**Governed Opacity: The EU AI Act's Unresolved Tension Between Prohibiting and Institutionalising Information Asymmetry**

Word Count: 3,874

The European Union's Artificial Intelligence Act (AIA) opens with a sweeping aspiration: to be applied "in accordance with the values of the Union enshrined in the Charter, facilitating the protection of natural persons, undertakings, democracy, the rule of law and environmental protection, while boosting innovation and employment and making the Union a leader in the uptake of trustworthy AI."[1] The ambition is to balance fundamental rights against innovation, and the Act's signature instrument for achieving this is a risk-based framework that grades obligations to the severity of potential harm.

Yet the Act contains a structural puzzle. Article 5 categorically prohibits certain AI practices—social scoring, subliminal manipulation, exploitation of vulnerabilities—treating them as wrong in and of themselves. Articles 57–63, by contrast, establish regulatory sandboxes and real-world testing regimes that explicitly permit experimentation with systems whose risks are not yet fully understood.[2] These sandboxes institutionalise supervised opacity for a wide range of systems, suggesting that permissibility turns not on the nature of a practice but on the conditions under which it is conducted.

The coexistence of these two logics is not merely an administrative detail. Article 5 forecloses assessment before it can occur; the sandbox framework assumes that deployment is necessary to generate the knowledge assessment requires. The Act makes this tension explicit, requiring that real-world testing proceed "without prejudice" to Article 5's prohibitions, and

[1]European Parliament, *Regulation (EU) 2024/1689*, Recital 2, p. 2.
[2]European Parliament, *Regulation (EU) 2024/1689*, pp. 88–91.

Articles 62 and 63 further calibrate who benefits from informational asymmetry through size-based exemptions. But the boundary between what is categorically prohibited and what is eligible for supervised experimentation is never defended—only enacted. This raises an important question: does the AI Act draw a coherent line between harmful and productive information asymmetry, or does it expose an unresolved tension between eliminating asymmetry where it threatens individual agency and preserving it where it drives innovation?

This essay will argue that the Act exposes an unresolved tension rather than drawing a principled line. A hybrid regime combining categorical prohibitions with graduated supervision is not inherently incoherent—many regulatory frameworks do exactly that. But it can only be principled if the framework owns the choice and exposes its criteria to challenge. The AIA does not. It treats certain practices as inherently unacceptable, while structurally similar opacity is permitted—even cultivated—through supervised sandboxes and exemptions. The question is no longer purely academic: as the European Parliamentary Research Service reported in March 2026, August 2025 data showed that only one Member State sandbox was operational, sixteen Member States had yet to communicate their implementation plans, and the 2 August 2026 deadline was then less than five months away.[3] The boundary between harmful and productive asymmetry is therefore not conceptual but procedural and distributive. The question is no longer what kinds of asymmetry are permissible, but who may retain opacity, under what controls, and at what cost to individual agency.

## The Firm-Model Architecture of Generative AI

Before evaluating the Act's specific provisions, it is worth establishing what it governs. Drawing on Mayer-Schönberger and Ramge's distinction, generative AI companies operate far more like

[3] Marcelin 2026, p. 2.

centralised firms than open markets. In a marketplace, individuals and firms "gather and provide information and make decisions for themselves," producing a system that is "flexible and dynamic" precisely because informational power is diffused throughout.[4] The firm, by contrast, operates through "centralised coordination," where "information stream[s] to its centre."[5] GenAI development maps almost entirely onto this second architecture. Training data, model weights, fine-tuning methodologies, and alignment techniques are retained rather than shared because their retention is the source of competitive advantage. This is acknowledged even in the law-and-economics literature: Antonella Zarra identifies what she calls the "AI explainability gap"—the structural inability of regulators and end users to "inspect or understand model weights, critical parameters, training data provenance or emergent behaviours."[6] The AIA is therefore not intervening in a sector where asymmetry is incidental; it is engaging with an industry structured around the principle that information must flow inward and opacity is load-bearing.

This distinction matters because it changes what regulatory transparency requirements can accomplish. In conventional markets, disclosure requirements correct informational failures and restore the conditions under which a market functions. In GenAI, however, disclosing training data or model weights does not restore market function—it dismantles the informational model of the firm entirely. Opacity is not a removable layer of value creation but its core mechanism.[7] Regulating that opacity is therefore not analogous to regulating a polluting externality; it is intervening in the constitutive logic of the industry itself.

---

[4]Viktor Mayer-Schönberger and Thomas Ramge, *Reinventing Capitalism in the Age of Big Data* (New York: Basic Books, 2018), p. 26.
[5]Mayer-Schönberger and Ramge, p. 28.
[6]Antonella Zarra, "Experimentalism beyond ex ante regulation: A law and economics perspective on AI regulatory sandboxes," *Cambridge Forum on AI: Law and Governance* 2 (2026), e58, p. 3.
[7]Zarra 2026, p. 4.

The AIA does build a substantial information-production regime outside of sandboxes. Articles 11–13 require providers of high-risk systems to maintain technical documentation, ensure logging and traceability, and supply transparency information to deployers and users—outward-facing flows that generate records suitable for post-hoc monitoring. The argument here is not that this regime is absent but that it operates asymmetrically. It disciplines deployers and produces records available to regulators. It does not redistribute the core epistemic asymmetry between developers and the public over training data, model weights, and methodology. The documentation Articles 11–13 require is defined and mediated by providers themselves, and the sandbox framework, as the following sections show, institutionalises a form of opacity that even this regime does not reach.

## Risk Assessment, Risk Management, and the Logic of Article 5

To evaluate how the Act navigates the asymmetry described above, Aven's distinction between risk assessment and management is useful. Assessment produces decision support, not decisions in and of themselves; as Aven notes, there is "always a need for managerial review and judgment, which sees beyond the results of the analysis and adds considerations linked to the knowledge and lack of knowledge on which the assessments are based."[8] The AIA describes itself as a risk-based framework, categorising AI systems by risk tier and applying graduated obligations accordingly. Article 6 confirms this default logic: high-risk systems trigger conformity obligations calibrated to harm potential, with information flowing outward toward regulators and users. Article 5, however, breaks from this logic entirely. Its prohibitions are categorical. There is no proportionality mechanism, no sandbox pathway, no supervisory

[8]Terje Aven, "Risk Assessment and Risk Management: Review of Recent Advances on Their Foundation," *European Journal of Operational Research* 253, no. 1 (2016): 1–13, at p. 8.

override—the provision forecloses assessment before it can occur. The contrast is not one of degree but of kind. Article 6 assumes sufficient knowledge to classify and condition. Article 5 assumes the normative judgment has already been made, and classification is beside the point. The Act's own architecture confirms this asymmetry: Genicot and Guimaraes Moraes observe that under the real-world testing mechanism in Article 60, "most requirements arising from the AI Act are not applicable"—but that the prohibitions in Article 5 "continues to apply" throughout testing in real-world conditions.[9]

This is not necessarily illegitimate. Many harmful outcomes from AI use may warrant normative prejudgment, and the European tradition of constitutional rights protection—particularly the Charter of Fundamental Rights—provides a defensible source for it. The recitals accompanying Article 5 do gesture at such a source, invoking human dignity, the Charter, and the harms of manipulation. But these are stated as given rather than as reasons open to challenge: the Act treats its categorisation as settled.

The point is worth pressing because it sits at the centre of an old debate. Does legitimate governance proceed by stipulating substantive ends, or by securing fair procedures? John Hart Ely's *Democracy and Distrust* offered the most influential modern defence of the procedural side of this quarrel, arguing that legitimate judicial review should not impose substantive values, but should police the procedural conditions under which democratic decisions are made—clearing the channels of political change and protecting groups whose interests are systematically excluded from those channels.[10] The appeal of the position lay in its modesty: courts would not

[9]Nathan Genicot and Thiago Guimaraes Moraes, "Exploring the boundaries of AI regulatory sandboxes under the AI Act: Flexibility and real-world testing," *Cambridge Forum on AI: Law and Governance* 1 (2025), e36, p. 10.
[10]John Hart Ely, *Democracy and Distrust: A Theory of Judicial Review* (Cambridge, MA: Harvard University Press, 1980), pp. 73–180.

need to defend contested first-order commitments, only to ensure that the procedures through which such commitments were debated and adopted remained open.

The AIA presents itself in essentially Elyan terms. Its organising rhetoric is procedural and risk-based—systems are classified, conformity is assessed, sandboxes generate evidence, market surveillance authorities supervise, and the whole apparatus is described as a calibrated mechanism for managing harm without committing to contested substantive judgments. The claim, both implicit and frequently explicit, is that the framework is value-neutral with respect to the AI applications it governs: harms are scaled to obligations, classifications are revisable in light of evidence, and the regulation does not pre-judge which technologies are good or bad.

Laurence Tribe's "The Puzzling Persistence of Process-Based Constitutional Theories" denied that the procedural modesty Ely defended was achievable: any procedural theory depends on prior substantive judgments about which interests count as worthy of representation, what fairness consists in, and which exclusions are intolerable.[11] Process cannot be specified without substance. The appearance of value-neutrality conceals the very normative commitments the procedural turn was designed to evade. The AIA exhibits this move almost paradigmatically. Article 5 is not the output of the risk-based apparatus. It is a set of substantive normative commitments that the procedural framework cannot generate from within itself, inserted and stipulated as foundational. The framework cannot tell us why social scoring is prohibited while a high-risk recruitment screening tool is admitted to a sandbox—only that the Union has decided in advance that one is intolerable and the other is not. The risk classification cascade in Articles 6 to 51 calibrates obligations for systems whose admissibility has already been settled by Article 5. It does no normative work of its own. As Tribe observed of process-based constitutional theories,

[11]Laurence H. Tribe, "The Puzzling Persistence of Process-Based Constitutional Theories," *Yale Law Journal* 89, no. 6 (1980): 1063–80.

the substantive commitments do not disappear when concealed behind procedural rhetoric—they merely become harder to contest. The risk-based framework can be revised in light of new evidence. The prior normative judgments on which Article 5 rests cannot: they are the settled boundary against which revision proceeds.

Article 5 forecloses certain practices categorically. Article 6 and the documentation regime in Articles 11–13 establish graduated obligations that channel information outward. Articles 57–63 invert this logic entirely. Article 5 assumes the normative judgment is settled in advance. Article 6's classification cascade assumes sufficient knowledge to grade obligations. The sandbox framework assumes deployment is required to generate that knowledge. The direction of information flow reverses, and the Act offers no account of how these regimes cohere. A hybrid regime combining categorical bans with supervised experimentation could be principled—but only if the Act explained why some practices warrant categorical closure while adjacent ones warrant supervised learning. The problem is not that the AIA is hybrid; it is that it treats the boundary between these modes as self-evident rather than justified, foreclosing the very question an accountable framework would need to answer.

## Sandboxes as Institutionalised Asymmetry

The sandbox framework is not a corrective to information asymmetry but its institutionalisation under regulatory supervision. The claim is not that sandboxes are unregulated; documentation requirements, testing-plan obligations, supervisory oversight, and the data-protection carveout all impose real constraints. The claim is rather that they regulate opacity without redistributing it to affected persons or the public. Article 57(5) establishes the sandbox as a "controlled environment that fosters innovation and facilitates the development, training, testing and validation of innovative AI systems… before their being placed on the market." The regime exists precisely

because conformity assessment cannot be completed in advance — an admission built into the structure of the regime, not incidental to it. Article 57(9)(d) makes this explicit, claiming that "evidence-based regulatory learning" is an objective of these sandboxes; they are more a form of supervised observation than oversight in the Article 5–6 sense. The data protection carveout in Article 59 reinforces this controlled information flow: further processing of personal data in sandbox contexts is permitted only under specified conditions and remains subject to safeguards, rather than becoming a vehicle for public disclosure. This inversion is openly acknowledged even by sympathetic accounts: Lanamäki and colleagues note that "the exact nature and implementation of regulatory sandboxes remain ambiguous," and Zarra's law-and-economics defence treats the regime as one whose primary role is to remove information asymmetries via structured engagement between developer and regulator. The sandbox is not merely a waiver for earlier obligations but a restructuring of the underlying information architecture, accumulating knowledge for government and industry use.

Article 60, meanwhile, permits real-world testing of high-risk systems outside sandboxes entirely, subject to a testing plan and market surveillance approval. It explicitly allows the developer to retain the opacity of their system while affording regulators observational access, institutionalising a form of productive asymmetry that does not extend to the market or to the public. The one formal boundary is that real-world testing must be conducted "without prejudice to the prohibitions under Article 5." Read hierarchically, the clause means Article 5 trumps the experimental regime where they would otherwise conflict. In practice, it functions as a drafting move: the two regimes coexist without the Act ever explaining why the Article 5 floor falls where it does. As Genicot and Guimaraes Moraes observe, this coexistence is workable only because Article 60 "allows an AI system to be tested in a real-life setting with no need for the

provider to carry out a conformity assessment beforehand," while most AIA requirements are simply not applicable during real-world testing, with Article 5 the lone hard floor.[12]

What, exactly, does sandbox participation produce? Alex Moltzau, a Seconded National Expert at the European AI Office, has noted that "regulatory sandboxes do not provide conformity assessment," and that providers seeking compliance certainty "should either consult their national notified body or perform a self-assessment."[13] The sandbox produces something else entirely: a plan that, as Genicot and Guimaraes Moraes describe it, "acts as a contract between the regulator and the regulated party," together with limited sanctions immunity under Article 57(12) for participants acting in good faith and an exit report that market surveillance authorities are obliged to take "positively into account" in subsequent conformity proceedings.[14] The arrangement is generous from the developer's side: structured engagement, regulatory familiarity, fine immunity, and reputational signalling. From the public's side, it produces no enforceable accountability output.

Articles 62 and 63 sharpen the point further by calibrating who benefits from this institutionalised asymmetry. Article 62 directs Member States to facilitate priority access to sandboxes for SMEs and start-ups, and Article 63 establishes derogations permitting smaller providers to satisfy certain documentation and quality-management requirements through simplified means. The intent is plainly to avoid regulatory capture. But viewed through the lens of information asymmetry it is striking: the Act distributes the right to retain opacity, treating opacity itself as a resource it allocates to favoured beneficiaries. This is the language of industrial policy, treating epistemic position as a tradeable good. The framework's underlying

---

[12]Genicot and Guimaraes Moraes, p. 10.
[13]Lanamäki et al., p. 9.
[14]Genicot and Guimaraes Moraes, p. 5; European Parliament, *Regulation (EU) 2024/1689*, Articles 57(7) and 57(12).

assumption—that opacity is a productive input which the regulator distributes by class of provider—is the same assumption Article 5 implicitly rejects when it forbids certain practices precisely because of the asymmetry they entail. The distributional question sharpens further once the jurisdictional dimension is added. The European Parliamentary Research Service warns that fragmented enforcement across Member States risks producing uneven sandbox capacities, with the consequence that "AI providers might therefore intentionally choose less stringent sandboxes"—a sandbox-arbitrage problem the implementing acts have not yet resolved. As the EPRS reported in March 2026, August 2025 data showed that only one Member State sandbox was operational, leaving the question of how this distribution will play out in practice almost entirely open.[15]

Taken together, Articles 57–63 do not correct the firm-model information architecture for the benefit of the public. They replicate it under regulatory supervision. The regulator assumes the position of the centre to which information streams, but the developer's competitive opacity over training data, weights, and methodology is preserved throughout. The closed architecture of generative AI is not opened by the sandbox; it is sustained within it. Article 5's categorical floor and the sandbox regime, meanwhile, remain in contest.

## Distributive Consequences and the Accountability Deficit

At its core, the gap between Article 5 and Articles 57–63 is not a drafting oversight but a regulatory choice with distributive consequences. The Act's risk-based approach implies a unified framework in which graduated obligations track the severity of potential harm, but this implication is misleading. The Act relies on two different regulatory logics. Article 5 bans some

[15] Marcelin 2026, p. 3.

practices outright on normative grounds. The sandbox regime allows supervised deployment in order to generate knowledge. The two approaches can coexist—hybrid frameworks can be principled—but the Act never explains why the boundary between them falls where it does.

This has distributive consequences. The sandbox framework preserves developer opacity throughout the knowledge-generation process. The informational risk of deployment is therefore borne not by the developer but by the subjects of testing and, ultimately, by the public. Article 5's categorical prohibitions protect against the most acute concentrations of this risk, but they do so by normative fiat rather than assessment. The Act identifies certain asymmetries as intolerable while quietly institutionalising adjacent ones. The stakes are sharpest in boundary cases: systems that fall just outside Article 5's prohibitions can enter the sandbox framework, retaining opacity while generating the regulatory knowledge on which future governance will depend. The public bears the epistemic cost of that knowledge production without any formal mechanism for sharing its benefits. The compliance economics sharpen the concern. A small enterprise of fewer than fifty employees and ten million euros in turnover can expect compliance costs of around three hundred thousand euros for a single high-risk AI system, an estimate Zarra cites from the CEPS study commissioned to support the Commission's impact assessment.[16] The Act treats these as categorically distinct by asserting that certain uses of the underlying architecture are intolerable prior to any assessment, but the basis for that assertion is never defended, only enacted.

The accountability gap this opens is not theoretical. Even Zarra, whose paper defends the sandbox concept, warns that "poorly designed or under-resourced sandboxes can become 'innovation theatres' where institutions celebrate innovation activities without fully evaluating their impacts," a failure mode in which sandbox participation becomes a reputational asset for

[16]Zarra, p. 13, citing Andrea Renda et al., *Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe* (Luxembourg: Publications Office of the European Union, 2021).

providers and a public-relations vehicle for regulators without producing meaningful governance.[17] Sandbox participation generates reputational and informational benefits for the developer. The regulator obtains familiarity with the technology. The public obtains an exit report whose abstract may be published; anything further is at the participant's discretion. As Genicot and Guimaraes Moraes observe, "it is unlikely that sharing only abstracts will be enough for proper public scrutiny."[18]

The Act is therefore not distinguishing between harmful and productive informational asymmetry on epistemic grounds but on normative ones. It does include mechanisms for evaluating and updating its risk classifications, so the boundary is not formally unchallengeable. But formal revisability is not the same as genuine contestability. Revision under the AIA is driven by expert bodies and Commission review; the public who bear the costs of those arrangements have only thin and indirect means of challenging them. Genuine contestability would require, at minimum, a public rationale explaining why a given practice falls within Article 5 rather than the sandbox-eligible category. That rationale would have to be open to challenge by those bearing its costs. No mechanism in the Act requires a public-facing justification of why a given practice sits on the Article 5 side of the line rather than the sandbox-eligible side.

Finally, the Act's instrument tension is partly a predictable artefact of the conditions under which the Regulation was produced. The Commission's 2021 proposal preceded the public release of ChatGPT by eighteen months, and the late-stage scramble to accommodate general-purpose AI—culminating in the bifurcated regime in Articles 51 to 55, substantially absent from

[17]Zarra, p. 17.
[18]Genicot and Guimaraes Moraes, p. 12.

the original proposal—reshaped the Regulation under acute political and temporal pressure. The sandbox provisions themselves bear marks of this compromise: Genicot and Guimaraes Moraes note that "in the initial version of the AI Act… the regime applicable to AI regulatory sandboxes did not provide any form of regulatory flexibility," with the fine-immunity provisions of Article 57(12) added later rather than designed by the Commission.[19] Zarra documents the broader political-economy context: intense lobbying by EU AI champions during the Act's final negotiations, and the Draghi Competitiveness Report's call for simplification, together produced "sweeping simplification efforts that impacted the newly adopted AI rules."[20] Capacity disparities across Member States compound this; one panellist on the regime observes that "many state agencies lack such expertise, often due to limited budgets or salary constraints—an issue especially pronounced in Eastern Europe."[21] A regulation that emerges from such a process is required to be different things to different audiences: a Charter-grounded instrument to rights-protective constituencies, a calibrated framework that licences experimentation to innovation-focused ones, and a regime that allocates regulatory burden lightly to Member States anxious about competitiveness. The categorical commitments must be unconditional to satisfy the rights-protective audience; the sandbox apparatus must be elastic to satisfy the innovation audience; and the boundary between them must remain unjustified to keep both in the coalition. The unjustified boundary is, in this sense, less an accident of drafting than a condition of the Act's existence—the price of producing a regulation at all under the conditions in which it was negotiated.

[19]Genicot and Guimaraes Moraes, p. 7.
[20]Zarra, p. 6.
[21]Lanamäki et al., p. 7 (Niemeyer's panel contribution).

## Conclusion

This essay has argued that the AI Act does not draw a coherent line between harmful and productive information asymmetry, but instead leaves the tension unresolved. The first section established why GenAI is structurally resistant to conventional transparency requirements: asymmetry is not incidental to the sector but constitutive of it. Regulatory intervention therefore engages not with a market pathology but with a firm-model architecture in which opacity is load-bearing. The second section showed that Article 5 does not operate on the risk-based logic the Act claims as its organising principle. It rests on prior normative commitments that risk assessment cannot generate. The third section showed that Articles 57–63 do not correct that architecture but replicate it under regulatory supervision, allocating the epistemic costs of knowledge production to the subjects of testing and deployment. The fourth section traced the distributive consequences and the accountability deficit that follow.

This is not to argue that the sandbox framework is illegitimate or that the Act's categorical prohibitions are unjustified. Supervised opacity may be the most viable regulatory posture available given the nature of the sector, and some harms may warrant normative prejudgment rather than graduated assessment. Rather, it is to suggest that a framework which depends on information asymmetry to govern information asymmetry cannot treat its own normative commitments as self-evident. What accountability would require is not the elimination of normative commitments from governance, but their legibility: criteria explicit enough to be contested, revised, and answered for by those who bear their costs. The AIA's deeper failure is not that it draws a line between prohibited and sandbox-eligible systems but that it declines to expose the reasoning by which the line is drawn. A framework that cannot explain why certain asymmetries are intolerable while institutionalising adjacent ones cannot plausibly claim to be

governing on behalf of the public whose informational agency it is meant to protect. The question the Act leaves open is not whether asymmetry in AI can be governed, but whether governance that depends on asymmetry can itself be held to account—and the answer depends entirely on whether the Act, and the political coalition that produced it, are willing to defend, rather than merely enact, the choices it makes.

Works Cited

Aven, Terje. “Risk Assessment and Risk Management: Review of Recent Advances on Their Foundation.” *European Journal of Operational Research* 253, no. 1 (2016): 1–13. https://doi.org/10.1016/j.ejor.2015.12.023.

Ely, John Hart. *Democracy and Distrust: A Theory of Judicial Review*. Harvard University Press, 1980.

Genicot, Nathan, and Thiago Guimaraes Moraes. “Exploring the Boundaries of AI Regulatory Sandboxes under the AI Act: Flexibility and Real-World Testing.” *Cambridge Forum on AI: Law and Governance* 1, no. Article e36 (2025): 1–17. https://doi.org/10.1017/cfl.2025.10013.

Lanamäki, Arto, Karin Väyrynen, Fanny Vainionpää, et al. “What to Expect from the Upcoming EU AI Act Sandboxes: Panel Report.” *Digital Society* 4, no. 42 (2025). https://doi.org/10.1007/s44206-025-00201-x.

Marcelin, Tristan. “AI Regulatory Sandboxes: State of Play and Implementation Challenges.” *European Parliamentary Research Service, At a Glance*, At a Glance briefing, PE 785.673, March 17, 2026. https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2026)785673.

Mayer-Schönberger, Viktor, and Thomas Ramge. *Reinventing Capitalism in the Age of Big Data*. Basic Books, 2018.

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) (Text with EEA Relevance), Official Journal of the European Union (2024). http://data.europa.eu/eli/reg/2024/1689/oj.

Tribe, Laurence H. “The Puzzling Persistence of Process-Based Constitutional Theories.” *Yale Law Journal* 89, no. 6 (1980): 1063–80. https://doi.org/10.2307/795856.

Zarra, Antonella. “Experimentalism beyond Ex Ante Regulation: A Law and Economics Perspective on AI Regulatory Sandboxes.” *Cambridge Forum on AI: Law and Governance* 2, no. Article e58 (2026): 1–23. https://doi.org/10.1017/cfl.2025.10039.