Barriers

Certain materials, such as technical standards from ISO were paywalled, preventing full-text access and verification. For these materials, analysis retrieval was usually conducted through limited previews, or references in other sources including excerpts or summaries. This may have resulted in the overrepresentation of open-access publications, whereas closed-access materials yielded fewer definitional artifacts despite being relevant.

Imbalance between AI model vs AI system

Definitions of *AI systems* appeared far more frequently than definitions of *AI models*. As a result, the corpus is inherently skewed toward system-level descriptions. This limited the comparative analysis across the two terms and their definitional boundary, and nuanced changes in the conceptualisation of AI models may have been overlooked.

D.2. Definitional Limitations and Future Work

This work aimed to clarify an important distinction between two concepts that have been used almost interchangeably for the past few years, despite being given different roles in the legislation. The main challenge was to align with common usage in the field, when it is this usage that has been ambiguous. Words do not have an inherent definition, as their meaning comes from usage, and the boundaries we have placed could be deemed arbitrary. Our focus was that these new definitions could disambiguate the concepts of AI models and AI systems, so that regulatory frameworks referencing these concepts could be applied with greater clarity.

However, the definitions do not resolve all ambiguities. Years of inconsistent and overlapping usage of the terms AI models and AI systems, be it in technical literature, standards, and industry practice mean that it is not always clear where the boundaries of these concepts should lie. While improved definitions can reduce ambiguity in many cases, they cannot provide definitive answers to all edge cases. Some questions may not have a single "correct" solution at present, as the field would first have to converge on shared understanding of certain fundamental distinctions. Rather than claiming to provide definitive answers to all edge cases, our work aims to propose operational definitions that reduce ambiguity where possible, while highlighting remaining areas where continued interpretation and debate among regulators, implementers, and the technical community will be necessary.

Furthermore, despite an attempt to be technology neutral and future-proof, it is impossible to predict whether radical shifts

in AI paradigms could give rise to new forms that would not fit neatly within the boundaries drawn by our definitions. Our proposed conceptual and operational definitions are grounded in current understanding of how AI models and systems are structured and deployed. While expect this framework to remain functional for at least 5-10 years considering current AI trajectories and past AI paradigm cycles (Dhar, 2023), future developments may be transformative enough to represent fundamental breaks from current paradigms, and could require reconceptualisation of these fundamental categories. In consequence, mechanisms for periodic review and refinement of regulatory definitions into the governance frameworks that rely on them will be necessary.

Several categories of ambiguity remain unresolved. First, questions about compositional architectures persist: In agent-based systems, should the tools available to the agent be considered part of the system? Likewise, should the data used by a RAG system be considered as part of the system, or external to the system? Second, determining when modifications create a new model rather than an updated version of an existing model remains challenging. While the EU AI Act provides specific thresholds, such as the guideline that modifications using one-third of the original training compute may constitute a new model (European Commission, 2025b), it remains unclear whether compute-based metrics adequately capture the relevant distinctions. The relationship between computational resources expended and functional changes is not straightforward: shallow fine-tuning with minimal compute can produce significant behavioral shifts (Volkov, 2024), such as circumventing safety guardrails, while extensive parameter updates might produce relatively smaller functional changes. Moreover, as the scale of AI training continues to grow, fixed compute thresholds may become less meaningful indicators of whether a model has been sufficiently transformed to warrant treatment as a new regulatory object. Third, technical questions about model identity remain open: does changing activation functions without modifying weights constitute a new model? More fundamentally, when regulatory obligations hinge on these distinctions, what criteria should determine whether a modified system triggers new compliance requirements?