Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Between legal aspiration and engineering reality: technical and legal barriers to implementing control safeguards for high-risk AI systems

Faztudo Languisse

flanguisse@ispt.ac.mz

Instituto Superior Politécnico de Tete

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11460713/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) is increasingly used in decisions that affect health, safety, welfare and rights, and regulators have responded by requiring “control safeguards”, such as human oversight, robustness and monitoring, for systems classed as high-risk. This article asks why such safeguards are hard to implement and why the difficulty matters for society. It combines doctrinal analysis of the European Union’s Artificial Intelligence Act (Regulation (EU) 2024/1689, as amended in 2026) with a structured critical review of the technical, human-factors and policy literature. Five technical barriers are identified (model opacity, speed and autonomy, brittle guardrails, limits of human attention and judgement, and the lack of tests for control effectiveness), together with six legal and institutional barriers (open-textured legal standards, uncertain classification, delayed standards and guidance, limited regulatory capacity, fragmented enforcement, and burdens on small organisations). The article argues that these barriers reinforce one another through three gaps: a feasibility gap between what the law demands and what technology can deliver, a verification gap between legal requirements and testable criteria, and a capacity gap between large and small actors and between providers and regulators. Mapping each oversight capability in Article 14(4) to these barriers shows where safeguards are most likely to be formal rather than effective. The article proposes ten recommendations for EU institutions, standard-setters, national authorities, providers and deployers, and considers lessons for lower-capacity jurisdictions, including in Africa, that are drawing on EU-style models. The central conclusion is that regulatory attention should shift from documenting oversight to demonstrating that it works.

1. Introduction

Governments around the world have converged on a risk-based logic for regulating artificial intelligence: the greater the potential of a system to harm health, safety or fundamental rights, the more demanding the duties placed on those who build and use it. The European Union’s Artificial Intelligence Act (AI Act), Regulation (EU) 2024/1689, is its most developed expression. For systems classified as high-risk, it requires a risk management system (Article 9), data governance (Article 10), technical documentation (Article 11), automatic logging (Article 12), transparency (Article 13), human oversight (Article 14), and appropriate accuracy, robustness and cybersecurity (Article 15), backed by a quality management system (Article 17), conformity assessment (Article 43) and post-market monitoring (Article 72) (European Parliament and Council 2024). Deployers carry complementary duties, including assigning oversight to people with the necessary competence, training, authority and support (Article 26).

This article uses the term *control safeguards* for the technical design features and organisational arrangements that keep human actors able to understand, monitor, override or interrupt an AI system, and that keep the system within acceptable performance and security limits. In the Act, the core of this bundle is Article 14, supported by Articles 9, 12, 13, 15 and

26. The social ambition is clear: no high-risk system should act as an unaccountable black box beyond human reach. Whether people affected by automated decisions are protected in practice depends on

whether these safeguards work, not only on whether they exist.

The path from that ambition to working systems is much less clear. Three developments make the question pressing. First, the legislative timetable has moved. Because standards, guidance and competent authorities were not ready, the Digital Omnibus on AI (Regulation (EU) 2026/1744) deferred the high-risk obligations for stand-alone (Annex III) systems to 2 December 2027, and for product-embedded (Annex I) systems to 2 August 2028 (European Parliament and Council 2026; European Commission 2025). Second, technical research has argued with growing force that the control assumed by regulators is often unattainable with current methods, particularly for opaque, autonomous and agentic systems (Manheim and Homewood 2025; Reuel et al. 2024). Third, early empirical work reports uncertainty among providers and deployers about what counts as an AI system, what counts as high-risk, and what evidence of compliance will be accepted (European Union Agency for Fundamental Rights [FRA] 2025).

These strands have mostly been studied apart. Legal scholars have examined the meaning and limits of human oversight (Enqvist 2023; Laux 2024), the conflation of trustworthiness with acceptable risk (Laux et al. 2024) and the dependence of the regime on standards (Veale and Zuiderveen Borgesius 2021; Ebers et al. 2021). Human-factors researchers have documented why human-in-the-loop designs fail (Bainbridge 1983; Parasuraman and Manzey 2010; Green 2022; Crootof et al. 2023). Machine-learning researchers have catalogued the weaknesses of alignment, filtering and monitoring (Casper et al. 2023; Wei et al. 2023; Hendrycks et al. 2021). What is missing is an integrated analysis that takes each element of the control bundle and asks which technical and which legal-institutional obstacles stand in its way, and how those obstacles amplify each other. Without it, regulators may keep tightening texts that engineers cannot satisfy, and engineers may keep producing documentation that regulators cannot evaluate.

The article addresses three research questions. **RQ1**: What are the principal technical obstacles to effective control safeguards in high-risk AI systems? **RQ2**: What features of the legal text, its standardisation machinery and its enforcement architecture obstruct their operationalisation? **RQ3**: How do these barriers interact, and what does that imply for regulatory design and engineering practice?

The contribution is threefold. It offers a taxonomy of eleven barriers grounded in both bodies of literature. It maps the oversight capabilities required by Article 14(4) to the barriers that most threaten each, which makes the abstract notion of an “implementation gap” concrete. And it derives recommendations aimed at specific actors rather than at “policymakers” in general. A secondary contribution is a discussion of what the EU experience implies for jurisdictions outside Europe, including African states developing AI governance under tighter capacity constraints.

Section 2 sets out the conceptual framework and methods. Section 3 reviews the literature. Sections 4 and 5 analyse the technical and the legal-institutional barriers. Section 6 examines their interplay. Section 7 discusses implications and limitations, Section 8 offers recommendations and Section 9 concludes.

2. Conceptual framework and methods

2.1 Key concepts

Three concepts structure the analysis. The first is the distinction between *oversight* and *control*. Manheim and Homewood (2025) argue that the two are routinely conflated: control means ex-ante or real-time operational constraints that prevent failures, whereas oversight means largely ex-post detection, remediation and governance. A system can be heavily overseen (logged, audited, reviewed) yet weakly controlled, because nothing stops a harmful action from completing before review. Article 14 uses the language of oversight, but its fourth paragraph describes capabilities, such as interrupting the system through a "stop" mechanism, that belong to control. This article follows the technical distinction because it shows where the legal wording asks for more than oversight practice can deliver.

The second concept is the *implementation gap*: the difference between the protective effect a requirement is meant to have and the effect it has once real organisations operationalise it with available technology, standards and institutions. The gap may be formal (the requirement is met on paper but not in practice) or substantive (it cannot be met at all in a given context). The distinction matters because the remedies differ: formal gaps call for better verification, substantive gaps for changes in design, scope or law.

The third is the *barrier*, defined as a condition that makes it materially harder, costlier or less reliable to implement a control safeguard as the law intends. Barriers are termed technical when they arise from properties of AI models, safeguard techniques or human–machine interaction, and legal-institutional when they arise from the content of legal rules, the standardisation machinery or the capacities of the bodies that apply and enforce them. The classification is analytic: as Section 6 shows, the most important barriers sit at the boundary of the two.

2.2 Research design

The study uses an interdisciplinary qualitative design that combines doctrinal legal analysis with a structured critical review of technical, socio-technical and policy literature. Its purpose is not to measure the performance of particular safeguards but to map the obstacles to their implementation and explain how they interact.

The doctrinal component is a close reading of Regulation (EU) 2024/1689 as amended by Regulation (EU) 2026/1744. It focuses on classification (Article 6 and Annexes I and III), the

requirements for high-risk systems (Articles 9–15), provider and deployer obligations (Articles 16 and 26), standardisation and the presumption of conformity (Articles 40 and 41), conformity assessment and substantial modification (Article 43), post-market monitoring and incident reporting (Articles 72 and 73), and the provisions on small and medium-sized enterprises (SMEs) and regulatory sandboxes. Recitals, the Commission's standardisation request to CEN and CENELEC (European Commission 2023), the Digital Omnibus proposal (European Commission 2025) and secondary legal literature were used to

interpret legislative intent and identify contested meaning. The legal position is stated as at 14 October 2026.

The literature component draws on peer-reviewed and preprint sources from computer science, human factors, law and public administration, retrieved from IEEE Xplore, the ACM Digital Library, arXiv, Scopus, Web of Science and SSRN, and from the websites of standardisation bodies and public authorities. Search strings combined terms for high-risk AI, human oversight, human control, technical safeguards, guardrails, automation bias, agentic systems, conformity assessment, harmonised standards and implementation barriers. Sources from 2015 to 2026 were considered, with priority to 2022–2026 because the high-risk regime is recent; older sources were retained where foundational. Sources were included if they bore directly on the technical feasibility of a control mechanism or on a legal, institutional or organisational obstacle to implementing the Act. Grey literature was included only where it contained substantive analysis or empirical findings.

The retained corpus was analysed thematically. Barriers were coded inductively, grouped under technical and legal-institutional headings, and cross-coded to the Article 14(4) capabilities and the Act's supporting provisions. Interactions were identified wherever a source described a barrier in one category as a cause, condition or amplifier of a barrier in the other. This procedure favours conceptual synthesis over exhaustive coverage and should not be read as a formal systematic review in the PRISMA sense.

2.3 Use of AI tools

A large language model (Claude, Anthropic) was used to assist in structuring and drafting parts of the text and in suggesting candidate literature. The author selected, read and checked the sources cited, verified the legal provisions against the official texts, revised all content and takes full responsibility for the final manuscript.

2.4 Limitations

The study relies on published materials and includes no primary data such as interviews with providers, deployers or market surveillance authorities. The EU regime is in transition: its high-risk obligations had not yet applied at the time of writing, so claims about enforcement are anticipatory and rest on early guidance, consultations and reports of preparation rather than case law. Standardisation work is moving quickly, and the status of individual standards may change after publication. Finally, the analysis concerns the EU AI Act; its conclusions travel to other jurisdictions only to the extent that they adopt comparable conceptual structures (Section 7.4).

3. Literature review

3.1 Legal analyses of the high-risk regime

Legal commentary has concentrated on the structure of the high-risk regime and its reliance on the New Legislative Framework, in which general essential requirements are given technical content by

harmonised standards. Veale and Zuiderveen Borgesius (2021) observed early that this model hands much of the substance of the regime to private standardisation bodies, whose processes are ill-suited to contested value judgements about fundamental rights. Ebers et al. (2021) raised related concerns about vague requirements and the self-assessment character of conformity assessment for most high-risk categories. Mökander et al. (2022) examined how conformity assessment and post-market monitoring could be given content through auditing, and noted that the available techniques are immature. Edwards (2022) criticised the Act for treating AI as a product certified at the point of sale although its risks emerge from use, adaptation and context.

Another body of work examines risk. Schuett (2024) and Fraser and Bello y Villarino (2024) analyse the risk management system of Article 9 and the notion of acceptable residual risk, and find little guidance on how much risk reduction is enough. Laux et al. (2024) argue that the Act conflates trustworthiness with the acceptability of risk, so that compliance may be mistaken for evidence of trustworthiness. Novelli et al. (2024) propose structured, scenario-based risk assessment to make classification and proportionality more predictable. Empirical findings from the FRA, based on interviews with providers, deployers and experts, show divergent interpretations of what counts as an AI system and as high-risk (FRA 2025).

3.2 Human oversight: law, human factors and socio-technical research

Koulu (2020) traced how European policy and AI ethics have treated human control as a safeguard against automation, and Enqvist (2023) analysed what, when and by whom oversight is to be exercised under the Act, concluding that the text leaves each question substantially open. Laux (2024) argues that Article 14 embodies institutionalised distrust, designed to counter the weaknesses of both machines and their supervisors, and that its legitimacy depends on how it is operationalised. Green (2022) argues, from evidence on how people use algorithmic advice, that policies requiring human oversight rest on two flawed assumptions: that people can perform the oversight function, and that oversight is a legitimate remedy; it may legitimise systems that should not be deployed. Crootof et al. (2023) develop a typology of the roles humans play in automated systems and caution against treating the “human in the loop” as a generic cure.

The empirical foundation of these critiques is human-factors research. Bainbridge (1983) described the “ironies of automation”: the more reliable the automation, the less practised the operator and the less able to take over when it fails. Parasuraman and Manzey (2010) and Goddard et al. (2012) synthesised evidence on complacency and automation bias, and Skitka et al. (1999) showed that commission and omission errors rise when decision support is available. Lee and See (2004) analysed how trust in automation is calibrated or miscalibrated by system feedback. In public administration, Alon-Barkat and Busuioc (2023) found selective adherence to algorithmic advice, with decision-makers following advice that matched their prior beliefs, and Busuioc (2021) examined the limits of explanation as a route to accountability. Elish (2019) described the “moral crumple zone”, in which humans formally in control

absorb responsibility for failures of systems whose behaviour they could not reasonably prevent. Sterz et al. (2024) identify conditions for effective oversight, including sufficient causal power, epistemic access and appropriate self-control, that current designs rarely guarantee.

3.3 Technical safeguards and their limits

Machine-learning safety research supplies the technical counterpart. Concrete problems such as reward misspecification, distributional shift and unsafe exploration were framed by Amodei et al. (2016) and extended in broader agendas (Hendrycks et al. 2021). Predictive uncertainty degrades when inputs depart from the training distribution (Ovadia et al. 2019), which bears directly on the performance and monitoring assumptions behind Article 15.

Adversarial examples (Goodfellow et al. 2015) and, for language models, automated adversarial suffixes (Zou et al. 2023) and systematic failures of safety training (Wei et al. 2023) show that input and output defences remain vulnerable. Casper et al. (2023) catalogue fundamental limitations of reinforcement learning from human feedback, the main alignment method for large models. Greshake et al. (2023) demonstrated indirect prompt injection against applications that combine language models with retrieval and tools, the architecture of many emerging agentic systems. Chan et al. (2023) and Kapoor et al. (2024) analyse the harms and evaluation difficulties of increasingly agentic systems, and Nolte et al. (2025) examine the legal and technical shortcomings of the Act's robustness and cybersecurity provisions.

On interpretability, Rudin (2019) argued that post-hoc explanation of black-box models is no substitute for inherently interpretable models in high-stakes settings. Adebayo et al. (2018) showed that some popular saliency methods fail basic sanity checks, Jain and Wallace (2019) questioned whether attention weights explain model behaviour, and Turpin et al. (2023) found that the stated reasoning of language models can be systematically unfaithful. Raji et al. (2020, 2022) show that accountability and functionality claims for deployed AI are often unsupported by adequate internal auditing.

3.4 Technical AI governance and the gap

The field of technical AI governance defines itself as the technical analysis and tools needed to support effective regulation. Reuel et al. (2024) provide a taxonomy of open problems in assessment, verification, security and operationalisation, and conclude that many regulatory aspirations outpace the technical means to implement and enforce them. International standards offer partial scaffolding, notably ISO/IEC 42001:2023 on AI management systems (ISO/IEC 2023b), ISO/IEC 23894:2023 on AI risk management (ISO/IEC 2023a) and the NIST AI Risk Management Framework (NIST 2023), but none is a harmonised standard conferring a presumption of conformity with the Act.

Taken together, these strands show that technical and legal barriers are individually well documented. The legal literature rarely engages with the technical feasibility of the capabilities it presumes, and the technical literature rarely maps its findings to the provisions that create the obligations. The present study addresses that disconnect.

4. Technical barriers

Technical barriers to effective control are both structural and dynamic. They arise from the properties of contemporary AI, the limits of available safeguard techniques, and the difficulty of embedding reliable human and technical controls in systems that operate at high speed, with high autonomy or at high complexity. Five barriers are distinguished (T1–T5). They do not merely delay compliance; they raise the question whether some forms of control that the law presupposes are attainable.

4.1 T1: Opacity of advanced models

Deep learning systems, and in particular large foundation models and their agentic extensions, generate outputs through high-dimensional computations that resist direct human interpretation. Article 14(4)(a) presupposes that overseers can "properly understand the relevant capacities and limitations" of the system and monitor its operation, and Article 14(4)(c) that they can correctly interpret its output. Opacity undermines both. When decision pathways cannot be inspected in real time, the operator evaluates surface results rather than underlying reasoning, which increases the risk of misplaced trust (Manheim and Homewood 2025).

Post-hoc explanation only partly mitigates the problem. Some saliency methods fail basic sanity checks (Adebayo et al. 2018), attention weights do not reliably show why a model produced an output (Jain and Wallace 2019), and the chain-of-thought explanations of language models can be unfaithful to the factors that influenced the answer (Turpin et al. 2023). Rudin (2019) concludes that in high-stakes settings the more defensible route is inherently interpretable models, a design choice the Act does not require and which may conflict with the performance claims of the most capable systems. Transparency under Article 13 is therefore necessary but not sufficient: documentation can describe a system's intended purpose and known limitations, but it cannot make the internal basis of an individual output legible.

4.2 T2: Speed, autonomy and the timing of intervention

Control in the technical sense requires either operational boundaries that can be enforced programmatically or a human able to intervene before irreversible effects occur (Manheim and Homewood 2025). In many high-risk contexts the time available is shorter than the time a person needs to perceive, evaluate and act, so the human can at best review outcomes after

the fact. Article 14(4)(e), which calls for the ability to intervene or interrupt the system "through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state", is feasible only where a halt is itself safe and can be exercised in time. For systems embedded in safety components of critical infrastructure, an abrupt halt may cause harm that continued operation would not.

Agentic systems aggravate the difficulty. Once an agent has tool access, memory and multi-step planning, actions that look individually permissible can combine into trajectories that violate system-level constraints before any review occurs (Chan et al. 2023; Reuel et al. 2024). Evaluating such systems

is itself hard: Kapoor et al. (2024) show that agent benchmarks often neglect cost, reproducibility and real-world conditions, which weakens the evidence for claims of reliable behaviour. The Act requires oversight measures “commensurate with the risks, level of autonomy and context of use” (Article 14(3)), but supplies no method for deciding how much control is achievable at a given level of autonomy.

4.3 T3: Brittle guardrails and monitoring

Existing safeguards reduce certain classes of harmful behaviour without eliminating them. Post-training alignment, filters and content classifiers remain vulnerable to adversarial inputs (Zou et al. 2023), to systematic weaknesses in how safety training generalises (Wei et al. 2023) and to the open problems of learning from human feedback (Casper et al. 2023). Guardrails implemented through prompts or static classifiers are particularly fragile when models operate with memory, retrieval and tools, because the effective decision context extends well beyond any single input; indirect prompt injection through retrieved content is a documented example (Greshake et al. 2023). Article 15(5) requires resilience against attempts by unauthorised third parties to alter a system’s use, outputs or performance by exploiting vulnerabilities, including data poisoning and adversarial examples, yet Nolte et al. (2025) conclude that the provisions leave substantial questions open about what resilience must be demonstrated.

Monitoring has its own limits. Performance can degrade silently under distributional shift, and uncertainty estimates are least reliable exactly when inputs diverge from training data (Ovadia et al. 2019). Monitoring tends to detect problems after they occur and struggles to supply the low-latency, high-reliability signals needed for real-time interruption. Defence-in-depth, which layers several imperfect controls, mitigates individual weaknesses,

but the residual risk of the layered system is hard to quantify and depends on context (Reuel et al. 2024). Article 9 requires that residual risk be judged acceptable; the difficulty of estimating it makes that judgement hard to ground.

4.4 T4: Human-factors limits on oversight

Even where a system could in principle be overseen, the people asked to do so may be unable to. Automation bias, anchoring and complacency are well documented (Skitka et al. 1999; Parasuraman and Manzey 2010; Goddard et al. 2012). Article 14(4)(b) requires that overseers “remain aware of the possible tendency of automatically relying or over-relying” on output, but awareness alone has not been shown to eliminate the effect. Bainbridge’s (1983) ironies persist: highly reliable systems erode operator skill and vigilance, so the human is least prepared when intervention matters most. Alon-Barkat and Busuioc (2023) show that decision-makers selectively follow algorithmic advice that fits their prior beliefs, so oversight may reinforce bias rather than correct it.

Interface and workflow design add constraints. Excessive or poorly prioritised information overwhelms rather than empowers the overseer, turning the human into a formal approval step. Where a person is nominally in control but practically unable to prevent failure, the result is Elish’s (2019) moral crumple

zone: responsibility attaches to the human while the system's behaviour was beyond their influence. Sterz et al. (2024) argue that effective oversight requires causal power, epistemic access and appropriate self-control; these are properties of the whole socio-technical arrangement, not of a stop button or a screen. Green (2022) concludes that evidence does not support the assumption that people reliably perform the oversight function envisaged by policy.

4.5 T5: Absence of testable metrics for control effectiveness

The final technical barrier is the lack of accepted methods for demonstrating that a control safeguard works. For accuracy and robustness, benchmarks and test protocols exist, though their relevance to real conditions is contested (Raji et al. 2022). For human oversight, there are few validated measures of whether overseers in fact understand, detect, interpret and override as Article 14(4) requires. For continuously learning systems, Article 15(4) requires that feedback loops be addressed, and Article 43(4) treats changes pre-determined at the time of conformity assessment as not constituting substantial modification, but the means of bounding and verifying permitted change are not settled. The Commission's standardisation

request to CEN and CENELEC (European Commission 2023) responds to this gap and work continues in the joint technical committee JTC 21 (CEN-CENELEC n.d.), but the same technical uncertainty that makes controls hard to specify in testable form hampers that work.

4.6 Synthesis

These barriers show that control safeguards cannot be treated as engineering add-ons. In many settings they confront hard limits arising from opacity, autonomy, speed and the immaturity of safeguard technology. Addressing them requires incremental technical progress and an explicit recognition, in engineering practice and in regulatory expectation alike, of the conditions under which meaningful control is, and is not, attainable.

5. Legal and institutional barriers

Legal and institutional barriers arise from the design of the regulatory framework, the interpretive problems it generates, and the administrative and organisational conditions under which it must be applied. The AI Act makes a clear normative demand for effective oversight, but its translation into practice is impeded by six interlocking difficulties (L1–L6).

5.1 L1: Open-textured statutory concepts

Key provisions use terms that call for judgement: "effective" oversight (Article 14(1)), "appropriate" levels of accuracy, robustness and cybersecurity (Article 15(1)), "state of the art" (Article 9), "reasonably foreseeable misuse" (Article 9(2)(b)), and oversight "commensurate with the risks, level of autonomy and context of use" (Article 14(3)). Such standards are a deliberate legislative technique that keeps the law

technologically neutral, but they presume that guidance, standards and case law will progressively fix their meaning. In the interim, providers and deployers must decide for themselves what level of control is sufficient and how to evidence it (Enqvist 2023; Fraser and Bello y Villarino 2024; Schuett 2024).

The difficulty is sharpest for Article 14, where the requirement that natural persons be enabled to understand, monitor, interpret, override and interrupt the system must be realised through design choices whose adequacy is not specified. Article 14(3) also divides the work between providers, who build in oversight measures “when technically feasible”, and deployers, to whom other measures are left. This creates a coordination problem: providers can satisfy their duty by designing the interface and documenting the measures, yet whether

oversight is effective depends on how deployers staff, train and empower the people using it (Article 26(2)). Each party can plausibly attribute shortfalls to the other (Enqvist 2023; Laux 2024).

5.2 L2: Classification uncertainty

The full set of safeguards applies only to high-risk systems. Article 6 and Annex III define the category, and Article 6(3) lets providers conclude that an Annex III system is not high-risk where it poses no significant risk of harm, subject to documenting the assessment (Article 6(4)). In practice, borderline and novel systems make classification contestable, and the FRA (2025) reports that uncertainty about the definition of an AI system and of high risk is already visible among market actors. Errors run in both directions. Over-inclusion imposes costs where safeguards add little, while under-inclusion, encouraged by the self-assessment character of the Article 6(3) derogation, leaves risky systems outside the control regime. The Commission was due to publish guidelines on the practical implementation of Article 6 by 2 February 2026, and the lack of timely guidance was among the reasons for deferring the high-risk obligations (European Commission 2025). Novelli et al. (2024) suggest scenario-based assessment as a remedy, but no such method has been adopted as a reference.

5.3 L3: Delayed standards and regulatory instability

The Act follows the New Legislative Framework: essential requirements are given technical expression through harmonised standards, compliance with which creates a presumption of conformity (Article 40), with common specifications as a fallback (Article 41). The Commission’s request to CEN and CENELEC (European Commission 2023) assigned the work to JTC 21, but delivery has run well behind the legislative timetable. Until standards are adopted and cited in the Official Journal, providers lack an authoritative route to demonstrate conformity, and many rely on international instruments such as ISO/IEC 42001 and 23894, which are valuable but were not designed to confer a presumption of conformity with the Act.

The gap is especially significant for control-related requirements, where engineering practice is evolving and consensus is incomplete. The Digital Omnibus acknowledges the problem in practice by moving the application dates to December 2027 and August 2028 (European Parliament and Council 2026). This gives organisations time, but it also introduces regulatory instability: providers that invested early must

adjust their plans, while those awaiting clarity have an incentive to defer. A moving timetable is a barrier in its own right, because investment in control safeguards is rational only if the target is stable enough to design against.

5.4 L4: Limited institutional capacity

Enforcement depends on market surveillance authorities and, for certain systems, notified bodies. Evaluating complex AI systems demands specialised expertise, dedicated audit capacity and methods suited to dynamic models. The delay in designating and equipping national competent authorities was one of the stated reasons for the Omnibus (European Commission 2025). Authorities with limited technical capacity cannot supply the authoritative interpretation that open-textured provisions require, and cannot readily distinguish substantive from merely documentary compliance. Mökander et al. (2022) stress that conformity assessment and post-market monitoring will be only as meaningful as the auditing capacities that stand behind them.

5.5 L5: Fragmented interpretation and enforcement

Although the AI Act is directly applicable, enforcement is decentralised, and national authorities may interpret it differently. Member States may reach different views about what constitutes effective oversight or an adequate risk assessment for the same system, producing uneven burdens and opportunities for forum shopping. Beyond the EU, providers serving several markets must reconcile the Act with other regimes and standards, including the NIST framework (NIST 2023) and the Council of Europe's Framework Convention on AI (Council of Europe 2024), whose vocabularies overlap but are not identical. Fragmentation raises the cost of developing safeguards that can be demonstrated across jurisdictions and weakens the signal regulators send about what works.

5.6 L6: Resource barriers and smaller organisations

The cumulative demands of a risk management system, documentation, oversight arrangements, conformity assessment, logging and post-market monitoring impose substantial administrative and financial burdens that fall disproportionately on SMEs and start-ups. The Act contains mitigating provisions, including simplified technical documentation (Article 11(1)), simplified quality management for microenterprises (Article 63), priority access to sandboxes (Article 62) and proportionate penalties (Article 99), and the Omnibus extends further simplifications for smaller enterprises (European Parliament and Council 2026). Nevertheless, relief from documentation does not supply expertise. An organisation without specialists in robustness testing, human factors and security will struggle to implement safeguards that are both effective and demonstrably compliant. The deployer side is affected too: public bodies and small firms must assign competent, trained and empowered oversight personnel (Article 26(2)), which many may be ill-equipped to do. Smaller actors are therefore more likely to resort to checklist compliance, a main route to formal-only safeguards.

5.7 Synthesis

The legal-institutional barriers share a feature: each defers the specification of control to a later, external source (guidance, standards, authorities, case law) that has not yet materialised in usable form. The result is an environment in which formal adherence is attainable but genuine risk reduction through effective control remains uncertain.

6. Interplay between technical and legal barriers

The two sets of barriers do not operate in parallel. They interact in ways that make formal compliance easier to achieve than substantive control. Three mechanisms explain most of the interaction: a feasibility gap, a verification gap and a capacity gap.

6.1 The feasibility gap: legal demand outruns technical capability

Standards such as “effective” oversight and “appropriate” measures presuppose that in a given context it is possible to design a system whose overseers can meaningfully understand, monitor and intervene. The technical literature indicates that opacity, speed and autonomy often make such intervention difficult or impossible in realistic conditions (Manheim and Homewood 2025; Green 2022). The law is thus calibrated to an ideal of controllability that the state of the art supports only in some settings. Because the text does not say what a provider should do when the ideal cannot be reached, providers can meet documentation and process requirements (instructions for use, logging, assignment of oversight personnel) while the underlying control stays weak. Critics note that human oversight may then legitimise rather than limit the use of systems that cannot in fact be controlled (Green 2022; Crootof et al. 2023). The qualifier “when technically feasible” (Article 14(3)) acknowledges the feasibility problem, but it is framed as an exception applied by the provider’s own judgement, not as a trigger for reconsidering whether the system should be deployed in that context.

6.2 The verification gap: standards cannot be written for what cannot be tested

The Act assumes that technical consensus can be reached on what adequate control looks like, so that standards can supply testable criteria. The same uncertainties that make safeguards hard to engineer make them hard to standardise. Without validated measures of oversight effectiveness (T5), standards bodies can specify processes and documentation more readily than outcomes, and a presumption of conformity may then attach to adherence to a process rather than to demonstrated control. The effect is circular: incomplete standards perpetuate interpretive ambiguity (L1, L3), and unresolved technical limits slow the standards that would resolve it (Reuel et al. 2024; Veale and Zuiderveen Borgesius 2021). The problem is sharpest for dynamic and agentic systems, whose behaviour may change after initial conformity assessment, so that a snapshot certificate says little about what is later deployed (Article 43(4); Mökander et al. 2022).

6.3 The capacity gap: unequal ability to bridge uncertainty

Institutional and organisational factors mediate between the two domains. Authorities with limited technical capacity cannot interpret open-textured requirements in light of engineering constraints and so cannot distinguish substantive from superficial compliance (L4). Large, well-resourced providers can build sophisticated internal control architectures, commission red-teaming and run their own human-factors studies, and are positioned to influence standards. Smaller providers and deployers rely on checklists and vendor assurances (L6). The result is a stratified landscape in which the quality of control depends on the resources of the actor, and in which the regulator is least able to detect weak control where it is most likely to occur. Fragmented enforcement (L5) multiplies the effect by weakening any common benchmark.

6.4 Mapping the Article 14(4) capabilities to the barriers

Table 1 applies the framework to the five capabilities that Article 14(4) requires a high-risk system to enable. Each capability is threatened by at least one technical and one legal-institutional barrier, and the strongest gaps arise where the two coincide.

Table 1
Article 14(4) capabilities mapped to the technical (T1–T5) and legal-institutional (L1–L6) barriers identified in Sections 4 and 5. Source: author's analysis.

Article14(4) capability	**Principaltechnical barriers**	**Principal legal-institutional barriers**	**Resulting implementation gap**
(a)Understand capacitiesand limitations; monitor operation, including anomalies	T1 opacity; T3 monitoring that detects problems late	L1 no definition of sufficient understanding; L3 no testable criteria	Documented limitations stand in for realunderstanding; monitoringis procedural
(b) Remain aware of automation bias	T4 bias persists despite awareness; deskilling	L1 and L6 training duty (Articles 4 and 26(2)) left unspecified; limiteddeployer capacity	Awareness training is deliveredand recorded; behaviour is unchanged
(c) Correctly interpret the output	T1explanations unfaithful or unstable; T4information overload	L1"appropriate" interpretability undefined;L3no standardfor output explanation	Interpretation is an assertion by the provider, untested with real users
(d) Decide not to use, disregard, override or reverse	T2 speed limits time to decide; T4 selective adherence	L1 provider–deployer split; L4 authorities cannot test override in practice	Override exists as a function but is not exercised, or only after the fact
(e) Intervene or interrupt("stop" mechanism)	T2 safe halt may be unattainable; agentic trajectories outrun review	L1 "safe state" undefined;L2 classification affects whether required	A stop control is present but may not prevent harm in time

Two patterns emerge. First, capabilities (a) to (c) concern epistemic access, and the principal obstacle is the combination of opacity and unfaithful explanation with the absence of a standard for sufficient understanding. Second, capabilities (d) and (e) concern causal power, and the principal obstacle is the combination of speed and autonomy with the fact that nothing in the law ties the stop and override functions to the time scale on which harm occurs. These correspond closely to the epistemic and causal conditions proposed by Sterz et al. (2024), and suggest that regulatory specification should target them directly.

6.5 The resulting implementation gap

The interaction of the three gaps produces a distinctive implementation gap. Because legal openness and technical limitation reinforce one another, neither legal compliance nor technical excellence is sufficient alone. Effective control becomes contingent on a continuing negotiation among what the law requires, what technology can deliver and what organisations can sustain. This is not simply a

transitional difficulty to be resolved by time; some elements, such as the limits of human supervision over fast agentic systems, reflect enduring properties of the technology. Progress therefore cannot come from tighter drafting alone or from incremental engineering alone. It requires mechanisms that let regulatory concepts and engineering realities inform one another iteratively, the subject of Section 8.

7. Discussion

7.1 From formal compliance to substantive control

The analysis supports a central claim: under present conditions, the Act's control safeguards are more likely to be met formally than substantively. This is a structural prediction, not an accusation of bad faith. When the legal standard is open-textured, the technical means are immature and the verifying institutions are stretched, rational actors will concentrate on what can be shown (policies, instructions for use, logs, training records and conformity documentation) rather than what is hard to show, namely that overseers in practice understand, detect and override. The risk is that documentary compliance is read as evidence of safety, an inference against which Laux et al. (2024) warn. The consequences fall unevenly: operators carry formal responsibility without effective power (Elish 2019), and affected persons are protected less than the existence of safeguards implies.

7.2 The deferral as a window of opportunity

The postponement of the high-risk obligations (European Parliament and Council 2026) is often presented as a concession to industry. It can also be read as a window in which the preconditions for effective control can be built: standards with testable criteria, guidance on classification and oversight, and authority capacity. Whether the window is used well depends on whether the time is spent on the problems identified here or merely on drafting more detailed documentation templates.

7.3 When control is not attainable

For some combinations of system and context, effective control may not be technically attainable at the required level. The Act addresses this only indirectly, through the proportionality language of Article 14(3) and the "when technically feasible" qualifier. A more candid approach would recognise that where meaningful control cannot be achieved, the correct response is to restrict autonomy, narrow the context of use, require additional safeguards outside the system, or refrain from deployment, rather than to certify the system with weaker oversight. This is consistent with the oversight–control distinction (Manheim

and Homewood 2025) and with Green's (2022) caution that human oversight can legitimise deployments that should not proceed.

7.4 Implications beyond the European Union

The findings matter outside Europe for two reasons. First, the AI Act is likely to exert influence beyond the EU, through market access and through the adoption of its vocabulary and risk categories by other

jurisdictions, a dynamic described as the Brussels effect (Bradford 2020; Siegmann and Anderljung 2022). Second, several regions, including the African Union through its Continental Artificial Intelligence Strategy (African Union 2024), are developing AI governance in contexts where regulatory and technical capacity is more limited than in the EU. The barriers identified here are, if anything, more acute in such settings. Authorities may lack technical audit capacity (L4), local developers and public deployers are typically small (L6), harmonised standards will be written elsewhere (L3), and many high-risk systems will be imported rather than built locally, so that control depends on what foreign providers have chosen to build in.

Three cautions follow for jurisdictions considering EU-style frameworks. A framework that copies the formal structure of the EU regime without its institutional supports risks producing documentary compliance with little protection. Capacity should be built in step with obligations: for example, by focusing initially on a narrow set of high-consequence uses, by relying on procurement conditions and deployer-side requirements where public bodies are the principal deployers, and by pooling technical expertise at regional level. And international standards such as ISO/IEC 42001 and 23894 and the NIST framework can serve as practical scaffolding, provided they are tested against the question this article poses: whether the resulting controls work in the actual operating context. Where systems are imported, provider documentation requirements (Articles 11 and 13) and contractual rights to information about oversight design become important levers for deployers.

7.5 Future research

The analysis is conceptual, and the empirical weight of each barrier is likely to vary by sector and system type. Future work should test the Article 14(4) mapping empirically through user studies of overseers in realistic conditions, examine early conformity-assessment practice once the obligations apply, compare how Member State authorities interpret open-textured

terms, and study the experience of deployers and providers in lower-capacity jurisdictions that adopt EU-inspired frameworks.

8. Recommendations

The recommendations follow from the three gaps in Section 6 and are addressed to specific actors. They are intended to reinforce one another: guidance (R1, R2) gives standards bodies targets; standards (R3, R4) give authorities and providers criteria; and capacity measures (R5–R7) allow those criteria to be applied beyond the largest organisations.

1. **R1. Specify control requirements by autonomy level and context (European Commission, AI Office).** Guidelines under Article 14 should translate “effective” oversight into graded, context-specific expectations. They should distinguish oversight (ex-post review) from control (ex-ante or real-time constraint), indicate for typical Annex III use cases which the system must support, and state what must happen when the time available for intervention is shorter

than a human needs to act: for example, programmatic limits on action scope, rate and reversibility, or human authorisation before irreversible steps (Manheim and Homewood 2025). They should also define a "safe state" for the Article 14(4)(e) stop mechanism and clarify the division of responsibilities between providers and deployers under Articles 14(3) and 26.

2. **R2. Provide authoritative and timely classification guidance (European Commission, national authorities).** Guidelines on Article 6 should include worked examples of borderline systems, criteria for the Article 6(3) derogation and a procedure for obtaining an authoritative view from an authority. Scenario-based assessment methods of the kind proposed by Novelli et al. (2024) could be piloted. Clarity at the classification stage determines whether control safeguards apply at all.
3. **R3. Develop and standardise control-effectiveness testing (CEN-CENELEC JTC 21, ISO/IEC JTC 1/SC 42, research funders).** Standards should move beyond process requirements to outcome-oriented tests: red-team protocols for guardrails against adversarial and multi-turn attacks, latency tests for interruption, human-factors protocols that measure whether trained overseers detect and override faulty outputs under realistic workload, and monitoring requirements tied to a system's exposure to distribution shift. Research funding should prioritise validated measures of oversight effectiveness (Sterz et al. 2024) and benchmarks for agentic systems that reflect real operating conditions (Kapoor et al. 2024).
4. **R4. Adopt dynamic standardisation with scheduled review (standardisation bodies, European Commission).** Because the state of the art moves quickly, standards and common specifications for control should be versioned, with fixed review cycles and annexes that can be updated without reopening the whole standard. Where harmonised standards are unavailable, the Commission should use its power to adopt common specifications (Article 41) rather than leave providers without a route to demonstrate conformity, and should publish its timetable so that investment targets are stable.
5. **R5. Build technical capacity in authorities and notified bodies (Member States, AI Office).** Authorities need technical staff, shared testing infrastructure and access to external expertise. A pooled European technical resource could provide evaluation support and common test environments, reduce divergence in enforcement (L5) and make it feasible to examine whether controls work as well as whether documents exist (Mökander et al. 2022). Peer review and joint guidance among authorities would help align interpretations of open-textured terms.
6. **R6. Lower the practical cost of effective control for small organisations (European Commission, Member States, industry bodies).** Beyond documentation relief, small providers and deployers need reusable assets: open reference implementations of oversight interfaces and logging, template risk assessments tied to common use cases, subsidised access to testing tools, and priority access to sandboxes (Article 62). Sandboxes should generate evidence about what works and feed it back into guidance and standards, rather than serve only as exemptions.

7. **R7. Shift deployer obligations from documentation to demonstrated competence (providers, deployers, authorities).** Providers should supply oversight-specific training materials and, where feasible, scenario-based exercises. Deployers should show, for example through periodic exercises and incident reviews, that designated personnel have the competence, authority and time envisaged by Article 26(2). Organisational design should avoid placing individuals in a moral crumple zone by ensuring that those held responsible have the means to intervene (Elish 2019).

8. **R8. Make the limits of control explicit and act on them (providers, deployers, legislators).** Provider documentation should state candidly the conditions under which control measures lose effectiveness, such as input types, speeds and autonomy levels. Risk management should treat non-attainable control as a reason to restrict autonomy, narrow deployment or decline to deploy. At the next review of the Act, legislators should consider whether specific high-consequence categories require stronger ex-ante constraints than oversight-based safeguards can provide.
9. **R9. Strengthen the evidence base through mandatory learning (authorities, AI Office, researchers).** Serious incident reporting (Article 73) and post-market monitoring (Article 72) should capture control failures, including near misses involving ineffective oversight, and make anonymised information available to researchers
and standards bodies. This closes the loop between practice, standards and guidance.
10. **R10. Align internationally and support lower-capacity jurisdictions (EU institutions, international organisations, regional bodies).** Mappings among the EU regime, ISO/IEC standards and the NIST framework should identify what evidence satisfies several regimes at once. Technical assistance, shared evaluation resources and regional cooperation, including within the African Union framework (African Union 2024), would allow jurisdictions with limited capacity to adopt proportionate controls without importing documentary requirements they cannot verify.

The recommendations differ in cost and timing. R1, R2 and R9 can be acted on within the deferral window using existing instruments. R3, R4 and R5 require sustained investment and cooperation among regulators, standardisation bodies and researchers. R8 and R10 involve political choices about the scope of the regime and international responsibility, which this article can inform but not settle.

9. Conclusion

This article asked why the control safeguards required for high-risk AI systems under the EU AI Act are difficult to implement, and how the obstacles interact. It identified five technical barriers (opacity, speed and autonomy, brittle guardrails and monitoring, human-factors limits, and the absence of testable control metrics) and six legal-institutional barriers (open-textured concepts, classification uncertainty, delayed standards with an unstable

timetable, limited authority capacity, fragmented enforcement, and resource constraints on smaller organisations). Its central finding is that these barriers form a mutually reinforcing system. A feasibility

gap arises where the law demands a degree of controllability that technology supports only in some settings. A verification gap arises where standards cannot specify what cannot yet be tested. A capacity gap arises where regulators and smaller organisations cannot bridge the resulting uncertainty as large providers can. Mapping Article 14(4) to these barriers showed that the most acute gaps occur where epistemic access and causal power are both required, and where nothing in the law ties the stop and override functions to the time scale on which harm occurs.

Without closer alignment between regulatory expectations and engineering realities, control safeguards risk remaining formal rather than protective: present on paper and in system interfaces, but not shown to work when it matters. The deferral of the high-risk obligations offers time to avoid this outcome. The recommendations proposed here are designed to use that time to narrow each of the three gaps.

Two broader lessons follow. For regulators, specifying what must be achieved matters more than extending what must be documented, and a requirement that cannot be tested cannot be enforced. For engineers and deployers, control is a property of the entire socio-technical arrangement and not of a single feature, so a stop button, a log or a trained operator is meaningful only if it can be shown to work in the real operating context. For jurisdictions outside Europe that are building AI governance frameworks, the experience suggests starting with a narrow set of high-consequence uses and building verification capacity alongside obligations, so that frameworks protect people in practice and not only in text. The most urgent empirical question for future research is whether human overseers can in fact perform the functions that Article 14(4) assigns them under realistic conditions.

Declarations

Data availability

No primary datasets were generated or analysed during this study. The research consists of doctrinal legal analysis of publicly available legislative texts and a structured critical review of published academic, technical and policy literature.

The principal legal sources analysed are:

- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), and

Regulation (EU) 2026/1744 amending that Regulation (Digital Omnibus on AI). These texts are publicly available from EUR-Lex (https://eur-lex.europa.eu).

- Related official documents of the European Commission (including the standardisation request to CEN and CENELEC and the Digital Omnibus proposal) that are cited in the reference list and available from the official EUR-Lex or European Commission websites.

All secondary sources (peer-reviewed articles, preprints, international standards, reports of public authorities and policy documents) are fully cited in the References section of the manuscript. They are publicly accessible via their DOIs, arXiv identifiers, or the websites of the respective organisations (ISO, NIST, CEN-CENELEC, European Union Agency for Fundamental Rights, African Union, etc.).

No confidential, proprietary, restricted or non-public materials were used. There are therefore no obstacles to sharing the materials underlying the analysis; all of them are already in the public domain.

Competing interests

The author declares no competing interests.

Ethical approval

Ethical approval was not required as no human participants were involved in this research.

Author contributions

F.L. conceived the study, carried out the analysis, wrote the manuscript and approved the final version.

References

1. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: Adv Neural Inf Process Syst 31
2. African Union (2024) Continental artificial intelligence strategy: harnessing AI for Africa's development and prosperity. African Union, Addis Ababa
3. Alon-Barkat S, Busuioc M (2023) Human–AI interactions in public sector decision making: automation bias and selective adherence to algorithmic advice. J Public Adm Res Theor 33(1):153–169
4. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv:1606.06565
5. Bainbridge L (1983) Ironies of automation. Automatica 19(6):775–779
6. Bradford A (2020) The Brussels effect: how the European Union rules the world. Oxford University Press, Oxford
7. Busuioc M (2021) Accountable artificial intelligence: holding algorithms to account. Public Adm Rev 81(5):825–836
8. Casper S, Davies X, Shi C et al (2023) Open problems and fundamental limitations of reinforcement learning from human feedback. Transactions on Machine Learning Research
9. CEN-CENELEC (n.d) Joint Technical Committee 21 Artificial Intelligence. European Committee for Standardization and European Committee for Electrotechnical Standardization. https://www.cencenelec.eu

10. Chan A, Salganik R, Markelius A et al (2023) Harms from increasingly agentic algorithmic systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), pp 651–666
11. Council of Europe (2024) Framework convention on artificial intelligence and human rights, democracy and the rule of law (CETS No. 225). Council of Europe, Strasbourg
12. Crootof R, Kaminski ME, Price WNII (2023) Humans in the loop. Vanderbilt Law Rev 76(2):429–510
13. Ebers M, Hoch VRS, Rosenkranz F, Ruschemeier H, Steinrötter B (2021) The European Commission's proposal for an Artificial Intelligence Act: a critical assessment by members of the Robotics and AI Law Society (RAILS). J 4(4):589–603
14. Edwards L (2022) Regulating AI in Europe: four problems and four solutions. Ada Lovelace Institute, London
15. Elish MC (2019) Moral crumple zones: cautionary tales in human–robot interaction
16. Engaging Science Technol Soc 5:40–60
17. Enqvist L, Law (2023) Innov Technol 15(2):508–535
18. European Commission (2023) Commission implementing decision C(2023) 3215 final of 22 May 2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence. European Commission, Brussels
19. European Commission (2025) Proposal for a regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). European Commission, Brussels. 19 November 2025
20. European Parliament and Council (2024) Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official J Eur Union L 2024/1689, 12 July 2024
21. European Parliament and Council (2026) Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Official J Eur Union, 24 July 2026
22. European Union Agency for Fundamental Rights (2025) Assessing high-risk artificial intelligence: fundamental rights risks. FRA, Vienna
23. Fraser H, Bello y Villarino J-M (2024) Acceptable risks in Europe's proposed AI Act: reasonableness and other principles for deciding how much risk management is enough. Eur J Risk Regul 15(2):431–447
24. Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127
25. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples
26. In: International Conference on Learning Representations (2015)

27. Green B (2022) The flaws of policies requiring human oversight of government algorithms
28. Computer, Law & Secur Rev 45:105681
29. Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M (2023) Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23)
30. Hendrycks D, Carlini N, Schulman J, Steinhardt J (2021) Unsolved problems in ML safety. arXiv:2109.13916
31. ISO/IEC (2023a) ISO/IEC 23894:2023 Information technology – Artificial intelligence – Guidance on risk management. International Organization for Standardization, Geneva
32. ISO/IEC (2023b) ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system. International Organization for Standardization, Geneva
33. Jain S, Wallace BC (2019) Attention is not explanation. In: Proceedings of NAACL-HLT 2019, pp 3543–3556
34. Kapoor S, Stroebl B, Siegel ZS, Nadgir N, Narayanan A (2024) AI agents that matter. arXiv:2407.01502
35. Koulu R (2020) Human control over automation: EU policy and AI ethics. Eur J Legal Stud 12(1)
36. Laux J (2024) Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. AI & Society 39
37. Laux J, Wachter S, Mittelstadt B (2024) Trustworthy artificial intelligence and the European Union AI Act: on the conflation of trustworthiness and acceptability of risk. Regul Gov 18(1):3–32
38. Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80
39. Manheim D, Homewood A (2025) Limits of safe AI deployment: differentiating oversight and control. arXiv:2507.03525
40. Mökander J, Axente M, Casolari F, Floridi L (2022) Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. Mind Mach 32(2):241–268
41. National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. U.S. Department of Commerce, Gaithersburg
42. Nolte H, Rateike M, Finck M (2025) Robustness and cybersecurity in the EU Artificial Intelligence Act. arXiv:2502.16184
43. Novelli C, Casolari F, Rotolo A, Taddeo M, Floridi L (2024) AI risk assessment: a scenario-based, proportional methodology for the AI Act. Digit Soc 3:13
44. Ovadia Y, Fertig E, Ren J et al (2019) Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Adv Neural Inf Process Syst 32
45. Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52(3):381–410

46. Raji ID, Smart A, White RN et al (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20), pp 33–44
47. Raji ID, Kumar IE, Horowitz A, Selbst A (2022) The fallacy of AI functionality. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22), pp 959–972
48. Reuel A, Bucknall B, Casper S et al (2024) Open problems in technical AI governance. arXiv:2407.14981
49. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
50. Schuett J (2024) Risk management in the Artificial Intelligence Act. Eur J Risk Regul 15(2):367–385
51. Siegmann C, Anderljung M (2022) The Brussels effect and artificial intelligence: how EU regulation will impact the global AI market. Centre for the Governance of AI, Oxford
52. Skitka LJ, Mosier KL, Burdick M (1999) Does automation bias decision-making?
53. International J Hum Comput Stud 51(5):991–1006
54. Sterz S, Baum K, Biewer S et al (2024) On the quest for effectiveness in human oversight: interdisciplinary perspectives. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24), pp 2495–2507
55. Turpin M, Michael J, Perez E, Bowman SR (2023) Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In: Adv Neural Inf Process Syst 36
56. Veale M, Zuiderveen Borgesius F (2021) Demystifying the draft EU Artificial Intelligence Act. Comput Law Rev Int 22(4):97–112
57. Wei A, Haghtalab N, Steinhardt J (2023) Jailbroken: how does LLM safety training fail? In: Advances in Neural Information Processing Systems 36
58. Zou A, Wang Z, Carlini N, Nasr M, Kolter JZ, Fredrikson M (2023) Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043