COMMENTARY PAPER

Pipeline Attestation and Digital Rights

Security-by-design: A Case for Digital Rights Enforcement Through Data Pipeline Attestations

Bhavyatta Bhardwaj[1]

[1]Independent Researcher & Consultant. Hamilton, ON, Canada
info@bhavyattabhardwaj.com

Keywords: data conditioning; pipeline attestation; Lovelace Threshold; digital rights enforcement; security-by-design; vulnerability management; rights-centric substrate
Abbreviations: CVE: Common Vulnerability and Exposure; GDPR: General Data Protection Regulation; ZTA: Zero Trust Architecture; DoD: Defence in Depth; OSI: Open Systems Interconnection; SbD: Security-by-design

Abstract

You cannot protect what you cannot see is a foundational cybersecurity axiom. This commentary extends that principle one layer upstream: digital-rights frameworks struggle to govern what they cannot observe, regardless of the technology at play. Current rights regimes, particularly the GDPR and the EU AI Act, retain a "Lovelace World" assumption that digital systems execute human instructions in ways that are fully legible and specifiable. This presupposition leaves upstream data-conditioning pipelines largely unregulated, rendering rights claims unenforceable before any regulated endpoint is reached. Drawing on practitioner experience in vulnerability management and applying Security-by-Design (SbD) logic to governance architecture, this paper introduces the *Lovelace Threshold* diagnostic lens to expose structural blind spots in rights enforcement. Rather than proposing detailed technical specifications, the analysis highlights how observability and attestation can ground governance in verifiable practice. Although artificial intelligence provides a timely illustration, the argument applies to any governance framework where enforcement mediates human rights. Ultimately, the paper calls for embedding observability and traceability at the heart of compliance design, ensuring that digital-rights enforcement becomes a development principle and a precondition for credible accountability of rights protection across data ecosystems.

Policy Significance Statement
Clarifying the role of the upstream data-conditioning substrate to policymakers is essential for creating layered governance for digital systems because at this layer, training data is selected, structured, and weighted. Using vulnerability management practices as an operative principle for human-rights-centric policymaking in a multi-dimensional digital landscape underscores the need for a binding regulatory architecture for data-conditioning pipelines. This intersectional approach is foundational to determining the operational parameters of these digital systems and the social realities they reproduce.

1. Introduction

How we name a problem determines how we solve it. Digital rights are human rights and the infrastructure that governs them is currently pre-instrumental to where the harm originates. We have an opportunity to raise the governance standard before the substrate consolidates; that window does not stay open indefinitely. The inequities now exposed by digital infrastructure are structural and long-standing. The rapid advances in the infrastructure make them visible at scale but calling them *emergent* obscures their origin, and origin determines remedy. Current frameworks govern the *right systems* for the *wrong era*.

This commentary introduces the *Lovelace Threshold* as a conceptual diagnostic device that exposes where the assumptions of current rights frameworks no longer match the operational logic of generative and autonomous systems. Using an intersectional approach across philosophy of computation, cybersecurity, and data laws, I derive this threshold between the *Lovelace World*, where systems merely execute predefined instructions, and the

Post-Lovelace World, where systems anticipate, generate, and act beyond explicit specification. The gap between the two is a structural vulnerability that operates beyond current rights frameworks' reach. When I apply the vulnerability management concept to Lovelace's Objection on the Babbage Analytical Engine (Turing, 1950; Lovelace, 1843), this perspective grounds the gap between the two and creates a policy vulnerability. Vulnerabilities are gaps between what frameworks assume systems do and what systems actually do.

Furthermore, using SbD reasoning to address the governance gap: effective controls must be embedded within the data substrate before deployment rather than remediated after harm appears. Pipeline attestation represents that principle operationalized at the data-conditioning layer. The underlying logic is straightforward: human vulnerability translates directly into governance exposure; when rights-based governance operates only after harm occurs, that lag becomes a policy vulnerability that malicious or negligent actors can exploit. Applying SbD to governance therefore protects individuals and systems holistically. (Miller and Piccone, 2015; Bygrave, 2022)

The *General Data Protection Regulation* (GDPR) (European Parliament, 2016) and other data regulations were built for the *Lovelace world.* That has ended especially since the rise of generative AI. And while the *EU AI Act* was prepared for exactly this, falls short and only governs up to the *threshold*. The foundational assumption of many current rights frameworks is no longer tenable in the *Post-Lovelace* world. Reactive governance was coherent when system behaviour was bounded and auditable. Now that the systems can anticipate, generate, and act beyond explicit specification, the harm horizon is no longer legible in advance. Governance needs to prepare for the unknown and not wait for a failure mode it cannot see coming.

Recent empirical work (Folkerts, L. et al., 2026) demonstrates directly that frontier AI models executing autonomous cyberattacks on purpose-built ranges show log-linear capability scaling with no observed plateau, completing attack sequences at speeds that outpace human expert response windows. The applied SbD view aligns with broader anticipatory governance literature, which emphasizes that emerging technologies require forward-looking institutional design rather than reactive rulemaking (OECD, 2024; Engler, 2024). It also reflects the growing recognition in cybersecurity practice that SbD principles must be embedded across the development lifecycle, not appended after deployment (Tsekmezoglou, 2024; IBM Institute for Business Value, 2024; Microsoft, 2025). Pipeline attestation is that principle operationalised at the conditioning layer so upstream governance clears all assumptions for data–no dataset, no transformation process, no conditioning pipeline is assumed clean by virtue of having been used before. In AI, for example, every input to a foundational model must verify its provenance, transformation logic, and withdrawal mechanisms before it is trusted with the substrate.

1.1. Methodology

To examine this conceptual gap, I adopt an analogical-abductive approach across domains as its primary method. I draw structural logic from cybersecurity's SbD principles and apply it to rights-based governance more broadly. The commentary proceeds in two parts. *First*, I map vulnerability management practice onto the Lovelace Test to derive the *Lovelace Threshold*. Using this lens, I explore the regulatory paradigm with GDPR and the EU AI Act to showcase that while improvements are intentional, current frameworks apply endpoint instruments to a post-endpoint problem. *Second*, I use data-conditioning pipelines as the structural proof that makes digital rights exercisable. The argument does not claim that data-conditioning pipelines are security vulnerabilities in a technical sense, but to show that operational and structural challenges in governance are identical: a framework designed for a bounded, auditable system that cannot see the substrate beneath it. This produces a *multi-dimensional policy landscape* spanning the technical, legal, and human-rights-centric layers simultaneously that could benefit from an intersectional approach as discussed in Section 4.1.

2. The Lovelace Threshold and Security-by-Design

Ada Lovelace's firm conviction about the Analytical Engine's fundamental nature is encapsulated in what has come to be known as "Lady Lovelace's Objection" or simply "Lovelace's Objection." In her Notes, (Lovelace, 1843, Note G) she stated:

> "The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has not power of anticipating any analytical relations or truths."

Turing later argued that machines can surprise us (Turing, 1950). Lovelace's objection was subsequently formalised as the "Lovelace Test", proposing that for a computer to genuinely exhibit true creativity, it must originate something truly novel and surprising (Bringsjord, Bello, Ferrucci, 2001). If a system merely remixes existing information or applies pre-programmed rules, even in incredibly complex ways, it still falls under Lovelace's objection. When

repurposed as a governance instrument, this distinction is critical for human-rights-centric data frameworks that are built on assumptions about system predictability.

At the base of every system we cannot predict all consequences of systems in advance. When this assumption fails, the framework's scope fails. Natale and Henrickson (2022) demonstrate that perceptions of machine creativity are socially constructed: what we believe machines can do shapes how we govern them, often before the technical reality catches up. That boundary between a system that executes what it is told and a system that operates beyond explicit determination marks a governance inflection point current governance frameworks have yet to address. The *Lovelace Threshold* identifies this point, where governance based on predictability fails.

In the *Post-Lovelace World*, systems anticipate, generate, and act beyond explicit specification. The assumption of fully determined, bounded, auditable behaviour no longer holds as a governance baseline. Therefore, if machines can surprise us, governance frameworks are structurally exposed and cannot wait for harm to occur before responding. Frontier AI governance capability removes the legible harm horizon that reactive human-rights frameworks depend on, especially agentic AI systems that are no longer fully determined. This shift is both conceptual and structural: machine creativity and generative capacity alter the conditions under which prediction and auditability remain reliable governance assumptions (Natale and Henrickson, 2022; Turing, 1950). The result is a regime where capability can outpace the visibility structures that rights frameworks depend on.

2.1. Vulnerability Management

The collapse of the Lovelace assumption creates a moment where the system behaves beyond what the governance frameworks are designed to see. This process gap demands governance response most legible through vulnerability management practice. I use it as a diagnostic analogy: the upstream layer of AI systems is not a software vulnerability in a technical sense, but the governance failure mode is structurally identical. In practice, I observe that the harm lies in the gap between what a framework assumes a system does and what it actually does under critical operations–also where governance rarely looks. The logic is to find the vulnerability, disclose it, patch it. Common Vulnerabilities and Exposures (CVEs) catalogue known weaknesses but only systems that are deployed and observable are in scope.

According to NIST, vulnerability management is the capability that identifies gaps (CVEs) on devices that are likely exploited and used as platforms for broader compromise. Data-conditioning pipelines are not currently registered within regulatory scope for rights enforcement, as technology vulnerabilities are within the CVE register. This commentary proposes importing similar integrity principles into data governance to close that structural gap. Adopting such instruments could support human-rights-centric approaches to digital rights. Individuals currently lack legal mechanisms to identify, challenge, or seek redress for conditioning-layer uses of their material.

2.2. Defense-in-Depth and Zero Trust as Operative Models

SbD embeds controls into the architecture before deployment rather than remediated after harm surfaces. Data-conditioning pipelines, the pre-instrumental layer of digital governance, operate upstream of regulated endpoints, beyond the reach of current compliance and audit mechanisms. They process and structure raw information before it becomes identifiable data, yet remain outside the scope of formal oversight or accountability procedures. Traditional cybersecurity governance assumed fully human-determined system behavior, reflected in rule-based access controls, network segmentation, and other static security configurations codified through manual policy rules and privilege hierarchies. These mechanisms presuppose that system logic is explicit and bounded, which is now an assumption that no longer holds for generative or adaptive systems. The architectural parallels with OSI *Defence-in-depth*[1] and *Zero Trust* highlights that governance must be credible at the conditioning stage as well as the deployment stage. For policy, it would mean that no dataset or transformation process should be presumed compliant merely because it has been used previously. Each must demonstrate verifiable provenance, transformation logic, and withdrawal mechanisms before it is trusted with the substrate. It is already operationalised in two relevant ways:

1. *Zero Trust Architecture operationalises it at the identity layer:* no actor, internal or external, is assumed safe by virtue of position. Trust is never inherited from prior access and every actor must verify before entry.

[1] The commentary doesn't claim to map governance onto OSI's model, but adapts layered governance to data-conditioning.

2. *Defence in depth operationalises it at the architectural layer:* each layer of a system must be governed independently, because a failure at one layer cannot be compensated for by controls at another.

Both principles share the same anticipatory logic that we must govern before harm, not after. These design principles also appear in contemporary governance guidance for AI and digital systems, where SbD approaches emphasize continuous validation, layered controls, and lifecycle accountability (Microsoft, 2025; IBM Institute for Business Value, 2024; NIST, 2023). Pipeline attestation extends it to rights-governance architecture by protecting the conditioning layer that the ethical and regulatory standards have not yet been stipulated to reach. The *Lovelace Threshold* doesn't consider prior use as provenance and asks **why are datasets assumed clean by virtue of prior use**; and showcases that ungoverned data conditioning reproduces exactly the failure mode *defence-in-depth* was designed to prevent by implying layered governance.

2.3. Layered Governance

Current digital-rights frameworks govern outputs while leaving the generative substrate unregulated. Data-conditions are dynamic, and require effective governance to manage variation rather than treat it as constant. In a *Post-Lovelace* world, that variation is the default.

Data-conditioning pipelines perform the formative function of transforming raw digital traces into policy-relevant evidence: determining what enters the system, how it is structured, what meaning it carries, and the legal basis for its subsequent use. The regulatory question is whether that process can be made bidirectional by requiring verifiable proof of the conditioning integrity before any output is produced. This employs the same anticipatory governance logic as vulnerability disclosure, identifying and addressing risks prior to exploitation. This logic has yet to be extended systematically to the data-conditioning layer to mandate verified and transparent conditions so rights frameworks can operate as intended and their outcomes remain equitable. But when conditioning varies without oversight, those variances are inherited by the frameworks themselves, producing uneven rights enforcement. Managing that inheritance after deployment does not constitute structural risk mitigation. Since every downstream rights claim depends on the integrity of this substrate, governing it is prerequisite for credible digital-rights enforcement.

3.The Regulatory Ceiling

When the *Lovelace Threshold* is applied as a diagnostic to any governance framework, the threshold asks: **was it designed for a world in which system behaviour is fully and holistically determined and auditable?** If yes, it is a *Lovelace-World* and the instruments are designed for a system whose behaviour is, in principle, fully legible. Otherwise, in the case of generative and agentic AI, that legibility assumption is what breaks the *Post-Lovelace* transition. Once systems operate beyond explicit specification, the framework's reach cannot extend to a conditioning layer whose outputs it can no longer predict or bound, however well designed for its own world.

The EU AI Act now requires high-risk AI systems to be designed and developed to achieve an appropriate level of accuracy, robustness and cybersecurity, and perform consistently as intended throughout their lifecycle (Regulation (EU) 2024/1689, art 15(1); Tsekmezoglou, 2024). Moreover, other design-focused requirements have been proposed which in effect demand logging by design, transparency by design and human oversight by design for high-risk AI systems. (Regulation (EU) 2024/1689, art 12(1), 13(1) and 14(1)). This is good practice to have foundational principles in regulations.

The EU AI Act has moved directionally closer to the upstream layer by requiring training data disclosure for general-purpose AI models (Regulation (EU) 2024/1689, art 53) This is a meaningful step toward transparency, yet architecturally insufficient because disclosing which datasets were used is not the same as attesting to how data was conditioned–the transformation, weighting, filtering, and structuring that occurs before raw data becomes training material. And one might trust the training data, the issue of recombining the trustworthy data into new answers as generative LLMs do in a new context may lead to misplaced trust (Li, 2023).

This matters because governance by design has increasingly been proposed as the only practical way to make complex AI systems accountable at scale (Upmann, n.d.; Burnham, 2025; Cihon, Maas and Kemp, 2023). The EU AI Act's lifecycle and transparency provisions are therefore important, but they still stop short of making the conditioning layer itself a direct object of enforceable rights governance.

For AI, while organizations work on estimating how dangerous a system can be (Bova, et al., 2024), deferring risk classification to declared use cases introduces capability-blindness which is a structural vulnerability rather than a deliberate policy decision. A capability that can execute high-risk functions becomes a governed risk the moment it enters the environment. From *SbD* standpoint, deferring classification to declared use-cases exposes the system to high‑impact failure modes the moment such tools are integrated. While it does not formally qualify as a "high‑risk" deployer under the EU AI Act (Regulation (EU) 2024/1689, Annex III), the systems and users are at a risk of data leakage, covert high‑risk decision support, over‑reliance which can become a vulnerability class, not a policy position.

Similarly, GDPR governs processing endpoints: consent, breach notification, individual rights of access and erasure. Mahieu et al. (2021) demonstrated that GDPR's right-to-access provisions measurably improved individual rights outcomes and accelerated regulatory convergence globally through the Brussels Effect. (Bradford, 2020) While meaningful, this evidently points to the structural ceiling of endpoint governance where a framework is designed for the processing layer. The *Lovelace Threshold* views this as well enforced yet insufficient for the conditioning process that precedes it. A *Post-Lovelace* system does not merely use disclosed data; it extends, recombines, and acts on it in more fundamental provenance disclosure and stipulates attestation to whether the sources themselves were equitable–who was included, who was excluded, and on what basis. Data cannot speak for itself (D'Ignazio and Klein, 2020) therefore a record of origin is not a proof of integrity but risk.

Current regulatory instruments approach but do not yet penetrate the pre‑instrumental layer which is the level at which systemic inequity becomes embedded. At this upstream layer, individuals recognised in law as rights‑bearing subjects are rendered legible only as data points. Their claims to dignity and redress cannot yet traverse the substrate that existing frameworks do not formally govern. The individuals treated as data points were rights-bearing subjects whose dignity was absorbed into a substrate that no current framework can see, reach, or redress.

4. Pipeline Attestations for Market Access

Security practitioners have a unique worldview, and we understand that it takes a particular mindset to understand risks. A much matured vulnerability management field now catalogues adversarial attack surfaces using tools like MITRE ATLAS (NIST, 2023) and AI Risk Management Framework (NIST, 2023) developed to address training data integrity. However, the reality of governance today reinforces security into thinking about justice (Miller and Piccone, 2015). Critical data studies by D'Ignazio and Klein (2020) focus on data justice approaches rather than data ethics which implies that security and justice are not separate from each other, promoting a human-rights-centric data governance. Therefore, pipeline attestation asks whose data, whose categories, and whose absences structure the datasets on which systems train. I am proposing to govern these inequities through attestation as an enforceable policy instrument. It's an anticipatory risk standard to attest upstream for data rights law.

Pipeline Attestations would function as a certification requirement analogous to pre-market conformity assessments. Attestations must document selection criteria, transformation logic, and withdrawal mechanisms, verified by an independent auditor before system deployment. This becomes a mandatory and transparent condition of market access. Specifically, what data was included and excluded from a training pipeline and on what criteria; what transformation, weighting, and filtering operations were applied and by whom; and what withdrawal mechanisms exist for data subjects whose material entered the pipeline. Without attestation, no certification, therefore, no market access. This commentary is consistent with wider calls to embed compliance and ethics in system design rather than rely on post hoc review (Cobbe, Lee and Singh, 2021; Binns and Edwards, 2025). It also complements data-rights scholarship that treats access, contestation, and provenance as central to meaningful rights enforcement rather than formalities alone (Ausloos and Veale, 2021; Mahieu et al., 2021).

Information technology–Artificial intelligence–Data life cycle framework (ISO/IEC 8183:2023) establishes that data lifecycle management from acquisition through decommissioning is technically standardisable. The technical infrastructure for pipeline attestation exists that provides machine-readable attestation standards for software supply chains (OWASP CycloneDX, 2024), establishes provenance verification for build pipelines (Google SLSA, 2023), and training data integrity (NIST AI RMF, 2023). Technical feasibility of these standards can make attestation legally consequential, but self-regulation without deterrence and transparent regulations are different instruments. Effective

digital rights enforcement requires all three as applicable and operational conditions. The *Lovelace Threshold* anticipates that a framework designed for a bounded system, however well enforced at its own layer, cannot reach the conditioning process that precedes it. We see an architectural spread globally through compliance (Bradford, 2020). This structural ceiling built into the instrument's design no longer applies.

4.1. Multi-dimensional Policy Landscape: Solutions at the Intersection

The policy problem is not that the threat landscape is too complex but that it is structurally fragmentated. Each governance domain–technical, legal, ethical–addresses only the layer it can see, and coordination between those disciplines struggles to produce holistic governance (Dafoe, 2018). The principle required is anticipatory coordination by aligning oversight functions before systemic failure. Pipeline attestation serves as the mechanism that makes such coordination actionable, translating intersectional insight into a compliance touchpoint across domains.

This commentary is a demonstration of that problem through method. The *Lovelace Threshold* is drawn from the philosophy of computation. The vulnerability management framework is drawn from cybersecurity practice. The rights enforcement analysis is drawn from data law. The argument only becomes visible from the intersection. Intersectional positioning is a structural governance requirement: the capacity to see what each discipline cannot see from inside its own domain is what the current governance moment demands. (Jasanoff, 2004) This is consistent with research showing that policymaking in multidimensional environments cannot be treated as a single-axis problem, because decisions in one domain spill into others in ways that change outcomes and accountability structures (Crosson, Invernizzi and Izzo, 2024).

Data policy is the load-bearing substrate for governance, where ethical concerns like eroded trust or erroneous decisions scale through mis- and disinformation. It can reinforce biases that systemic ignorance allows to perpetuate through training data (Li, 2023). When data policy is treated as one discipline rather than as the ground everything else stands on, the architecture is unstable. Moving between the systemic pattern and the specific decision, between structural cause and individuals whose rights are at stake is a governance requirement that now gets attention, yet operational designs overlook. The decisions being made now about which data conditions foundational systems carry have consequences beyond any single policy cycle. What is conditioned now determines what future governance can see, what future rights claims can be evidenced, and whose history survives as a legible record.

5. Conclusion

Digital-rights governance now faces a structural visibility problem. As systems grow anticipatory and generative, reactive enforcement loses coherence. Applying *SbD* logic to the governance substrate provides a path toward enforceable, anticipatory rights protection. The *Lovelace-to-Post-Lovelace* transition marks the point where the foundational assumption of current digital-rights frameworks begin to fracture. The harm horizon is no longer predictable and under those conditions reactive governance becomes structurally inefficient.

Vulnerability management offers an honest precedent and I applied it analogically to data conditioning pipelines, transforming digital rights from aspirational recourse into an exercisable infrastructure because anticipating the unanticipated is no longer a survival condition. Except for the *Lovelace Threshold*, this commentary doesn’t claim to introduce a new framework, but offers an intersectional insight that the tools for holistic governance are available but not applied to the root of the vulnerability. However, when applied one layer upstream of where regulation has been willing to stop, digital rights can truly be grounded in human rights.

We stand at an inflection point and what is conditioned into foundational systems now determines what future governance can see, what future rights claims can be evidenced, and whose history survives as a legible record. We have an opportunity to secure human dignity from being absorbed by ungoverned data conditioning into a substrate that no current framework can see to govern. This is beyond biased outputs and has a long-term impact that determines what enters the historical record and whose experience becomes evidence for future systems. With such multigenerational stakes dressed as technical decisions, (Benjamin, 2019) the window to govern the substrate before it consolidates does not stay open.

Funding statement. None. This work received no specific grant from any funding agency, commercial or not-for-profit sectors.
Competing interests. Author declares none.
Data availability statement. N.A. The commentary is solely based on the included references, bibliography, and Author's industry experience.
Author contributions. Conceptualization: B.B. Methodology: B.B. Formal analysis: B.B. Investigation: B.B. Writing, reviewing, and editing[2]: B.B. Visualization: B.B. Author approved the final submitted draft.

References

Ausloos, J. and Veale, M. (2021) 'Researching with data rights', Technology and Regulation, 2020, pp. 136–157. https://doi.org/10.26116/TECHREG.2020.010

Benjamin, R. (2019) Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge: Polity Press.

Binns, R. and Edwards, L. (2025) 'ChatGPT tells fibs about me: Are data protection and libel adequate tools to protect reputation in the LLM era?', in Hacker, P., Engel, A., Hammer, S. and Mittelstadt, B. (eds.) The Oxford Handbook of the Foundations and Regulation of Generative AI. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198940272.013.0017

Bova, P., Di Stefano, A. and Han, T.A. (2024) 'Quantifying detection rates for dangerous capabilities: A theoretical model of dangerous capability evaluations', arXiv [cs.AI]. http://arxiv.org/abs/2412.15433

Bradford, A. (2020) The Brussels Effect: How the European Union Rules the World. Oxford: Oxford University Press.

Bringsjord, S., Bello, P. and Ferrucci, D. (2001) 'Creativity, the Turing Test, and the (better) Lovelace Test', Minds and Machines, 11(1), pp. 3–27. https://doi.org/10.1023/A:1011206622741

Bygrave, L.A. (2022) 'Security by design: Aspirations and realities in a regulatory context', Oslo Law Review, 8(3), pp. 126–177. https://doi.org/10.18261/olr.8.3.2

Cihon, P., Maas, M. and Kemp, L. (2023) 'Fragmentation and the future: Investigating architectures for international AI governance', Global Policy, 14(1), pp. 52–63.

Cobbe, J., Lee, M.S.A. and Singh, J. (2021) 'Reviewable automated decision-making: A framework for accountable algorithmic systems', in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), pp. 1–15. New York: ACM. https://doi.org/10.1145/3442188.3445921

Crosson, J.M., Invernizzi, G.M. and Izzo, F. (2024) 'Learning in a complex world: How multidimensionality affects policymaking'. Available at: https://giovannainvernizzi.com/wp-content/uploads/2024/05/Multidimensional_policy_experimentation-5-2.pdf; last accessed 29 March 2026.

Dafoe, A. (2018) AI Governance: A Research Agenda. Oxford: Future of Humanity Institute, University of Oxford.

D'Ignazio, C. and Klein, L.F. (2020) Data Feminism. Cambridge, MA: MIT Press. Available at: https://data-feminism.mitpress.mit.edu/

Engler, A. (2024) 'The state of implementation of the OECD AI Principles: Horizontal analysis across sectors', Policy and Society, 43(2), pp. 112–130.

[2] During the preparation of this manuscript, the author's use of AI was limited to Perplexity for policy vocabulary. On March 29-30, 2026, author used Claude Sonnet 4.6 for grammar feedback and organizing literature. All research, synthesis, and writing remains the author's own.

European Parliament (2016) General Data Protection Regulation (EU) 2016/679. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

European Parliament and Council of the European Union (2024) 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), OJ L 2024/1689. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

Folkerts, L. et al. (2026) 'Measuring AI agents' progress on multi-step cyber attack scenarios', arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2603.11214

Google (2023) Supply chain levels for software artefacts (SLSA) framework. Available at: https://slsa.dev ; last accessed 28 March 2026.

IBM Institute for Business Value (2024) Secure by design, smarter with AI: Redefining cyber resilience for the age of intelligent threats. Available at: https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-secure-design-cyber-resilience; last accessed 28 March 2026.

ISO/IEC 8183:2023 (2023) Information technology — Artificial intelligence — Data life cycle framework. Geneva: International Organization for Standardization. Available at: https://www.iso.org/standard/83002.html

Jasanoff, S. (2004) States of Knowledge: The Co-Production of Science and Social Order. London: Routledge.

Li, Z. (2023) 'The dark side of ChatGPT: Legal and ethical challenges from stochastic parrots and hallucination', arXiv [cs.CY]. https://doi.org/10.48550/arXiv.2304.14347

Lovelace, A.A. (1843) 'Notes by the translator', in Menabrea, L.F., 'Sketch of the Analytical Engine invented by Charles Babbage Esq.', Scientific Memoirs, Selected from the Transactions of Foreign Academies of Science and Learned Societies, 3, pp. 666–731.

Mahieu, R. et al. (2021) 'Measuring the Brussels Effect through access requests: Has the European General Data Protection Regulation influenced the data protection rights of Canadian citizens?', Journal of Information Policy, 11, pp. 301–349. https://doi.org/10.5325/jinfopoli.11.2021.0301

Microsoft (2025) 'Secure AI by design series: Embedding security and governance across the AI lifecycle', Microsoft Defender for Cloud Blog, Microsoft Tech Community. Available at: https://techcommunity.microsoft.com/blog/microsoftdefendercloudblog/secure-ai-by-design-series-embedding-security-and-governance-across-the-ai-lifec/4457200; last accessed 18 March 2026.

Miller, A. and Piccone, T. (2015) 'No security without justice', Brookings, 18 June. Available at: https://www.brookings.edu/articles/no-security-without-justice/; last accessed last accessed 18 March 2026.

Burnham, K. (2025) 'This new framework helps companies build secure AI systems', MIT Sloan Management Review, 22 July. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/new-framework-helps-companies-build-secure-ai-systems; last accessed 18 March 2026.

MITRE (2023) ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. Available at: https://atlas.mitre.org; last accessed 29 March 2026.

Natale, S. and Henrickson, L. (2022) 'The Lovelace effect: Perceptions of creativity in machines', New Media and Society, 26(4), pp. 1909–1926. https://doi.org/10.1177/14614448221077278

NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1; last accessed 29 March 2026.

OECD (2024) Framework for anticipatory governance of emerging technologies. OECD Publishing. Available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/04/framework-for-anticipatory-governance-of-emerging-technologies_14bf0402/0248ead5-en.pdf; last accessed 18 March 2026.

OWASP CycloneDX (2024) Authoritative guide to attestations. OWASP Foundation. Available at: https://cyclonedx.org/guides/OWASP_CycloneDX-Authoritative-Guide-to-Attestations-en.pdf; last accessed 28 March 2026.

Tsekmezoglou, F. (2024) 'Getting started with the secure by design approach', Government Digital and Data Blog, 7 February. Available at: https://cddo.blog.gov.uk/2024/02/07/getting-started-with-the-secure-by-design-approach/; last accessed 18 March 2026.

Turing, A.M. (1950) 'Computing machinery and intelligence', Mind, 59(236), pp. 433–460.

Upmann, P. (n.d.) 'Governance by design: Embedding compliance and ethics in AI development', AI Governance Network. Available at: https://aign.global/ai-governance-insights/patrick-upmann/governance-by-design-embedding-compliance-and-ethics-in-ai-development/; last accessed 29 March 2026.

Bibliography

Cath, C. et al. (2018) 'Artificial intelligence and the "Good Society"', Science and Engineering Ethics, 24(2), pp. 505–528.

Cisco (2025) Framework foundations: Zero Trust models — CISA, DoD, and NIST solution brief, 20 November. Available at: https://www.cisco.com/c/en/us/products/collateral/security/zero-trust-cisa-dod-nist-sb.html

Daly, A., Devitt, S.K. and Mann, M. (eds.) (2019) Good Data. Theory on Demand, no. 29. Amsterdam: Institute of Network Cultures. Available at: https://networkcultures.org/blog/publication/tod-29-good-data/

Ebers, M. (2023) 'Regulating AI through risk-based accountability: The false promise of the EU AI Act', European Journal of Risk Regulation, 14(3), pp. 562–579.

Edwards, L. and Veale, M. (2017) 'Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for', LawArXiv. https://doi.org/10.31228/osf.io/97upg

Ganguli, D. et al. (2024) 'The EU AI Act: A risk-based approach to regulating artificial intelligence', International Journal of Law and Information Technology, 32(1), pp. 45–67.

Hagendorff, T. (2023) 'The ethics of AI ethics: An evaluation of guidelines', Minds and Machines, 33(2), pp. 213–235.

Knight, M. (2026) 'The 2026 guide to creating and maintaining your data governance policy', DATAVERSITY, 20 March. Available at: https://www.dataversity.net/articles/creating-a-data-governance-policy/; last accessed 18 March 2026.

Roberts, H. et al. (2022) 'Achieving a "good AI society": Comparing the aims and progress of the EU-28 and the US', AI and Society, 37(1), pp. 251–273.

Veale, M. and Zuiderveen Borgesius, F. (2022) 'Adtech and real-time bidding under European data protection law', German Law Journal, 23(2), pp. 226–256.

Wachter, S., Mittelstadt, B. and Russell, C. (2017) 'Counterfactual explanations without opening the black box', Harvard Journal of Law and Technology, 31(2), pp. 841–887.