THE DIGITAL ECONOMIST

Date of Publication | May 21, 2026

Expert Insights

Kapil Bareja and Imen Ameur

AI's Next Failure: Governance, Not Models

AI GOVERNANCE | SYSTEM RISK | CONTROL ARCHITECTURE

Table of Contents

Introduction

Most executives are asking the wrong question about AI risk. They ask whether the model is safe, accurate, or reliable enough to deploy. The more important question is whether the organization can govern the full AI system around that model once it is connected to enterprise data, third-party tools, customer interactions, and business decisions. That is where the real risk now sits.

AI is no longer being deployed as a standalone capability but increasingly embedded into workflows, copilots, and enterprise systems (Brynjolfsson et al. 2023; McKinsey Global Institute 2023). It is embedded into copilots, digital products, autonomous workflows, service channels, and decision-support systems. As soon as that happens, risk expands beyond the model itself to the broader ecosystem: data pipelines, retrieval layers, APIs, vector stores, plug-ins, vendors, human inputs, and machine-triggered actions. In that environment, a strong model is not enough. If the surrounding system is weak, the enterprise is exposed.

This is why the traditional split between AI safety and cybersecurity is becoming unworkable as generative AI systems introduce overlapping governance, safety, and security risks (Amodei et al. 2016; Brundage et al. 2018; Microsoft 2023). In practice, they are now the same executive problem. A prompt injection attack is both a cyber event and an AI control failure (Greshake et al. 2023; Perez & Ribeiro 2022; OWASP Foundation 2025). A hallucinated output that triggers a downstream action is both a reliability issue and a governance breakdown. A retrieval layer that exposes sensitive data is both a privacy failure and a security architecture flaw. Organizations that continue to manage these risks in separate lanes will create blind spots between them.

1.

From Model Risk to System Risk

The strategic shift is clear: AI governance must move beyond model oversight and become ecosystem governance. What remains underdeveloped in many enterprises is not awareness of AI risk but the ability to translate governance into system behavior. Too often, governance is still treated as an external layer of policies, external layer of policies and compliance processes rather than embedded operational controls (Kroll et al. 2017; Floridi et al. 2018), review boards, and compliance processes rather than embedded controls inside live workflows. This creates a structural gap: systems can act faster than governance can observe, interpret, or intervene.

That gap is not only a control problem. It is a performance problem tied to organizational productivity and technological integration challenges (Brynjolfsson, Rock, and Syverson 2017). When governance is weak, AI systems generate more exceptions, require more human review, and trigger more frequent overrides, all of which increase latency, rework, and operating cost. Poorly governed systems also allow errors to propagate farther across workflows before they are detected, reducing decision quality and making failures harder to contain. Organizations deploy AI to improve speed, improve productivity, and operational efficiency (Acemoglu and Restrepo 2020; Autor 2015; McKinsey Global Institute 2023). But when governance remains external to the system, those gains are diluted by manual escalation, inconsistent execution, compliance friction, and declining trust in automated outputs.

As a result, many enterprises are scaling capability faster than control. In practical terms, they are accelerating AI adoption while increasing the probability that failures will be harder to detect, attribute, and contain.

As AI-related incidents become more operational and interconnected, failures are less likely to appear as isolated model defects. They are more likely to emerge from interactions across data systems, tools, vendors, users, and downstream decisions. The implication is straightforward: risk now sits at the system level, not the model level. The next phase of AI deployment will depend less on model capability and more on governance architecture.

That means treating AI as a living socio-technical system rather than isolated software infrastructure (Leveson 2011; Perrow 1984; Reason 1990), not as another software feature. It means governing how models interact with data, tools, users, and third parties over time, not just how they perform in testing. It also means recognizing that the most serious failures will not come from the model in isolation but from how the full system is assembled, connected, and authorized.

2.

From Principles to Operational Governance

A practical governance foundation already exists. The National Institute of Standards and Technology (NIST) AI Risk Management Framework provides an enterprise governance structure for identifying and managing AI risks (NIST 2023) through four continuous functions: Govern, Map, Measure, and Manage. Its value is not only structural. It helps organizations define accountability, identify harms, test for real-world risk, and maintain oversight after deployment. For generative AI, that baseline needs to be specialized. Large language models introduce risks that traditional controls were not designed to handle: persuasive false outputs, responses to adversarial inputs, sensitive data leakage, and variable behavior when connected to retrieval systems or external tools, including hallucinations, adversarial prompting, and sensitive data leakage (OpenAI 2023; Anthropic 2023). These are not fringe issues. They are mainstream deployment realities.

That is why the NIST Generative AI Profile matters as organizations operationalize governance for generative systems (National Institute of Standards and Technology 2024). It helps translate broad AI governance principles into operational expectations for generative systems, including red teaming, incident readiness, misuse resistance, and controls for privacy and misinformation. The OWASP Top 10 for LLM Applications adds operational specificity for real-world deployment vulnerabilities (OWASP Foundation 2025) by identifying the failure modes that matter most in live deployments, from prompt injection and sensitive information disclosure to improper output handling, excessive agency, and vector database weaknesses.

But frameworks are guidance, not enforcement. They can clarify what good governance should look like, but they do not by themselves ensure that controls are operationalized in production systems. In practice, that translation into architecture is often the real bottleneck. Many organizations can articulate AI principles, risk categories, and ethical aspirations but struggle to operationalize them into enforceable controls (OECD 2019; The White House Office of Science and Technology Policy 2022; European Parliament & Council of the European Union 2024). Fewer have the technical capability, operating discipline, and cross-functional ownership required to embed those principles into system design, deployment gates, monitoring, and runtime controls.

That is the implementation gap executives should focus on. The challenge is no longer simply knowing which frameworks to reference. It is converting governance intent into enforceable technical and operational controls.

This is where the OWASP Top 10 for LLM Applications becomes valuable. It brings specificity to the risks that matter most in live deployments: prompt injection, sensitive information disclosure, supply chain vulnerabilities, data and model poisoning, improper output handling, excessive agency, vector database weaknesses, misinformation, and unbounded consumption. These are not technical footnotes. They are the control domains that should shape architecture decisions, deployment gates, vendor requirements, and operating procedures.

3.

Governing the AI Ecosystem

The implication for leadership is straightforward. AI governance should operate at three interconnected levels of organizational and institutional control (ISO/IEC 2023; Kroll et al. 2017).

First, enterprise governance: who owns AI risk, who sets risk appetite, and who has the authority to approve, restrict, or retire systems. Without clear decision rights, AI risk becomes everyone's concern and no one's responsibility.

Second, lifecycle governance: what evidence is required before deployment, how systems are monitored in production, and when they must be reassessed. AI should not move from pilot to scale on the basis of enthusiasm alone.

Third, ecosystem governance: how model providers, data suppliers, tool vendors, and integration partners are brought inside the control perimeter. In most enterprise deployments, the point of failure will sit somewhere in that chain.

This also requires more disciplined risk tiering. Not every AI use case deserves the same governance. A low-risk internal assistant should not be treated like an agentic workflow with access to enterprise systems or decision authority in customer-facing processes. As autonomy increases, governance should tighten. Systems with higher agency should be governed more like critical

infrastructure, more like high-risk socio-technical infrastructure systems (Perrow 1984; Leveson 2011) or productivity tools. For executives, the near-term agenda is practical. Establish a single AI governance model that integrates safety and cybersecurity. Define risk tiers based on impact, exposure, and autonomy. Require evidence-based lifecycle gates. Institutionalize red teaming for prompt injection, data leakage, tool misuse, and retrieval poisoning. Treat model outputs as untrusted unless validated. Constrain tool permissions and agent autonomy. Extend contractual and monitoring requirements across the vendor ecosystem.

For this shift to work, governance cannot remain only a policy overlay detached from operational architecture (Floridi et al. 2018; Mittelstadt et al. 2016). It must be designed directly into production systems. That means controlled data access, validation layers for model outputs, constrained tool permissions, continuous monitoring, and clear mechanisms for human override. Outputs should be treated as untrusted by default unless verified, especially when they trigger downstream actions, decisions, or customer-facing responses.

As AI systems move from assistive tools to agentic workflows, governance must scale with autonomy. Systems with greater decision authority should operate under stricter controls: narrower permissions, stronger authentication, explicit escalation paths, auditability, and enforced checkpoints for human intervention. Governance, in this context, is not a documentation exercise. It is a system capability that determines whether AI can be trusted at scale.

Conclusion

Organizations that do this well will gain more than risk reduction. They will move faster with greater confidence because they will know which systems they can trust, under what conditions, and with what controls. That is the real competitive advantage.

The next major AI failure in most enterprises is less likely to happen because the model was imperfect than because leadership allowed an AI ecosystem to scale faster than it could be governed, reflecting broader systemic governance and institutional coordination failures rather than isolated technical defects (Rittel & Webber 1973; Reason 1990).

The winners will not be the organizations with the most AI. They will be the ones with the most control over how AI is deployed, connected, monitored, and contained.

References

1. Acemoglu, D., and Restrepo, P. 2020. "Artificial Intelligence, Automation, and Work." *Journal of Economic Perspectives* 34 (1): 3–30. https://doi.org/10.1257/jep.34.1.3.
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. 2016. "Concrete Problems in AI Safety." arXiv. https://arxiv.org/abs/1606.06565.
3. Anthropic. 2023. "Constitutional AI: Harmlessness from AI feedback." https://arxiv.org/abs/2212.08073.
4. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., and Roberts, K. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology, 2024.
5. Autor, D. H. 2015. Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives* 29 (3): 3–30. https://doi.org/10.1257/jep.29.3.3.
6. Brundage, M., Avin, S., Clark, J., et al. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *Future of Humanity Institute*.
7. Brynjolfsson, E., Li, D., and Raymond, L. R. 2023. "Generative AI at Work" (NBER Working Paper).
8. Brynjolfsson, E., Rock, D., and Syverson, C. 2017. "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics." NBER Working Paper No. 24001. https://doi.org/10.3386/w24001.
9. European Parliament and Council of the European Union. 2024. "Regulation (EU) 2024/...Laying Down Harmonised Rules on Artificial Intelligence" (Artificial Intelligence Act).
10. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. 2018. "AI4People—An Ethical Framework for a Good AI society." *Minds and Machines* 28, 689–707. https://doi.org/10.1007/s11023-018-9482-5.
11. Greshake, K., et al. 2023. "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Prompt Injection." arXiv. https://arxiv.org/abs/2302.12173.
12. ISO/IEC. (2023). ISO/IEC 42001: Artificial intelligence—Management System.

13. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. 2017. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (3): 633–705.
14. Leveson, N. 2011. "Engineering a Safer World: Systems Thinking Applied to Safety." MIT Press.
15. McKinsey Global Institute. 2023. "The Economic Potential of Generative AI: The Next Productivity Frontier."
16. Microsoft. 2023. "Securing AI Systems: Best Practices for Generative AI Applications."
17. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2). https://doi.org/10.1177/2053951716679679.
18. National Institute of Standards and Technology. 2023. "Artificial Intelligence Risk Management Framework [AI RMF 1.0][NIST AI 100-1]." Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-1.
19. National Institute of Standards and Technology. 2024. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile [NIST AI 600-1]."
20. OECD. 2019. "OECD Principles on Artificial Intelligence." https://oecd.ai/en/ai-principles.
21. OpenAI. 2023. "GPT-4 System Card." https://cdn.openai.com/papers/gpt-4-system-card.pdf.
22. OWASP Foundation. 2025. "OWASP Top 10 for Large Language Model Applications."
23. Perez, F., and Ribeiro, I. 2022. "Ignore Previous Prompt: Attack Techniques for Language Models." arXiv. https://arxiv.org/abs/2211.09527.
24. Perrow, C. 1984. Normal Accidents: *Living with High-Risk Technologies*. Princeton University Press.
25. Reason, J. 1990. Human Error. Cambridge University Press.
26. Rittel, H. W. J., and Webber, M. M. 1973. "Dilemmas in a General Theory of Planning." *Policy Sciences* 4 (2): 155–169. https://doi.org/10.1007/BF01405730.
27. The White House Office of Science and Technology Policy. 2022. "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People."

Author

Kapil Bareja

Kapil Bareja is a cybersecurity and risk leader with more than two decades of experience operating at the intersection of information security, governance, and institutional strategy. As Cyber & Strategic Risk Consulting Leader at Deloitte, he advises boards and executive leadership on cybersecurity transformation, identity risk, and enterprise resilience across highly regulated sectors, including government, financial services, healthcare, and higher education. A recipient of the Cyber Under 40 award, he is recognized for translating complex cyber risk into actionable, board-level decision frameworks. His contributions to global cybersecurity and privacy discourse include leadership roles with the International Association of Privacy Professionals and IEEE, as well as advisory engagements with Harvard Business Review and MIT Sloan School of Management. He is also a member of the Forbes Technology Council, where he contributes perspectives on cybersecurity, risk, and digital trust. As a Senior Executive Fellow in Tech Policy and Governance and Cyber Studio at The Digital Economist, he brings a practitioner's perspective to systems-level questions of cybersecurity governance, institutional trust, and responsible technology adoption, bridging technical security leadership with policy development and executive accountability.

Co-Author

Imen Ameur

Imen Ameur is a technology governance and policy expert specializing in AI systems, healthcare innovation, and institutional transformation. She is the founder of ATL Engines and IA Consulting, where she designs governance-native architectures, automated systems, and agentic AI solutions that embed accountability, ethics, and decision traceability directly into operational workflows. She serves as Senior Executive Fellow and Co-Chair of Tech Policy and Governance at The Digital Economist, contributing to global efforts shaping responsible AI adoption, technology policy, and implementation frameworks. Imen is Professor of Practice at Hult International Business School, where she focuses on emerging technologies, AI governance, and entrepreneurship. She has taught at Columbia University and conducted research within leading academic and policy environments, including Harvard Kennedy School. Her work centers on healthcare AI, productivity systems, and institutional governance, with a focus on embedding oversight and accountability directly into technology systems. She also serves as Vice President of Research and Development at the Africa Digital Cluster Think Tank, advancing applied research on digital transformation, AI governance, and inclusive economic systems.

About

The Digital Economist, headquartered in Washington, D.C. with offices at One World Trade Center in New York City, is the world's foremost think tank on innovation advancing a human-centered global economy through technology, policy, and systems change. We are an ecosystem of 40,000+ executives and senior leaders dedicated to creating the future we want to see—where digital technologies serve humanity and life.

We work closely with governments and multi-stakeholder organizations to change the game: how we create and measure value. With a clear focus on high-impact projects, we serve as partners of key global players in co-building the future through scientific research, strategic advisory, and venture build out.

We engage a global network to drive transformation across climate, finance, governance, and global development. Our practice areas include applied AI, sustainability, blockchain and digital assets, policy, governance, and healthcare. Publishing 75+ in-depth research papers annually, we operate at the intersection of emerging technologies, policy, and economic systems—supported by an up-and-coming venture studio focused on applying scientific research to today's most pressing socio-economic challenges.

CONTACT: INFO@THEDIGITALECONOMIST.COM

Our Initiatives

Center of Excellence on Human-Centered Global Economy

The Center of Excellence convenes leaders across technology, policy, and industry to address critical global challenges and shape the systems, standards, and governance of the digital economy through research, convening, and applied initiatives.

Executive Fellowship

A selective leadership program integrating senior professionals into the Center of Excellence to advance global policy dialogue, systems innovation, and cross-sector collaboration.

Participation

- 24-hour annual commitment
- Contribution to two publications annually
- Access to 500+ global convenings (World Bank / IMF Spring Meetings, UNGA, Davos Week, The Digital Economist Virtual Summits)
- Speaking opportunities across major policy and innovation forums

Thematic Workgroups

- Applied Artificial Intelligence
- Digital Assets & Blockchain
- Sustainability in Technology
- Tech Policy & Governance
- Quantum Computing
- Cyber Studio
- Healthcare Innovation

Institutional Research Network

A cross-sector research platform convening institutions at the intersection of markets, policy, academia, and capital to shape the frameworks guiding the next phase of the global digital economy. Through structured collaboration, participating institutions co-author high-signal research and engage in executive roundtables and policy dialogues that inform governance design, regulatory development, and long-term economic architecture.

The Digital Economist Ventures

Applied platforms translating research into real-world implementation and governance innovation.

- **Tech for Transparency:** Financial integrity and anti-corruption frameworks in the digital economy
- **The Ostrom Project:** Governance models for digital commons and shared infrastructure
- **ANER-G:** Decentralized energy infrastructure and next-generation energy systems
- **Africa Coalition:** Continental coordination across strategic sectors, including infrastructure, finance, and technology

Collaborate with The Digital Economist

Advance leadership and institutional collaboration shaping technology, policy, and economic systems transformation.

partnerships@thedigitaleconomist.com
thedigitaleconomist.com

Scan the QR code to explore participation pathways.


THE
DIGITAL
ECONOMIST