Scaling AI with Adaptive Governance
Deck: Maximizing the value of AI investments while managing new and emerging risks requires embedding flexible oversight processes across the technology lifecycle.

By Gianvito Lanzolla, Margherita Pagani, and Christopher L. Tucci

About the authors:

Gianvito Lanzolla is a professor of strategy at Bayes Business School, City St George's, University of London.
Margherita Pagani is a professor of AI for Business at SKEMA Business School and Université Côte d'Azur and director of the SKEMA Center for Artificial Intelligence (SKEMA Business School)
Christopher L. Tucci is a professor of digital strategy & innovation at Imperial Business School, Imperial College London.

Author version 06 February 2026, edited 13 February 2026

Leaders with even a cursory understanding of AI know that while this powerful technology can help them to improve productivity and capture new opportunities, it can also expose their organizations to many risks. Those who know a bit more are aware that surfacing and mitigating those risks requires adopting responsible AI practices. And for those scaling AI implementation at their organization, it will become obvious that ad hoc attention to those practices is inadequate, and they will require a systematic capacity to govern AI at scale.

Yet building that capacity is proving far harder than most executives expect. They know what they need to accomplish: frameworks from governments and regulators[1] define important guardrails and principles such as transparency, fairness, and accountability. But to implement controls and principles into day-to-day workflows and decision-making, organizations must rethink AI governance. They must frame that task not as a compliance obligation, but as a strategic, adaptive capability that evolves as AI systems scale, use cases expand, and risks shift over time.

In this article, we show how leading organizations are doing exactly that. We introduce an approach to adaptive AI governance built on two principles: matching governance controls to the type of AI system and risk involved, and embedding those controls directly into workflows, decision rights, and accountability structures.

The fundamentals of AI risk

To design effective AI governance, leaders must first understand the multiple ways in which AI can fail and the resulting risks. The nature and severity of these risks depend on the type of system, its level of autonomy and the scope of domains affected by its decisions. The central challenge, therefore, is to design controls that anticipate how risks emerge and that evolve as AI systems operate in practice. Even as conditions, inputs, and expectations change, AI must remain reliable, safe, and aligned with the organization's values and goals.

In practice, most AI risks emerge at two moments that require very different governance responses: during development and after deployment**.** Development risks include using

biased or incomplete training data; failing to adequately align the model to the task requirements; and following inadequate validation processes. For example, an early credit-limit increase model at a bank we studied showed that small input changes could lead to unexpected decision shifts.

Deployment risks arise when models interact with dynamic environments and human operators: sustaining legitimacy, judgment, and accountability once AI systems operate at scale in real time is a central challenge. Over time, model quality may degrade as the statistical properties of input data change over time, a phenomenon termed data drift. A model may generate plausible but false outputs, or be overly trusted by users who lack the means to detect errors. At Nasdaq, AI-driven market-surveillance systems monitor trading activity for suspicious patterns, generating hundreds of alerts per second, but may fail to accurately flag activity because the boundary between abnormal and illicit behavior is often hard to spot in practice; illegitimate behavior may be deliberately shaped to pass as compliant by exploiting model learning patterns.

Fit-for-Purpose Controls

The kinds of controls employed depend not only on when risks arise in the AI lifecycle, but also on what kind of AI system is involved and how widely its decisions propagate. Artificial intelligence systems can be broadly divided into two categories: those based on bounded-learning (or static) models and those that learn and adapt in deployment. (See figure 1.)

Bounded-learning systems operate within a fixed set of rules and parameters, improving performance by optimizing how those rules are applied rather than changing them. Credit-scoring models, for example, refine risk estimates based on income or payment history, but they do not alter how those variables relate to one another. Many generative AI models are "pre-trained" (static) and do not update during use. Adaptive learning systems, by contrast, evolve as production data extends their training data, and by updating their internal representations and relationships between variables. Algorithmic trading platforms and dynamic fraud-detection systems illustrate this approach.

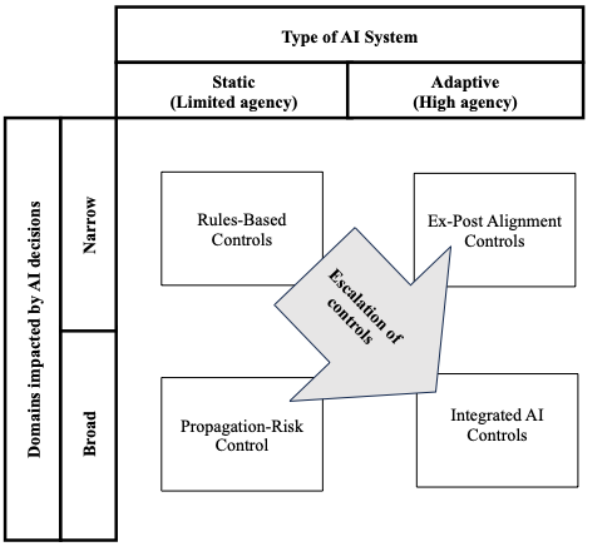
Just as salient to the type of control required is the scope of domains affected by AI decisions, shown on the vertical axis of Figure 1. This dimension determines how far and how fast risks can travel once a system goes wrong. At one extreme are narrow-scope systems, where errors remain contained within a specific function or task (for example, anomaly detection within a single transaction stream). At the other are wide-scope systems that shape outcomes across multiple functions, geographies, or even industries, such as cross-border supply-chain optimization platforms. The difference is not incremental but exponential: as system reach expands, small errors interact, propagate, and amplify into second-order effects.

Based on our typology of AI systems, we argue that rules-based controls provide the baseline safeguards for all narrow, static AI systems. When such static systems operate at a wider scope, additional propagation risk controls must be layered on to address broader downstream effects.

For adaptive learning systems, baseline safeguards remain necessary, but they must be complemented by ex-post alignment controls, particularly those focused on explainability and legitimacy. When adaptive systems also have a wide scope, they require the most comprehensive approach: integrated controls that combine baseline rules-based measures

with propagation risk management and alignment mechanisms. Let's take a closer look at how each of these works in practice.

Figure 1: Types of controls for adaptive AI governance



Rules-Based Controls

Rules-based controls are designed to prevent and correct errors in systems that operate within clearly defined parameters. They are particularly effective in narrow decision domains where logic is explicit and outcomes auditable, such as credit scoring, fraud detection, or customer service chatbots. Rules-based controls embed relevant norms (such as ethical guidelines and industry standards) and compliance requirements into models, using these as design constraints. Rules-based controls also include processes such as validation tests or anomaly monitoring.

Consider the example of a bank's credit-limit increase decision model. A senior AI leader at the bank explained that it uses a statistical model rather than deep learning so that decisions remain interpretable. Before deployment, the analytics team produces documentation called a model card that covers three aspects of AI risk management. Data checks indicate whether the training data is complete, recent, and balanced, and how the team will detect data drift over time. Decision logic and edge cases are checked to see how scores translate into approve/don’t approve decisions, including explicit analysis of thresholds where a customer tips from no increase to an increase, so that customers in the “grey zone” are not unfairly treated. Finally, bias and discrimination tests are undertaken to check that the model does not overfit to particular customer profiles or systematically disadvantage certain groups.

This documentation undergoes quality-assurance review by an independent model-risk unit, with input from credit-domain and regulatory experts. Internal audit later verifies that these steps were followed. Only then does the model go live.

Human judgment is central even in rules-based settings. In one organization, , each new lending model for mid-market clients underwent sample testing before deployment. Risk teams selected 100 existing client files and ran them through the model. Relationship managers then compared the model's recommended lending decisions with their own assessments. Where recommendations diverged, the model team investigated whether the model had uncovered a genuine insight or was overfitting to idiosyncrasies in the data. Only once the sample review showed an acceptable level of alignment between model outputs and the judgments of the domain experts involved —and the sources of disagreement were understood — did the bank approve the model for live use. After launch, periodic sample reviews continued as part of the standard risk-and-control cycle.

Rules-based controls are effective because they make critical decision boundaries explicit, reviewable, and contestable across domains. They are adaptive because they can be recalibrated over time. Divergences between model outputs and expert judgment are treated as learning signals, feeding back into updated model thresholds, assumptions, and review routines as data, models, and decision contexts evolve.

Ex-Post Alignment

The complexity of advanced AI systems, particularly those based on deep neural networks, render traditional traceability and explainability methods less effective. Rules-based controls depend on the ability to specify decision logic, yet that logic becomes increasingly opaque as models grow more complex. As a result, stakeholders must ensure that outcomes remain reliable, fair, and aligned with organizational and regulatory expectations. When such systems operate with significant autonomy, this need for explainability becomes especially critical, since decisions may be made and acted upon without immediate human review. Generative AI introduces an additional layer of difficulty because of its stochastic behavior, where the same "prompt" may yield different outputs.

This is where ex-post alignment controls come in. They reveal not how a decision was made, but whether its outcomes remain legitimate. They assess AI decisions against ethical, regulatory, and domain-specific standards. While some techniques carry over from rules-based approaches, the emphasis shifts from preventing errors upfront to detecting misalignment as systems operate, learn, and scale.

Organizations operationalize ex-post alignment through layered evaluation processes that test outcomes against reference expectations. Microsoft, for instance, has developed a structured evaluation pipeline in which high-stakes models are assessed against libraries of expert-defined policies—such as what constitutes a "fair" or an "acceptable" outcome. Evaluators annotate model outputs against these policies, while independent reviewers validate where the system falls short. In some cases, these evaluations can be partially automated—for example, when AI systems are continuously assessed against predefined policy benchmarks, fairness constraints, or risk thresholds, with automated monitors flagging deviations for human review.

This is why algorithmic auditing is a critical component of ex-post alignment. Audits systematically examine a model's behavior after deployment to detect hidden risks, evaluate fairness and performance across affected groups, and verify that outcomes align with organizational policy and ethical standards. In practice, auditing proceeds in two steps. First, auditors identify plausible failure scenarios and define the full use case, including who the

system serves, who is affected by its decisions, and for what purpose it operates. They then monitor these risks by assessing decision outputs, input data, and internal logic against predefined criteria. This process helps organizations surface unintended consequences, such as disparate impacts, document recurring risk patterns, and trigger corrective action before harm proliferates. Frameworks for auditing algorithmic risk, such as those articulated in Cathy O'Neil's work on auditing AI systems,[2] provide practical tools and metrics to operationalize this approach and strengthen accountability. In this way, auditing functions both as a diagnostic mechanism and as a foundation for continuous improvement.

A key part of ex-post alignment is ensuring that people do not treat AI outputs as unquestionable truths. Because many AI recommendations are inherently probabilistic, organizations need to train users to interpret those recommendations as informed signals rather than final decisions. Helping managers understand when to rely on the system, when to challenge it, and how to spot unexpected or biased outputs is essential for keeping AI use legitimate, accountable, and aligned with organizational values over time.

Managing misalignment at scale can be a particular challenge, especially for systems that are designed to filter, prioritize, and escalate alerts in real time. Nasdaq's AI-driven market surveillance, for example, monitors trading activity for irregularities—such as unusual volumes, price anomalies, or potential manipulation—and can generate hundreds of high-risk alerts per second. Cross-functional teams of compliance officers, data scientists, and domain experts review flagged activity through structured case workflows. Each alert is assessed to determine whether it reflects genuine market manipulation or a false positive triggered by unusual but legitimate trading behavior. Investigators document the rationale for their conclusions, and these outcomes are fed back to model developers to recalibrate thresholds, refine detection features, and reduce recurring noise in future alerts.

Escalation committees intervene when investigations suggest coordinated bad actors or when anomalies indicate broader systemic risk. Audit trails capture key elements of this process, including the original alert, supporting data signals, the human decision taken, and any subsequent model adjustments. Periodic governance reviews then evaluate patterns in false positives and missed detections to ensure accountability, regulatory compliance, and continuous improvement of surveillance rules. Even so, surges in alert volumes can place severe strain on teams, overwhelming response capacity and increasing the risk of error.

One effective response is to redesign workflows around AI outputs. This approach is illustrated by a global bank's experience with AI-driven fraud detection. Executives found that the main challenges did not stem from errors in the model predictions, but from breakdowns in how fraud alerts were interpreted, routed, and acted upon across teams. Inconsistent handoffs between compliance, risk, and frontline staff often led to delayed responses, duplicated effort, or missed follow-up, undermining the system's effectiveness despite technically sound outputs. Fraud analysts, fraud-operations staff, data scientists, and customer-service teams all worked from the same alert system, but errors often emerged at the handoffs between them. For example, alerts were sometimes routed to the wrong team, duplicated across units, or left unresolved because no group clearly owned the next step. Customer-service staff occasionally contacted clients based on alerts that fraud teams had not yet validated, while high-risk cases were delayed because escalation criteria were unclear.

To address this, the bank mapped the alert workflow step by step and reassigned responsibilities at each decision point. Fraud analysts were given clearer authority to close

low-confidence alerts, fraud operations focused on rapid escalation of confirmed cases, and customer-service teams were engaged after fraud review determined that outreach was necessary. Decision rules were standardized—for instance, specifying when an alert should be suppressed, investigated further, or escalated—reducing delays, mis-escalations, and alert overload.

Ex-post alignment focuses on evaluating AI decisions after they are made, testing outcomes against ethical, regulatory, and domain-specific expectations rather than reconstructing internal decision logic. Ultimately, successful ex-post alignment does not eliminate risk; it sustains legitimacy by ensuring that high-agency AI outcomes remain contestable, correctable, and aligned with the standards that matter over time. Unlike traditional risk management, ex-post alignment accepts that some misalignment is inevitable—and focuses governance on detection, contestability, and correction rather than prevention alone.

Propagation-Risk Controls

Rules-based controls and ex-post alignment mechanisms share an important limitation: they tend to treat risk as largely confined, focusing on discrete errors or individual outputs. This approach can be effective when AI systems operate in relative isolation, but it produces incomplete outcomes once systems are interconnected through real-time data flows, APIs, and automated decision-making. The rise of agentic AI is a case in point. As AI systems increasingly initiate actions autonomously, coordinate with other systems, and pursue objectives across multiple domains, errors or misalignments originating in one system can propagate across others rather than remain contained. The relevant concern, therefore, is interdependence and propagation risk, creating downstream effects that traditional, output-focused controls may overlook.

Regulators are increasingly recognizing the importance of propagation risk and the need for robust testing and oversight. The Bank of England, for example, has highlighted the risks posed by “deep trading agents,” AI-driven strategies that could amplify external shocks or coordinate in ways that evade human detection. In healthcare, biased diagnostic models can spread flawed heuristics across hospitals and insurers. In supply chains, algorithmic procurement platforms can amplify pricing errors across entire supplier networks. Similar dynamics arise in any digitally interconnected system.

Propagation-risk controls represent a third layer of governance, designed to surface second- and higher-order effects before they overwhelm downstream functions. In our framework, rules-based controls safeguard narrow and relatively static processes; alignment mechanisms address complex systems whose decisions are opaque; and propagation controls focus on interconnected systems. These controls are concerned not only with what happens within a system, but with what occurs when systems interact. Their central challenge is invisibility: failures travel laterally, exploiting hidden interdependencies that often become apparent only once disruption occurs. A minor logistics API error, for example, may be harmless in isolation, but when combined with a cyber incident affecting a payment gateway, it can contribute to systemic breakdown.

A governance framework built around a firm-centric view of risk is poorly suited to track such cross-boundary dynamics. Because propagation risks unfold across interconnected systems, often beyond the visibility or control of any single organization, managing them requires a shift from a firm-centric to an ecosystem-aware perspective.

In practice, this shift involves three complementary activities: mapping interdependencies, monitoring shared infrastructures, and institutionalizing anticipatory oversight. Together, these practices help surface risks that remain invisible when controls focus only on isolated systems or individual outputs. The European Central Bank's sector-wide cyber-resilience stress tests shows how ecosystem-level propagation-risk controls can be enacted in practice. These exercises map interdependencies across clearing houses, payment systems, and financial institutions; monitor shared infrastructures for cross-firm vulnerabilities; and simulate how localized disruptions could cascade through the financial system. These practices generalize beyond regulation to any highly interconnected environment.

Organizations can enact propagation-risk controls by redistributing visibility, accountability, and decision rights across ecosystems, rather than relying solely on firm-level rules or ex-post interventions. Because propagation risks are inherently cross-boundary, effective governance depends as much on coordination across organizations as on internal control. For some organizations, it can require a shift in cultural norms to encourage data sharing, coordination on standards, and co-investment in oversight infrastructures with partners, competitors, regulators, and, in some cases, open-source communities.[3] It requires an understanding that resilience is no longer something a single firm can achieve on its own, but is a property of the system the firm depends on.

As ecosystems become more densely interconnected, these risks are likely to intensify. The rise of agentic AI—capable of autonomously initiating transactions, negotiating contracts, or reallocating resources across networks—extends this logic, increasing both the speed and reach with which failures can propagate. In finance, logistics, and healthcare alike, errors may not simply spread; they may increasingly do so with limited human intervention.

Implementing adaptive AI governance

Once leaders have identified the AI risks that are salient to their organizations and the corresponding controls that they need to have in place, the challenge becomes integrating those controls into processes and systems, working within them, and continuously adapting them. Doing so involves three key practices: embedding controls into workflows and incentives, building cross-domain fluency, and institutionalizing governance as a living learning system. Here is how to do that.

1. Embed risk control protocols into operations

Risk protocols must be designed and hard-wired into workflows, accountability structures, and incentives. Oversight should flow directly into planning, audits, and leadership reviews—not sit as a separate compliance layer. Only when governance becomes part of the operating fabric can AI be scaled with confidence. This is a necessary condition.

A global bank we interviewed embedded AI controls into its standard lending workflow rather than treating them as a separate compliance step. For each approved AI use case, the bank’s AI use-case committee documented: (1) the risk tier (high, medium, low) based on customer impact, regulatory impact, data sensitivity, and model type; (2) the mandatory controls associated with that tier (for example, independent model validation, sample testing by relationship managers, frequency of post-deployment reviews); and (3) the decision rights

(who could approve model changes and under what conditions). These requirements were then encoded directly into the credit-approval process and systems. Relationship managers could not bypass model-validation steps or deployment reviews; exceptions required explicit sign-off from both the business and risk. Oversight surfaced in regular decision-making cycles, not as ad-hoc committees or audits.

2. Enable conclusive judgment across heterogeneous expertise and risk profiles

Adaptive AI governance does not require consensus or shared judgment. Quite the opposite: it requires mechanisms that enable conclusive judgment across heterogeneous expertise, methods, and risk profiles. This is often the hardest—and most decisive—step. As AI risks shift across categories and cut through organizational silos, accountability cannot reside within any single function. Differences across domains are not a flaw but a feature: they reflect distinct expertise, evaluative methods, and risk tolerances. The governance challenge is therefore not to homogenize these perspectives, but to create the conditions under which organizations can translate them into conclusive decisions at scale—while avoiding both judgment homogenization and uncritical rubber-stamping of AI outputs. In the following, we outline the central challenges organizations face—and the responses required—to enable this form of conclusive judgment and institutionalize it as a durable governance capability.

Rules-based controls are often undermined by siloed knowledge, when various domain experts don't share a common frame. To overcome those barriers, share knowledge across domains via joint model reviews, documentation such as the model cards described earlier, and routine cross-functional validation sessions that make decision logic, assumptions, and thresholds explicit and contestable. In ex-post alignment controls, the challenge is not only one of knowledge silos, but also of misaligned risk tolerance and methodological approaches. Alignment can break down when different teams operate with different implicit risk thresholds—stopping judgment too early on the one hand or falling into analysis paralysis on the other—and when they rely on divergent methods to reconcile expected outcomes with observed results (e.g., analytical validation, controlled experiments, or case-based judgment). In such settings, disagreement is not simply about what the model recommends, but about how much risk is acceptable and what constitutes sufficient evidence that the model is performing as intended.

A critical response, therefore, is not merely to “build trust” in AI recommendations, but to establish shared evaluative routines that surface and reconcile differences in both risk tolerance and methodological approach. Systematic post-deployment evaluation anchors discussion in observed system behavior rather than abstract beliefs about model quality.

Organizations can do this through structured review routines that combine incident and near-miss analysis, performance-drift monitoring, and explicit comparison between intended use cases and actual decision outcomes. Crucially, these routines create common reference points—agreed risk thresholds, shared evidentiary standards, and comparable metrics—through which analytically driven teams, experimentation-oriented groups, and use-case owners can jointly assess whether the model is functioning as intended. Over time, this enables assumptions, thresholds, and controls to be recalibrated, reducing both premature shutdowns driven by excessive caution and analysis paralysis driven by methodological disagreement.

Propagation risk control depends on a fundamental shift in mindset: from treating risk as a firm-centric problem to governing it as an ecosystem-level phenomenon. As with digital business ecosystems,[4] risks in AI systems propagate unevenly across actors with different roles, incentives, and degrees of interdependence. Mapping these interdependencies beyond firm boundaries is a necessary first step—and often a wake-up call—but it is insufficient on its own.

As research on ecosystem strategy shows, coordination breaks down when accountability is diffuse, incentives remain locally optimized, and no actor is explicitly responsible for orchestrating cross-boundary tradeoffs.[5] Similar dynamics undermine AI propagation-risk controls. Teams remain incentivized to focus narrowly on their own systems, risk ownership is fragmented across organizational units and external partners, and downstream or reputational risks are treated as someone else's responsibility.

Without leadership support for ecosystem-level accountability—and governance mechanisms that differentiate risk ownership by type of interdependence—interdependency mapping risks becoming a one-off analytical exercise rather than a sustained governance capability. An ecosystem mindset requires not only visibility into connections, but also shared rules of engagement, escalation rights, and decision authority to manage how risks propagate across organizational and technological boundaries over time.

Overcoming these barriers is essential to create the conditions for conclusive judgment that respects differences in expertise, methods, and risk tolerance, rather than collapsing them into a single acritical evaluative frame.

3. Institutionalize governance as a learning system

AI governance cannot be static: risks mutate, so controls must evolve.[6] Effective governance therefore requires organizations to establish learning loops with clear roles for capturing lessons from incidents and near-misses, and for translating those lessons into updated standards, thresholds, and controls.

Rather than relying solely on controls, systems, or large-scale governance platforms, effective adaptive AI governance depends on building the right mindset and embedding practical learning loops into everyday oversight. This involves assigning explicit responsibility for reviewing incidents and near-misses, systematically documenting what went wrong, and ensuring that insights are translated into revised policies, recalibrated thresholds, or strengthened controls. Over time, governance shifts from protocols and systems toward institutionalized continuous improvement—ensuring that AI systems remain aligned with organizational intent as models evolve, contexts shift, and new risks emerge.

<Conclusion>

Taken together, these steps mark a fundamental shift. Adaptive AI governance is not about multiplying controls, committees, or checklists. It is about identifying fit-for-purpose controls and hard-wiring them into how the organization works, decides, and learns—into workflows and incentives, into shared frames of judgment, and into living systems that continuously absorb and act on experience.

Organizations that treat governance as static will inevitably fall behind systems that learn, adapt, and propagate risk in real time. By contrast, firms that institutionalize governance as a learning capability—one that connects strategy, execution, and oversight—can turn AI governance from a constraint into an enabler of scale. In the age of intelligent systems, advantage will come not from adopting AI faster, but from governing it better—by embedding oversight where decisions are made, risks propagate, and value is created.

****

The Research

Between 2022 and 2025, the authors conducted in-depth, semi-structured interviews with senior leaders and practitioners responsible for AI governance, risk, compliance, data, and product decisions.

- Core interviews were conducted at Microsoft, Barclays, Kyriba, Nasdaq, Lloyds, Danske Bank, and the Abu Dhabi Finance Department. Interviews focused on how governance works in practice: where it breaks down, how controls are enacted, and what organizational trade-offs leaders face as AI systems scale.
- The authors collected additional evidence on AI governance at more than 40 other financial institutions by drawing on public disclosures, regulatory filings, and practitioner documentation. These additional cases were used to validate the generalizability of consistent themes emerging from core interviews.

---

[1] For example, 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. https://doi.org/10.6028/NIST.AI.100-1.

[2] O'Neil, C. (2018). Audit the algorithms that are ruling our lives. *Financial Times*, 30 July. https://www.ft.com/content/879d96d6-93db-11e8-95f8-8640db9060a7. See also O'Neil, C., Sargeant, H., & Appel, J. (2024). Explainable fairness in regulatory algorithmic auditing. *West Virginia Law Review*, 127(1), 79-133.

[3] Pagani M. & Davenport T. (2025). How AI changes partner collaboration. MIT Sloan Management Review. https://sloanreview.mit.edu/article/how-ai-changes-partner-collaboration/. See also Pereira, J. R., Viscusi, G., & Tucci, C. L. (2026). Managing digital crowds for generativity: The role of scalability and forking. *Strategic Management Review*. https://strategicmanagementreview.net/assets/articles/Pereira,%20Viscusi,%20and%20Tucci.pdf.

[4] Staudenmayer, N., Tripsas, M., & Tucci, C. L. (2005). Interfirm modularity and its implications for product development. *Journal of Product Innovation Management*, 22(4), 303-321.

[5] Lanzolla, G., & Markides, C. (2022). How to choose the right ecosystem partners for your business. *Harvard Business Review*. https://hbr.org/2022/03/how-to-choose-the-right-ecosystem-partners-for-your-business. Lanzolla, G., & Markides, C. (2025). *Diversification in the World of Data and AI*. Cambridge: Cambridge University Press.

[6] Pagani, M., & El Sawy, O. A. (2026). Challenges of boundaryless metaverses: implications for IS research. *European Journal of Information Systems*, Forthcoming. https://doi.org/10.1080/0960085X.2025.2612487.