

ELTE Társadalomtudományi Kutatóközpont



**MTA Law Working Papers
2026/14**

**Risk Management and Accountability under the EU AI Act:
From High-Risk AI Systems to Agentic AI**

Mezei Kitti

ISSN 2064-4515

<http://jog.tk.mta.hu/mtalwp>

*ELTE Társadalomtudományi Kutatóközpont – MTA Kiválósági Kutatóhely
ELTE Centre for Social Sciences – MTA Centre of Excellence*

**Risk Management and Accountability under the EU AI Act:
From High-Risk AI Systems to Agentic AI²**

1. Introduction

In the risk-based regulatory design of the AI Act, the intensity of legal intervention is determined by the level and nature of risk: the applicable obligations and legal consequences are calibrated to the system's risk category, ranging from lighter requirements through ex ante compliance obligations to outright prohibition. This approach leaves the identification, assessment and mitigation of concrete risks largely to the regulated actors, while seeking to make that assessment reviewable through documentation and regulatory oversight. A central question is therefore whether the self-assessments conducted by providers of AI systems are substantively reviewable. Because the risks associated with the same technology depend on its intended purpose, deployment context, and the safeguards in place, the regulatory framework cannot rely on ex ante classification alone.

For high-risk AI systems, this approach is reflected in a set of interrelated requirements concerning risk management, data governance, transparency and human oversight. Although formally distinct, these obligations are functionally part of the same lifecycle-based process, running from the identification and mitigation of risks through the transmission of information and human oversight to ex post review. This model is built in part on a co-regulatory logic: provider-side risk management, standardisation, conformity assessment and regulatory supervision operate as interrelated elements. Such arrangements may provide flexibility and draw on private expertise, but they also raise a central accountability problem: regulatory responsibility is distributed across public and private actors, while the legitimacy of the framework depends on whether the exercise of regulatory judgment remains transparent, reviewable and subject to external scrutiny.³ The challenge is therefore to preserve the advantages of private participation in regulatory implementation without weakening the level of accountability expected of public regulation. In the context of AI systems, this requires not only identifying the actor formally responsible for compliance, but also ensuring that the relevant risk management decisions and assumptions can subsequently be reconstructed and assessed. This study examines the AI Act's risk-based approach primarily through the requirements applicable to high-risk AI systems, since it is for these systems that the AI Act lays down lifecycle-wide requirements linking risk management to the ongoing reviewability of compliance.

¹ Senior Research Fellow, ELTE Centre for Social Sciences, Institute for Legal Studies.

² This research was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

³ Ronit Justo-Hanani, 'Risk-Based Approach to EU AI Act: Benefits and Challenges of Co-Regulation' (2026) 9(2) Policy Design and Practice 171, 176–177.

While the literature has examined in detail the AI Act’s risk-based model,⁴ fundamental-rights risk management,⁵ human oversight⁶ and, more recently, the regulatory challenges posed by AI agents,⁷ comparatively less attention has been paid to connecting these strands. This study, therefore, examines how the AI Act’s formal compliance requirements can support accountability when the information, decision-making authority, and technical capacity to intervene needed to manage risk are distributed across multiple actors in the AI value chain. Agentic AI systems constitute a stress test for this problem. Because of their modular architecture, their use of external tools and data sources, and their variable access permissions, their risk profile cannot be derived solely from the properties of the underlying model. Risk assessment must therefore also extend to the system’s current operational configuration. The study advances two related claims. First, the effective operation of the AI Act’s risk-based framework presupposes lifecycle-based risk management accountability. Second, legal obligations must be properly aligned with the relevant information, decision-making authority and technical capacity to intervene. Its contribution lies in translating this control–accountability alignment problem into an AI Act-specific analytical framework that distinguishes legal, informational, decision-making and technical dimensions. The analysis combines a doctrinal and functional examination of the relevant provisions of the AI Act with the literature on AI risk management, accountability and AI agents.

2. The Risk-Based Model and the Problem of Fundamental Rights Risk

Building on the distinction introduced above, the legal risk classification must be distinguished from the concrete risk assessment: the former determines the scope of applicable requirements, while the latter assesses the specific risks arising from the system’s intended purpose, operation, and deployment context.

Under the AI Act, high-risk classification is not determined by technological characteristics alone. Under Article 6, an AI system may qualify as high-risk in two main ways. First, it may fall within the product-safety route, where the AI system functions as a safety component of a product covered by Annex I, or itself constitutes such a product, and the product is subject to a third-party conformity assessment before being placed on the market or put into service. Second, it may fall within one of the use cases listed in Annex III, including biometrics, critical infrastructure, education and vocational training, employment and workers’ management, access to essential private and public services and benefits, law enforcement, migration, asylum and border control management, and the administration of justice and democratic processes. For systems falling under Annex III, however, Article 6(3) allows for an exception from high-

⁴ See Tobias Mahler, ‘Between Risk Management and Proportionality: The Risk-Based Approach in the EU’s Artificial Intelligence Act Proposal’ [2022] *Nordic Yearbook of Law and Informatics* 247; Michael Ebers, ‘Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU’s AI Act’ (2025) 16(2) *European Journal of Risk Regulation* 684.

⁵ See Karen Yeung, ‘Can Risks to Fundamental Rights Arising from AI Systems Be “Managed” Alongside Health and Safety Risks? Implementing Article 9 of the EU AI Act’ (2025) SSRN Working Paper; Claudio Novelli and others, ‘AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act’ (2024) 3 *Digital Society* 13.

⁶ Lena Enqvist, ‘“Human Oversight” in the EU Artificial Intelligence Act: What, When and by Whom?’ (2023) 15(2) *Law, Innovation and Technology* 513; Melanie Fink, ‘Human Oversight under Article 14 of the EU AI Act’ in Gianclaudio Malgieri and others (eds), *AI Act Commentary: A Thematic Analysis* (Hart–Bloomsbury 2026).

⁷ Thomas Ødegård Henriksen and Thomas Hoffmann, ‘Modular and Agentic AI Design, Jailbreaking, and Scheming Behaviours as Challenges to the EU AI Act’ (2026) 18(1) *Law, Innovation and Technology* 141; Luca Nannini and others, ‘AI Agents Under EU Law: A Compliance Architecture for AI Providers’ (2026) arXiv; Philipp Hacker and Matthias Holweg, ‘A Pragmatic Approach to Regulating AI Agents’ (2026) arXiv.

risk classification where the system does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons — including where it does not materially influence the outcome of decision-making — and where at least one of the further conditions set out in that paragraph is met. Article 6(4) requires the provider to document that assessment.⁸ High-risk classification, therefore, depends not only on the system's technological characteristics but also on its intended purpose and regulatory context. It is also distinct from the subsequent assessment of the concrete risks arising from the system's operation and use.

This distinction also matters because the AI Act's risk model extends the health- and safety-centred approach of traditional product-safety regulation. Alongside these risks, the AI Act also incorporates adverse effects on fundamental rights into risk classification and, particularly for high-risk AI systems, into risk management. In doing so, the AI Act retains the risk management logic inherited from product-safety law while adding a fundamental-rights dimension.⁹ This extension is methodologically significant because it requires providers to integrate the logic and techniques of product-safety risk management with the distinct normative framework of fundamental-rights protection.¹⁰

Article 3(2) of the AI Act defines risk as a combination of the probability of occurrence of harm and its severity, while Article 9 requires the identification and analysis of known and reasonably foreseeable risks, as well as the estimation and evaluation of risks arising in the context of the intended purpose and of reasonably foreseeable misuse. The inclusion of fundamental rights, however, raises a distinct methodological problem, since the assessment of fundamental-rights risks cannot be reduced to quantitative estimates of probability and severity alone. Scenario-based approaches likewise emphasise the joint examination of the specific context of use, the values and vulnerabilities at stake, and the available risk-mitigating responses.¹¹

The assessment of fundamental-rights risks often cannot be reduced to a single, quantifiable risk indicator, since it must also examine whether adverse effects disproportionately affect particular persons or groups. The fact that the AI Act brings fundamental-rights risks within the scope of risk management does not, in itself, prescribe a uniform method for identifying and mitigating them; methodological guidance may nonetheless be drawn from risk- and impact assessment standards, as well as from practical guidance supporting the AI Act's implementation.¹² The overall favourable performance of an AI system may, for example, mask the fact that erroneous outputs disproportionately disadvantage a particular group.¹³

The fundamental-rights dimension, therefore, also modifies the method of assessment. Under Article 9(5), the relevant residual risk associated with each hazard, as well as the overall

⁸ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L, 2024/1689 ('AI Act'), annex III, point 4(a). Under the third subparagraph of art 6(3), AI systems listed in Annex III that carry out profiling of natural persons must, in all cases, be considered high-risk.

⁹ AI Act, Recitals 7–8 and 47–48; Article 9(2)(a). See also Marion Ho-Dac, 'The EU AI Act and the Challenge of Protecting Fundamental Rights' (2025) 62(5) *Common Market Law Review* 1299.

¹⁰ Yeung (n 3) 4–5.

¹¹ Novelli and others (n 3) 10–14; Yeung (n 3) 35–38.

¹² ISO/IEC 42005:2025, *Information Technology — Artificial Intelligence (AI) — AI System Impact Assessment*; ISO/IEC 23894:2023, *Information Technology — Artificial Intelligence — Guidance on Risk Management*; ISO/IEC TR 24027:2021, *Information Technology — Artificial Intelligence (AI) — Bias in AI Systems and AI Aided Decision Making*; IEEE 7003-2024, *IEEE Standard for Algorithmic Bias Considerations*; Ministerio para la Transformación Digital y de la Función Pública, *Guide 5: Risk Management — European Artificial Intelligence Act* (2026) 16–27, 36–43.

¹³ European Union Agency for Fundamental Rights (FRA), *Assessing High-Risk Artificial Intelligence: Fundamental Rights Risks* (Publications Office of the European Union 2025) 10–11.

residual risk of the high-risk AI system, must be judged acceptable, yet the AI Act does not specify the standard governing that judgment. Because the concept of risk also encompasses harms to fundamental rights, acceptability cannot be derived simply from a numerical combination of probability and severity. Relevant considerations in that assessment may include the importance of the right at stake, the reversibility of the harm, the vulnerability of those affected, and the social distribution of the risk.¹⁴ The acceptability of fundamental-rights risks, therefore, cannot be assessed independently of the EU fundamental rights framework, in particular the Charter of Fundamental Rights of the European Union and the relevant case law. A system's residual bias, for instance, cannot necessarily be regarded as acceptable merely by reference to a low statistical frequency or a high aggregate accuracy rate if it systematically disadvantages a protected group.¹⁵

Article 9(5) does not, in itself, impose a general duty to state reasons. However, lifecycle-based risk management is accompanied by documentation requirements: the risk management system must be documented, and the technical documentation must include a detailed description of it. In this context, the assessment of acceptability can become substantively reviewable only if the basis of that assessment, the residual risks and the risk-mitigating measures chosen can be reconstructed ex post.¹⁶

The AI Act seeks a high level of protection for health, safety and fundamental rights, while operationalising the assessment of residual risk primarily through provider-side risk management (Article 9), conformity assessment (Article 43) and, where applicable, harmonised standards under Article 40. This provides flexibility, but may also result in fundamental rights and societal considerations being marginalised within an assessment built around technical metrics and predetermined thresholds.¹⁷

The assessment of fundamental-rights risks is therefore also an accountability problem. Where the provider exercises judgment in identifying relevant hazards, selecting appropriate mitigation measures, and assessing the acceptability of residual risk, the basis and normative standard for those decisions must be reconstructable and reviewable ex post.¹⁸ The next question is how the AI Act's specific requirements for high-risk systems connect this exercise of judgment to a risk management and accountability process spanning the system's entire lifecycle.

3. Lifecycle-Based Risk Management and Accountability for High-Risk AI Systems

¹⁴ Fabian Teichmann, 'Risk, Reasonableness and Residual Harm under the EU AI Act: A Conceptual Framework for Proportional Ex-Ante Controls' (2026) 17(2) *European Journal of Risk Regulation* 611, 615–616.

¹⁵ Teichmann (n 12) 615–616.

¹⁶ AI Act Recital 65; Article 9(1), 9(5) and 11(1); Annex IV, point 5.

¹⁷ Carsten Orwat and others, 'Normative Challenges of Risk Regulation of Artificial Intelligence' (2024) 18 *NanoEthics* 11, 13–16.

¹⁸ Himanshu Verma, Kirtan Shyamal Padh and Eva Thelisson, 'Can AI Be Auditable?' in Daniele Quercia and Marios Constantinides (eds), *Operationalising Responsible AI* (Cambridge University Press 2026). The authors identify the accessibility and traceability of decisions and processes, documentation and logging, and the anchoring of assessment to a predetermined normative or legal standard as preconditions of auditability. see further Jennifer Cobbe, Michelle Seng Ah Lee and Jatinder Singh, 'Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems' (FAccT '21, 2021) 599, 603–605. The authors identify targeted documentation and logging as preconditions of reviewability, linked to each stage of the decision-making process; for impact assessments, they emphasise the documentation of the assessment, its outcome and the mitigating measures chosen, and, for evaluation metrics, the recording of how the normative goals applied were defined and operationalised.

The study uses the term risk management accountability to denote a lifecycle-based approach in which decisions concerning the identification, assessment, mitigation and review of risks are documented, reconstructable and verifiable ex post, and in which those decisions are matched by the information, decision-making authority and capacity to intervene needed to manage the risks in question.¹⁹

Under Article 9(1) and (2) of the AI Act, risk management for high-risk AI systems is framed not as a one-off compliance exercise but as a documented, continuous process spanning the entire lifecycle. It encompasses the identification and assessment of risks, their mitigation, the provision of relevant information, human oversight, and the incorporation of operational experience through post-market monitoring. Article 9 is accordingly best understood together with Article 8(1): the requirements applicable to high-risk AI systems — data governance, transparency, human oversight, and the requirements on accuracy, robustness and cybersecurity — must be calibrated to the system’s specific risk profile. The risk management system thus functionally links these otherwise discrete requirements and calibrates their application accordingly.²⁰

Article 9(2)(a) requires the identification and analysis of known and reasonably foreseeable risks to health, safety or fundamental rights associated with the intended purpose, while point (b) requires the estimation and evaluation of risks that may arise in the context of both the intended purpose and reasonably foreseeable misuse. It follows from this wording that Article 9 does not tie the risk-identification duty solely to the provider’s actual knowledge. Account must also be taken of risks that a professionally diligent provider ought to have recognised in light of the state of the art, sectoral experience, incidents involving comparable systems, the intended context of use, or typical user behaviour.²¹ Reasonable foreseeability is a case-specific standard: it does not require consideration of every conceivable deviation, but of situations that can realistically be expected in light of the intended purpose, the context of use, known behavioural patterns and interactions with other systems.²² The treatment of foreseeable misuse is not confined to the risk assessment under Article 9. Article 13 requires the instructions for use to contain relevant performance- and risk-related information, while Article 14 links human oversight to preventing or minimising the risks that remain in the context of both intended purpose and foreseeable misuse. The AI Act thereby links identification, mitigation, the transmission of information and human oversight as successive, mutually reinforcing elements of the same lifecycle-based risk management process.²³

3.1. Risk Mitigation, Information and Intervention

Article 10(1)–(4) of the AI Act lays down data-governance and data-quality obligations for managing these risks. The provider must identify and address biases that may have adverse consequences for health, safety or fundamental rights, or that may lead to discrimination prohibited under EU non-discrimination law. Datasets must be relevant, sufficiently representative, and, to the best extent possible, free of errors and complete in view of the intended purpose.

¹⁹ The study uses the term ‘risk management accountability’ as its own analytical concept, building on the literature connecting AI risk management, the lifecycle perspective and accountability. see OECD, *Advancing Accountability in AI: Governing and Managing Risks Throughout the Lifecycle for Trustworthy AI* (OECD Digital Economy Papers No 349, 2023); Gudela Grote, Sharon K Parker and Kevin Crowston, ‘Taming Artificial Intelligence: A Theory of Control–Accountability Alignment among AI Developers and Users’ (2026) 51(2) *Academy of Management Review* 278.

²⁰ Yeung (n 3) 1–6, 11–18.

²¹ Jonas Schuett, ‘Risk Management in the Artificial Intelligence Act’ (2024) 15(2) *European Journal of Risk Regulation* 367, 375–376.

²² Teichmann (n 12) 617.

²³ *Ibid* 617.

The requirements of representativeness and data quality must, however, be assessed by reference to the specific intended purpose, the persons or groups concerned, and the context of deployment.²⁴ The assessment of algorithmic fairness, therefore, cannot be reduced to an examination of statistical indicators or fairness metrics²⁵: the context of use and the impact on the persons or groups concerned must also be taken into account.²⁶ In this respect, Article 10 illustrates that the risk assessment under Article 9 must also be reflected in concrete technical and organisational measures; data governance is therefore not merely a stand-alone compliance requirement but also a risk-mitigating tool in risk management.²⁷

Risks cannot, however, always be mitigated solely at the level of system design or data. Article 9(5) therefore makes appropriate mitigation and control measures, as well as the provision of information under Article 13, part of risk management, alongside design and development measures. The instructions for use required under Article 13 connect the provider's risk management, development, and validation knowledge with the deployer's operational practice and human oversight; their quality, therefore, directly affects whether the deployer is able to recognise the system's limitations. Generic warnings are not, by themselves, sufficient: the deployer must understand what errors to expect, when their risk increases, and what intervention is warranted.²⁸

Articles 13 and 14 are closely connected. Without adequate information, human oversight may remain merely formal: the capacity to override is of limited practical value to the overseer if he or she does not understand the system's capabilities, limitations, typical failure modes, or the circumstances under which reliability declines. Article 13 primarily provides the information necessary for understanding,²⁹ Article 14 complements this informational function with the capacity to intervene, and, under Article 26, the deployer must assign persons with the necessary competence, training and authority.³⁰ The provider's task is, accordingly, to establish the technical and informational preconditions for oversight, while the deployer must ensure that the assigned persons have appropriate authority and the ability to intervene within the specific organisational context. The risk management process thus extends from identification and mitigation, through the transmission of information and human oversight, to the feeding back of operational experience.

3.2. Logging and Monitoring

²⁴ AI Act Article 10(3) extends the requirement of appropriate statistical properties of the data to the persons or groups concerned, while art 10(4) requires that the specific geographical, contextual, behavioural and functional setting be taken into account.

²⁵ For instance, the fairness metrics 'equalised odds', 'demographic parity' (or statistical parity), 'equal opportunity' and 'treatment equality' are different ratios built upon rates of false positives, false negatives, true positives or true negatives. see Ninareh Mehrabi and others, 'A Survey on Bias and Fairness in Machine Learning' (2021) 54(6) ACM Computing Surveys 115.

²⁶ Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI' (2021) 41 Computer Law & Security Review 105567.

²⁷ Fabian Lütz, 'The AI Act, Gender Equality and Non-Discrimination: What Role for the AI Office?' (2024) 25 ERA Forum; Luca Deck and others, 'Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness' (2024) arXiv.

²⁸ Thomas Gils, Frederic Heymans and Wannes Ooms, 'From Policy to Practice: Prototyping the EU AI Act's Transparency Requirements' (Knowledge Centre Data & Society, January 2024) 20, 22–24.

²⁹ AI Act Article 13(3)(b)(ii)–(iii).

³⁰ AI Act, Article 13(3)(d), 14(3)–(4) and 26(2)–(3). Article 14(4) expressly links the understanding of capabilities and limitations, the interpretation of output, its override, and intervention in the system's operation. Article 26(2), in turn, requires competence, training, adequate authority and support on the part of the persons entrusted with oversight.

By recording information on the system's operation, the logging capabilities required under Article 12 support not only ex post review but also the identification of new risks arising during operation and the incorporation of those risks into the risk management process.³¹

The provider must reassess the risks arising from the system's operation throughout its entire lifecycle and adapt the technical and organisational controls accordingly. New information derived from testing, use, monitoring and incidents must be fed back into the risk management process; this is underpinned in particular by the post-market monitoring and incident-reporting obligations under Article 9(2)(c), Article 17(1)(h)–(i), and Articles 72–73. These provisions do not establish compliance as a static condition but require continuous review and feedback. Post-market monitoring, therefore, serves not only to identify incidents or performance deterioration, but also to test whether the assumptions underlying the initial risk assessment remain valid as the system's performance and operational environment evolve.³² The importance of continuous reassessment is also illustrated by the feedback loops referred to in Recital (67): a system's outputs may influence the inputs to its subsequent operation, thereby perpetuating or amplifying existing biases. Risk assessment must therefore also extend to such patterns as they emerge during operation.

In this context, logging and monitoring provide the informational basis for feedback within the risk management process. Lifecycle-based risk management thus extends from identification and assessment through mitigation, the provision of information, and oversight to monitoring and reassessment. The following chapter examines the conditions under which these requirements secure effective control over the system's operation.

4. From Formal Compliance to Effective Control

4.1. Information as a Precondition for Control

The effective exercise of human oversight requires information that is accessible, interpretable and relevant to the decision at hand; in its absence, the possibility of human intervention may remain merely formal. Effective oversight rests on three preconditions — authority, understanding and operational context: the overseer must have adequate authority, relevant knowledge of the system, and the time, training and decision-making latitude necessary to exercise those powers.³³

Automation bias, however, cannot be attributed solely to the system's technical characteristics. It may also be influenced by the design of the user interface, the manner in which information is presented, the decision-maker's training and workload, and the amount of time and latitude the organisation allows for the review of machine output. The same system may therefore present a different risk profile in different organisational settings. Mandating human oversight does not, by itself, guarantee risk mitigation: its effectiveness also depends on the overseer's knowledge, decision-making latitude, and capacity to intervene.³⁴

4.2. Intervention as a Precondition for Control

³¹ AI Act Article 12(1)–(2), 9(2)(c) and 72(1)–(2).

³² See Michael Clark, 'Post-Market Monitoring for AI Systems: Obligations, Architecture, and Implementation Under the EU AI Act' (2026) SSRN Working Paper.

³³ Fink (n 4) 10–13; Ana Maria Corrêa, Sara Garsia and Abdullah Elbi, 'Better Together? Human Oversight as Means to Achieve Fairness in the European AI Act Governance' (2025) 1 Cambridge Forum on AI: Law and Governance 12.

³⁴ Ben Green, 'The Flaws of Policies Requiring Human Oversight of Government Algorithms' (2022) 45 Computer Law & Security Review.

Information can perform a control function only if it is coupled with the capacity to intervene, which can be exercised in a timely manner. Article 14 links two related functions: monitoring the system's operation and interrupting it where necessary, and interpreting, disregarding, overriding or reversing its output. Oversight is therefore not necessarily reducible to a single point of approval and does not require ex ante human approval of every output. The regulatory significance of intervention depends on whether it can be exercised when the relevant risk can still be prevented, mitigated, or corrected. Routine approval of a machine-generated recommendation does not, in itself, substantively reduce risk, particularly where the overseer lacks adequate information, time or authority.³⁵

4.3. Documentation and Reconstructability

The role of documentation in supporting control depends not merely on the recording of data but on its accessibility, interpretability, and the ex post reviewability of the decision-making process. This requires, in particular, adequate access to documentation and logs, traceability of data, decisions, and processes, and identifiability of relevant system states.³⁶ At this point, the logging and monitoring requirements examined in section 3.2 are relevant not only for feeding risk information back into the process, but also as preconditions for ex post verifiability. Reconstructability is thus one of the preconditions for reviewability: it enables an ex post review to extend beyond the mere existence of documents to the process and the soundness of risk assessment, the mitigating measures chosen, and the decisions taken to intervene.

These functions can usefully be distinguished: documentation and logging support the reconstruction of the underlying decision-making process; this, in turn, enables the verification of compliance-related facts and the substantive review of risk-management judgments.

4.4. The Limits of Formal Compliance

The literature also points out, however, that the AI Act relies to a significant extent on provider self-assessment, documentation and internal compliance processes to support compliance by high-risk systems; although the AI Act grants affected persons certain direct rights, albeit to a limited extent, ex ante compliance mechanisms and regulatory supervision continue to play a decisive role in its enforcement architecture.³⁷ Whereas Article 5(2) GDPR enshrines accountability as an express principle, and Article 24 obliges the controller to ensure and demonstrate compliance, the AI Act contains no comparable general accountability clause. The logic of accountability nonetheless appears indirectly in the AI Act, through the interconnected requirements on risk management, documentation, logging, transparency, human oversight and conformity assessment. A feature common to both regulatory regimes is that compliance is not exhausted by the formal fulfilment of obligations, but also presupposes its demonstrability and ex post verifiability.

This concern also extends to the methodology of risk assessment itself. Where providers retain discretion over how probability, severity and their combination are operationalised, methodological choices may materially affect the resulting risk assessment or classification.

³⁵ Enqvist (n 4) 513, 525. The literature refers to this problem using the term 'liability sponge'; Fink (n 4) 10–13.

³⁶ Verma, Padh and Thelisson (n 16) 133; Jakob Mökander, 'Auditing of AI: Legal, Ethical and Technical Approaches' (2023) 2 Digital Society 49; Jakob Mökander and Luciano Floridi, 'Ethics-Based Auditing to Develop Trustworthy AI' (2021) 31(2) Minds and Machines 323–327.

³⁷ Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22(4) Computer Law Review International 97, 105–106; Céline Castets-Renard and Philippe Besse, 'Ex Ante Accountability of the AI Act: Between Certification and Standardization, in Pursuit of Fundamental Rights in the Country of Compliance' in Céline Castets-Renard and Jessica Eynard (eds), *Artificial Intelligence Law: Between Sectoral Rules and Comprehensive Regime* (Bruylant 2023); Francesca Palmiotto, 'The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation' (2025) 16(2) European Journal of Risk Regulation.

Recent literature describes the strategic use of such choices to obtain a lower risk assessment as “risk hacking”.³⁸

The accountability function of Article 9 depends on whether the provider’s identification of risks, choice of mitigating measures, and judgment on residual risk are substantively reviewable. It is at this point that the relationship between the legal obligation and the preconditions of control becomes significant. Where the information, decision-making authority and technical capacity to intervene needed to discharge an obligation are distributed among different actors, accountability cannot be secured merely by designating the addressee of that obligation. It must also be examined whether the actor subject to the obligation has access to the information necessary to perform the task, possesses adequate decision-making authority, and is able to give effect to the necessary intervention. This problem is particularly visible in systems with a dynamic and modular configuration. The following chapter, therefore, examines agentic AI systems as a stress test for the AI Act’s risk management and control requirements.

5. Agentic AI as a Stress Test for the AI Act’s Risk-Based Model

Agentic AI systems provide a particularly useful stress test for the AI Act’s risk-based model because their modular architecture, use of external tools and data sources, and variable access permissions may cause their risk profile to change over time. Their operation also makes particularly visible the relationship between legal obligations and the information, decision-making authority and technical capacity to intervene required to discharge them. More broadly, their modular and variable configuration highlights limitations in applying a regulatory framework in which assessment is premised on a relatively stable system state.

5.1. The Different Regulatory Levels of the AI Model and the AI System

The AI Act does not define the AI agent, or the agentic AI system, as a separate legal category. According to the non-binding interpretation published on the European Commission’s AI Act Service Desk, a common feature of AI agents is their capacity to process inputs received from their environment and, on that basis, to carry out actions that engage with, or affect the state of, that environment. According to that guidance, an AI agent typically incorporates at least one general-purpose AI model and qualifies as an AI system under Article 3(1) of the AI Act. This study accordingly uses the term agentic AI system to denote an AI system that, building on model-based inference, plans and executes one or more actions in pursuit of a defined goal, in the course of which it may draw on external tools, data sources or services, and whose operation may affect its digital environment.³⁹ Recital (97) of the AI Act distinguishes a general-purpose AI model from an AI system: the model may be an essential component of the system, but it does not, in itself, constitute an AI system, as this requires additional system-level components and functions. The general-purpose AI model and the AI system built upon it likewise represent distinct regulatory levels. Under Recital (100), a system integrating such a model may qualify as a general-purpose AI system where, through that integration, the system itself becomes capable of serving a variety of purposes. An agentic system designed for a specific purpose is therefore not necessarily a general-purpose AI system merely because it uses a general-purpose model. Agentic operation does not, by itself, give rise to high-risk

³⁸ Gustavo Gil Gasiola, Sarah H Cen and Frederike Zufall, ‘Qualifying and Quantifying Risk Under the EU AI Act’ (2026) arXiv.

³⁹ European Commission, ‘How Are AI Agents Addressed within the AI Act?’ (AI Act Service Desk) <https://ai-act-service-desk.ec.europa.eu/en/ai-act/faq/how-are-ai-agents-addressed-within-ai-act-0>.

classification; the system's classification under the AI Act continues to be determined by its intended purpose and the AI Act's risk categories.

This distinction matters for risk management purposes because the risks of the system as a whole cannot be automatically derived from the properties of the underlying model. The same model, embedded in different system architectures and operational configurations, may give rise to different capabilities and risks.

5.2. Modular Architecture and Variable Operational Configuration

The regulatory assessment of agentic systems must also account for their modular architecture. Such a system does not necessarily operate as a single, unchanging technical unit, but may consist of several interconnected components. Such components may include, for example, the underlying model, memory, the orchestration layer, sub-agents, data sources, external tools, or the access permissions attached to them. These elements may be exchanged, reconfigured, removed, or supplemented with further components over the system's lifecycle. Therefore, assessment of the system's risk profile cannot be derived solely from the separate examination of individual components. The connections between components, and their combined operation, may give rise to capabilities and risks that do not follow from any single element in isolation. A technically minor-seeming change — such as the addition of a new tool, the expansion of access permissions, or a modification of the orchestration logic — may materially transform the system's operation, autonomy, the range of actions it can perform, and its risk profile, even where the underlying model remains unchanged.⁴⁰

Because an agent carries out its task in a sequence of interdependent steps, an intermediate error may propagate to later stages of the process. It may also occur that each individual step is acceptable in isolation, yet the sequence of actions as a whole fails to produce an appropriate outcome. Risk assessment must therefore examine the full execution of the task, the connections between actions, the accumulation of errors, and the possible consequences, taken together.⁴¹ The tools and access rights that determine the agent's scope for action, therefore, require separate examination.

5.3. Tools, Access and Permissions

A functional-technical analysis examines which external systems the agent can access and what actions it can perform within them. From this perspective, the AI agent's distinguishing feature lies not only in its capacity to draw inferences but also in its ability to translate those inferences into actions that are executable within external systems. The system's risk is therefore significantly shaped by the range of data, tools, permissions and actions available to the agent.⁴² Tools and permissions are, therefore, independently relevant elements of the risk profile. The same underlying model and system-level logic may pose different risks depending on whether the system has only read access or can modify data, send messages, initiate transactions, or invoke other systems. The range of actions and access available to the AI agent is therefore not merely a technical characteristic; it directly shapes the system's risk profile and the scope of controls required. A change in the available tools or access permissions may materially alter that profile even without any modification to the underlying model, and may require the risk assessment to be reviewed.

5.4. Dynamic Configuration and Lifecycle-Based Risk Assessment

An agentic system's risk profile may continue to change after it has been put into service. For high-risk AI systems, examining this is directly connected to the AI Act's lifecycle-based

⁴⁰ Henriksen and Hoffmann (n 5) 141–163.

⁴¹ Nannini and others (n 5) 29–30; Hacker and Holweg (n 5) 21–22.

⁴² Hacker and Holweg (n 5) 3–5.

approach. Article 9 requires the risk management system to operate as a continuous iterative process throughout the system's entire lifecycle, while Article 72 requires a post-market monitoring system. The addition of a new external tool or data source, the expansion of access permissions, an increase in the level of autonomy, or the repurposing of the system for a new task may alter which persons or groups the system can affect, what data it can use, and what actions it can perform. Such changes are therefore not merely technical modifications: they may have implications for the system's risk profile and for the applicable risk management measures.⁴³

Risk management, therefore, cannot be concluded once the system's initial state has been assessed. In the course of an AI agent's operation, the range of tools, information and execution paths available to it may change, and the risk assessment must be updated to reflect those changes.⁴⁴

In this sense, agentic systems constitute a stress test for the AI Act's lifecycle-based risk model. Compliance cannot be tied exclusively to the underlying model, to the system's initial state, or to a single initial conformity assessment, if subsequent modifications can also alter the system's capabilities and risk profile. The subject matter of risk assessment, therefore, cannot be confined to the underlying model or the system's initial state, but must extend to the system's current operational characteristics, including in particular the relevant components, tools, data sources, access permissions and execution capabilities.

6. The Alignment of Control and Accountability in the AI Value Chain

The analysis of agentic systems has shown that assessing a system's risk profile cannot rest on an examination of the underlying model or of a given system state alone. The next question is how the information, decision-making authority, and technical capacity to intervene needed to manage the system's changing risks are distributed among the actors in the AI value chain, and how they relate to the legal obligations set out in the AI Act.

6.1. Separating Legal Role from the Preconditions of Control

The legal classification of actors participating in the AI value chain under the AI Act must be distinguished from the governance-oriented inquiry into which actors hold the information, decision-making authority, and technical capacity needed to intervene in the system's operation. The former determines the addressees of the AI Act's obligations, while the latter examines how the knowledge, decision-making authority and capacity to intervene needed to discharge those obligations are distributed across the value chain.

The provider of the model typically has more detailed knowledge of the model's capabilities, limitations, and known failure modes, but does not necessarily know into which agentic architecture, with what tools, data sources, and access permissions, the model will be integrated. The provider of the agentic AI system typically has more detailed knowledge of the system's technical design, capabilities and intended purpose, whereas the deployer knows the specific context of use, the available organisational data, the access arrangements and the operating conditions. The knowledge and capacity to intervene necessary for a comprehensive assessment of the system's risks are therefore distributed among several actors; as a result, the

⁴³ Henriksen and Hoffmann (n 5) 151–155.

⁴⁴ Ibid 151–155; Nannini and others (n 5) 15–16.

risk management of agentic systems cannot necessarily be carried out at a single level of the AI value chain.⁴⁵

In complex agentic architectures, the distribution of decision-making, informational, and technical capacities may give rise to what the literature describes as the “many hands problem”.⁴⁶ The system’s operation may be shaped by several third-party actors and technical components, while responsibility and the relevant knowledge are likewise distributed among them. As a result, no single participant necessarily possesses all the information and authority needed for a comprehensive assessment and management of the risks.⁴⁷

6.2. Four Dimensions of the Alignment Inquiry

To analyse this problem, this study distinguishes four interrelated dimensions. The legal dimension identifies which actor the AI Act designates as the addressee of a given obligation. The informational dimension concerns which actor holds the relevant knowledge of the system’s capabilities, limitations, current operational characteristics and specific context of use. The decision-making dimension identifies who is authorised to determine or modify the system’s intended purpose, architecture, tools employed, access permissions, or conditions for human approval. Finally, the technical dimension refers to which actor can restrict the system’s operation, revoke access, block an action, or interrupt its operation.

These four dimensions provide the basis for examining whether a given legal obligation is matched by adequate information, decision-making authority and technical capacity to intervene.

6.3. The Alignment Problem Between Control and Accountability

The literature describes these problems using several related, but distinct, concepts. Responsibility-gap approaches identify different forms in which responsibility may become difficult to attribute or exercise in AI-mediated socio-technical systems.⁴⁸ Hindriks and Veluwenkamp use the term “control gap” to describe situations in which the causal control exercised by an autonomous machine falls short of the guidance control it ought to emulate.⁴⁹ Veluwenkamp’s related concept of “control misalignment” focuses instead on deficiencies in the distribution of control within an artificial or socio-technical system.⁵⁰ Grote, Parker and Crowston, through their concept of “control–accountability alignment”, emphasise that mitigating AI risks requires control and accountability to be properly aligned among the stakeholders involved in AI development and use.⁵¹

Building on these approaches, the present study examines whether the information, decision-making authority and technical capacity to intervene needed to discharge the legal obligations laid down by the AI Act are properly aligned with those obligations. The distribution of these dimensions among several actors does not, in itself, amount to a deficiency. An alignment problem arises when the addressee of a legal obligation lacks the information, decision-making authority, or technical capacity necessary to intervene to discharge the obligation, and that gap

⁴⁵ Kathrin Gardhouse, Amin Oueslati and Noam Kolt, ‘Regulating AI Agents’ (2026) SSRN Working Paper 11–12; Nicola Fabiano, ‘Subject Roles in the EU AI Act: Mapping and Regulatory Implications’ (2025) arXiv 15–16.

⁴⁶ Gardhouse, Oueslati and Kolt (n 43) 62.

⁴⁷ Chris Reed and others, ‘Regulating Autonomous and Agentic AI’ (2026) arXiv 6.

⁴⁸ Filippo Santoni de Sio and Giulio Mecacci, ‘Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them’ (2021) 34 *Philosophy & Technology* 1057.

⁴⁹ Frank Hindriks and Herman Veluwenkamp, ‘The Risks of Autonomous Machines: From Responsibility Gaps to Control Gaps’ (2023) *Synthese* 201.

⁵⁰ Herman Veluwenkamp, ‘What Responsibility Gaps Are and What They Should Be’ (2025) 27 *Ethics and Information Technology*.

⁵¹ Grote, Parker and Crowston (n 17) 278.

is not offset by the flow of information among the actors, an adequate allocation of decision-making authority, or an available means of intervention.

The gap may also arise in the opposite direction: an actor may be capable of intervening in the system's operation without access to the relevant risk information, without possessing the necessary decision-making authority, or without clarity on the conditions under which intervention is required. Agentic systems render this particularly visible, since the relevant information, decision-making authority and technical capacity to intervene may reside with different actors across the AI value chain.

This problem can also arise in the regulation of high-risk AI systems. The duty to provide an explanation under Article 86 rests, in certain cases, on the deployer, while some of the information about the system's operation needed to discharge that duty may be held by the provider. Although Article 13 obliges the provider to supply information enabling the interpretation and proper use of the system's output, the two provisions do not, in every case, expressly resolve how the deployer is to obtain all the information necessary to fulfil the obligation under Article 86. This illustrates how separating the legal obligation from the informational position can raise a distinct accountability question.

Risk management accountability, therefore, does not require that all information, decision-making authority, and technical capacity to intervene be concentrated in a single actor. What matters is that their distribution does not impede the identification, assessment and management of risks.

7. Operationalising Risk Management Accountability

Building on the alignment problem set out in the preceding chapter, this chapter examines what technical and organisational solutions can support the operational implementation of risk management requirements in agentic systems.

In this context, the analysis examines access and execution limits, risk-calibrated human approval, logging and configuration tracking, and the feedback of operational information. These solutions do not, in every case, appear in the AI Act as separately named technical requirements; rather, they illustrate how the AI Act's requirements can be given effect in agentic systems.

7.1. Access and Execution Limits

The risk profile of AI agents is also shaped by the range of accessible data, tools, external systems and executable actions. For risk management purposes, it is therefore appropriate to identify which data, tools, and external systems the AI agent can access, which actions it can perform, and which rights or interests these may affect. In defining permissions, it is also appropriate to take account of the risk implications of the actions that can be performed: accessing information carries a different risk from modifying it, just as preparing an action carries a different risk from executing it autonomously. The effectiveness of defined limits can be strengthened if they are not confined to natural-language instructions or organisational rules governing the AI agent, but are also technically enforced by the execution infrastructure, restricting unauthorised access and actions.⁵²

7.2. Risk-Based Human Approval and Intervention

Human oversight under Article 14 applies to high-risk AI systems. Beyond that, for agentic systems, it may also be appropriate, from a risk management perspective, to determine in

⁵² Nannini and others (n 5) 12–13.

advance which actions may be performed autonomously and which require human approval or other forms of intervention.

Given the speed and multi-step nature of AI agent operation, continuous human review of every execution step is unrealistic; an ex post audit, however, does not provide the same level of oversight. A risk-based approach, therefore, requires determining for which actions, under what circumstances, and by which decision-maker, intervention is warranted. Recording approval, rejection, or intervention can also support ex post verifiability.⁵³

This does not mean that every action must be subject to ex ante human approval. The intensity of control should instead be calibrated to the risk associated with the action. Relevant factors in this regard may include, in particular, the nature and value of the action, the situation or vulnerability of the persons affected, the model's uncertainty, any detected deviations in the system's operation, and the risk pattern emerging from prior actions.⁵⁴ For higher-risk actions, prior human approval may accordingly be warranted.

One technical means of oversight is to make the execution of certain actions conditional on human approval; for instance, initiating a payment or modifying a database might be technically possible only after human confirmation. In such an arrangement, it is not sufficient for the AI agent merely to be instructed to seek permission before certain actions; the system component executing the action must also be designed so that it will not carry out the action in the absence of the required approval. A payment API, for example, can be configured to authorise a transaction exceeding a predetermined risk threshold only when human approval has been obtained and the system has recorded this fact in a verifiable manner.⁵⁵

7.3. Logging and Configuration Tracking

A further precondition for reviewability is that the system's actions, state changes and any related human interventions can be reconstructed ex post. The logging requirement discussed in section 3.2 assumes particular significance for agentic systems because a given outcome is often the product of several interdependent actions, tool calls and intermediate decisions. In multi-agent systems, this may also encompass the handover of tasks between AI agents, as well as the flow of data and permissions between them.⁵⁶ Ex post verifiability, therefore, cannot be confined to recording individual outputs or isolated tool calls; the relevant connections within the execution process must also be reconstructable.⁵⁷

To support reconstructability, it may therefore be appropriate to record the connections between actions: on what information the system based its choice of a particular tool, which earlier steps influenced later actions, under what permissions a given action was performed, whether human approval or intervention occurred, and what external state changes resulted. It may be particularly important that logging does not rely solely on information generated by the AI agent itself, but is performed independently of the agent, for example, by recording API calls, the permissions used, and relevant state changes in external systems.⁵⁸

In dynamically changing agentic systems, in addition to logging individual actions, it may also be necessary to track system changes relevant from a compliance perspective. To this end, it may be appropriate to record, in versioned form, in particular, the models and sub-agents employed, the available tools, the access permissions, the orchestration rules, the compliance-relevant state of memory, and the range of actions subject to human approval. Comparing successive system states can help to establish whether a subsequently observed change in

⁵³ Nannini and others (n 5) 13–14, 16.

⁵⁴ Ibid 28.

⁵⁵ Ibid 28.

⁵⁶ See Lewis Hammond and others, Multi-Agent Risks from Advanced AI (2025) arXiv.

⁵⁷ Nannini and others (n 5) 30.

⁵⁸ Ibid 30.

behaviour or risk can be linked to a technical or operational modification. Configuration tracking can thus, alongside ex post review, indicate whether the system's operational state has diverged from the state on which the earlier risk or conformity assessment was based.⁵⁹

7.4. Monitoring and Reassessment of Controls

Access limits, approval points and logging arrangements should be periodically reviewed in light of new information arising from the system's operation. Monitoring should therefore also extend to new models, sub-agents, tools, data sources, access permissions and changes in autonomy, insofar as these may alter the previously assessed risk profile. Not every configuration change carries the same legal consequence. The significance of a change depends on how it affects the system's intended purpose, autonomy, access, executable actions, risk profile, or regulatory compliance. The combined effect of smaller, interrelated modifications may accordingly also be relevant. Where the system's current state materially diverges from the state previously assessed, it should be assessed whether the access limits, human approval points, logging requirements and other technical or organisational controls remain adequate to the new risk profile. For high-risk AI systems, it should also be assessed separately whether the change may constitute a substantial modification under Article 3(23) of the AI Act.⁶⁰ Information arising from monitoring, incidents, system modifications, and operational experience should therefore inform risk management and, where necessary, lead to adjustments to access, execution, or human-intervention controls.

8. Conclusion

The AI Act's risk-based model extends beyond the classification of systems and the formal fulfilment of compliance requirements. The assessment of risks to fundamental rights also requires normative and legal judgment, the basis of which must remain reconstructable and reviewable ex post. In this sense, risk management accountability requires legal obligations to be matched, across the AI value chain, by adequate information, decision-making authority and capacity to intervene.

Agentic AI systems provide a particularly revealing stress test for this framework. Their modular architecture, changing access rights and capacity for autonomous action illustrate how legal obligations, relevant information, decision-making authority and technical capacity to intervene may be distributed among different actors. Such distribution is not problematic in itself. An accountability problem arises where these elements are insufficiently aligned, so that the identification, assessment or management of relevant risks is impeded.

The analysis therefore suggests that the alignment of control and accountability is an important operational condition for effective risk management under the AI Act. In agentic systems, such alignment may be supported by access and execution limits, risk-calibrated human approval, logging and configuration tracking, and mechanisms for feeding operational information back into the risk-management process.

This does not imply that the AI Act requires the same technical or organisational controls for every agentic AI system. The appropriate measures depend on the system's legal classification, intended purpose, level of autonomy, access rights and specific deployment context. Their significance lies instead in ensuring that the AI Act's requirements on risk management, oversight and ex post reviewability can be given practical effect within the architecture and operational configuration of the system.

⁵⁹ AI Act Article 3(23), 25 and 43; Nannini and others (n 5) 15–16; Henriksen and Hoffmann (n 5) 151–155.

⁶⁰ Fabiano (n 43) 14.

© Mezei Kitti

MTA Law Working Papers

**Kiadó: ELTE Társadalomtudományi Kutatóközpont (MTA
Kiválósági Kutatóhely)**

Székhely: 1097 Budapest, Tóth Kálmán utca 4.

Felelős kiadó: Boda Zsolt főigazgató

Felelős szerkesztő: Kecskés Gábor

Szerkesztőség: Hoffmann Tamás, Lux Ágnes, Mezei Kitti

Honlap: <http://jog.tk.mta.hu/mtalwp>

E-mail: mta.law-wp@tk.mta.hu

ISSN 2064-4515