sustainability

MDPI

Article

Educating Managers to Govern Artificial Intelligence

Viacheslav Osadchyi [1], Anton Shantyr [2,3,*], Olha Zinchenko [3], Andrii Bondarchuk [1], Nataliia Lashchevska [3] and Kateryna Osadcha [4,5]

[1] Faculty of Economics and Management, Borys Grinchenko Kyiv Metropolitan University, 04053 Kyiv, Ukraine; v.osadchyi@kubg.edu.ua (V.O.); a.bondarchuk@kubg.edu.ua (A.B.)
[2] Leviathan Security Group, Tukwila, WA 98188, USA
[3] Educational and Scientific Institute of Information Technologies, State University of Information and Communication Technologies, 03110 Kyiv, Ukraine; o.zinchenko@duikt.edu.ua (O.Z.); n.lashchevska@duikt.edu.ua (N.L.)
[4] Vrije Universiteit Brussel, 1050 Brussels, Belgium; k.osadcha@iitlt.gov.ua
[5] Institute for Digitalisation of Education of the National Academy of Educational Sciences of Ukraine, 04060 Kyiv, Ukraine
* Correspondence: anton.shantyr@gmail.com

Abstract

Artificial intelligence (AI)-related harms are increasingly attributed to governance failures rather than to isolated technical malfunctions. This article reframes AI governance as a core managerial competence grounded in leadership authority, accountability design, and organizational communication. The study addresses a persistent gap in higher education and managerial training, namely the insufficient preparation of future leaders to govern AI-mediated decision systems responsibly. Using a structured conceptual synthesis grounded in socio-technical systems theory and the organizational governance literature, the paper identifies recurring governance failure modes, including authority drift from human decision-makers to automated systems, diffusion of accountability, governance debt accumulation, and reliance on average-case performance metrics that obscure worst-case risks. To illustrate early governance readiness, an exploratory survey of senior university students—representing early-stage managerial cohorts—was conducted, resulting in the AI Governance Readiness Composite Score (AGRCS). The findings illustrate preliminary patterns in self-assessed governance readiness among early-stage managerial cohorts, without implying statistical generalization or population-level conclusions. The study does not seek statistical generalization but uses empirical signals to support conceptual arguments. The main contribution lies in positioning leadership authority, intervention capacity, and governance-related communication as central pillars of sustainable AI governance. The article translates these governance principles into an educational agenda, proposing sustainable pedagogy practices such as authority mapping, escalation rehearsals, worst-case simulations, and governance-focused learning environments. By framing AI governance as a leadership and communication challenge rather than a narrow technical problem, the study contributes to sustainable organizational development, responsible decision-making, and long-term societal trust aligned with the United Nations Sustainable Development Goals.

Keywords: sustainable pedagogy; managerial education; artificial intelligence governance; organizational accountability; algorithmic decision-making; AI risk management; educational transformation; socio-technical systems; decision-making authority; AI accountability and oversight; technology-enhanced education

Academic Editor: Anabela Carvalho Alves

Received: 19 April 2026
Revised: 25 May 2026
Accepted: 26 May 2026
Published: 2 June 2026

1. Introduction

When companies rely on artificial intelligence (AI) to guide their decisions, the real challenge often lies elsewhere—leaders might possess advanced technology yet struggle to apply it properly [1,2]. This article reconceptualizes AI failure as a foreseeable governance issue and defines the competencies managers must acquire to mitigate such failures [3].

Consider an organizational setting in which a computer system is used to detect and prevent cyber threats. For example, AI-based monitoring systems may incorrectly flag employees as security threats due to flawed training data. Without clearly defined oversight and intervention mechanisms, such systems can escalate errors autonomously, undermining trust and accountability. The company needs to ensure the system works in a way that benefits everyone [2]. Real-world incidents of this kind underscore the urgent need for effective AI governance [4]. Preventing these outcomes depends on whether managers are trained to define decision boundaries, assign accountability, and intervene when risks emerge—not only on whether models are accurate [1,3].

AI-related harms are often misattributed to model error, bias, or data quality, implicitly framing failures as technical malfunctions. However, AI systems operate within socio-technical decision environments where human and algorithmic elements are tightly coupled. Failures therefore emerge when authority is unclear, accountability is diffused, and intervention mechanisms are absent. These patterns reflect deficiencies in governance design rather than isolated technical breakdowns [5,6].

As AI systems increasingly shape decisions across finance, healthcare, safety, rights, and other critical domains, boards and executives face accountability for governance, not merely technical performance [1,7]. Courts and regulators are now focused less on output predictability and more on whether harm was foreseeable in the operational context, and whether responsible parties exercised appropriate care in system design, deployment, and oversight [3,8].

Correspondingly, legal and regulatory discourse has shifted from narrow assessments of technical malfunction toward standards emphasizing foreseeability and appropriate care. Under this emerging framework, organizational leaders are evaluated not only on system accuracy but on whether governance arrangements reasonably anticipated potential harms, assigned clear responsibility, and enabled timely intervention. Managerial accountability thus increasingly depends on governance design choices rather than on post hoc claims of technical compliance.

This shift creates significant governance challenges for organizational leaders. Organizational leaders remain accountable for AI-driven outcomes even when systems are complex, opaque, or difficult to interpret. When authority is dispersed, accountability becomes blurred, and no one retains the ability to halt systems in real time, organizations invite AI-related risks. Boards should recognize that as AI's impact expands, personal liability increases when governance lacks clarity and foresight [4,7]. Linking predictable harms to legal responsibility encourages executives to take accountability seriously.

When AI systems are used without a plan for who is responsible, harmful organizational and social consequences may occur. Poor AI governance can produce harmful organizational and social outcomes, including unfair or opaque decisions in high-stakes contexts.

Accordingly, this article analyzes recurring governance failure modes driving AI-related harm and derives the resulting implications for managerial education. We are considering a decision-architecture framework that synthesizes major organizational failure patterns and translates them into a manager education agenda and governance competencies.

AI governance is important for more than protecting organizations; it also affects how we can support sustainable social outcomes [9–11]. As AI starts to control who can get

a job, get money, see a doctor, use services, and access the internet, bad governance can make things unfair, and people will not trust the system. Gaps in managers' knowledge of AI are a global problem, not a company-specific one. We need to teach managers how to make AI-assisted decisions and to stop them when something goes wrong. Fair outcomes start with checks like these, protecting those most at risk from bias creeping into machine decisions. Success in the digital age becomes possible when no one is left behind by smart tools that run everything.

The transformation of education extends beyond technical literacy. It includes equipping current and future leaders with governance capabilities necessary for responsible AI deployment. Within this context, AI governance education becomes a core component of educational transformation, bridging corporate learning, executive development, and formal higher education systems.

Recent research on sustainability education emphasizes integrating AI into higher education curricula to foster sustainability consciousness and responsible technological development [12]. At the same time, emerging work on generative AI literacy highlights the importance of equipping learners with social, ethical, and sustainability-oriented competencies for responsible AI adoption [13]. These perspectives reinforce the need to conceptualize AI governance education not merely as technical training but as a foundational component of sustainable organizational leadership.

Despite the growing importance of AI governance, most studies focus on technical accuracy, ethical principles, or compliance frameworks rather than on how leadership authority, accountability structures, and communication practices shape decision outcomes. Furthermore, few academic programs prepare future managers to define decision boundaries, assign responsibility, or intervene effectively in AI-driven processes.

1.1. Contribution and Positioning Within AI Governance Research

Existing research on AI governance has primarily focused on regulatory compliance frameworks, ethical principles, technical safeguards, and organizational control mechanisms. While these contributions provide valuable guidance on oversight structures and responsible AI requirements, they frequently treat leadership authority, intervention capacity, and organizational communication as secondary or supporting concerns rather than as core governance mechanisms.

This study advances AI governance research by reframing governance failure as a leadership and decision-architecture problem rather than a technical malfunction or compliance deficit. Drawing on socio-technical systems theory and organizational governance literature, the article identifies recurring governance failure modes— including authority drift from humans to automated systems, diffusion of accountability, governance debt accumulation, and reliance on average-case performance metrics—that predictably produce AI-related harm.

The primary contribution of this study lies in (1) conceptualizing AI governance as a managerial capability centered on authority allocation, accountability design, intervention readiness, and communication clarity; (2) demonstrating how AI systems function as governance stress tests that expose latent organizational weaknesses rather than creating novel risks; and (3) translating these governance principles into an educational agenda for preparing managers to govern AI-mediated decision systems responsibly.

By positioning AI governance as a leadership competence embedded within decision systems, this study complements existing governance, ethics, and compliance scholarship while extending it toward practical managerial education and sustainable organizational development.

1.2. Academic Integrity as an AI Governance Challenge

Academic integrity challenges related to AI usage are frequently conceptualized as individual ethical violations or technical detection issues. From a governance perspective, however, failures in academic integrity indicate broader deficiencies in institutional authority allocation, communication, and oversight mechanisms. When organizations implement AI without explicitly communicating acceptable use policies, enforcement responsibilities, and escalation procedures, breaches of integrity become predictable rather than exceptional. The adoption of AI reveals underlying governance weaknesses, such as ambiguous leadership responsibilities and fragmented communication among policy designers, educators, and learners. Effectively addressing academic integrity in AI-enabled environments requires governance structures that clarify decision-making authority, align incentives, and establish transparent communication channels, rather than relying exclusively on technological monitoring or punitive measures.

Table 1 clarifies how this study extends prior work by shifting the focus from principles, compliance, and technical risk toward leadership-driven governance embedded in decision systems.

Table 1. Contribution of this study relative to prior work.

Domain	Prior Focus	Limitation	This Paper
AI ethics & responsible AI	Principles (fairness, transparency, accountability)	Normative; weak operationalization in organizations	Operationalizes governance through authority, accountability, and intervention design
AI governance & risk management	Compliance frameworks, oversight, model risk	Model-/compliance-centric; leadership treated as secondary	Reframes governance as a leadership and decision-architecture problem
Socio-technical systems	Human–AI interaction, systemic risk	Limited translation into managerial practice	Identifies recurring governance failure modes (e.g., authority drift, accountability diffusion)
Management & education/sustainability	AI literacy, skills, ethics awareness	Lacks focus on governance responsibility and intervention capacity	Defines AI governance as a core managerial competence and proposes governance-focused pedagogy

2. Materials and Methods

This study employs a primarily conceptual research design grounded in organizational theory, socio-technical systems analysis, and AI governance scholarship. The core contribution consists of a structured analytical synthesis of recurring governance failure patterns in AI-mediated decision systems. To illustratively support the conceptual argument, an exploratory empirical sub-study was conducted using a survey of senior university students.

The empirical component does not aim to generate generalizable or inferential findings. Instead, senior students are positioned as early-stage managerial candidates whose governance orientations are shaped during formal education. The survey therefore provides an indicative signal of emerging AI governance readiness prior to professional managerial practice. Participant selection followed a convenience sampling approach within two Ukrainian higher education institutions, which limits external validity and restricts generalization beyond early-stage managerial cohorts. Future research may extend the proposed framework by incorporating comparative empirical contexts from other national educational systems.

Ethical considerations were addressed through voluntary participation, informed consent, anonymity, and the absence of personal data collection. Given the minimal risk and non-intrusive nature of the survey, formal ethical approval was not required under applicable institutional regulations.

This article uses a structured conceptual analysis to examine why AI-related harms persist despite technical advances. We treat AI deployment as a decision system composed of models, workflows, incentives, oversight mechanisms, and intervention pathways. We analyze governance breakdowns by identifying recurring patterns in (1) authority allocation, (2) accountability assignment, (3) escalation and intervention design, and (4) risk measurement practices (average-case vs. worst-case). We then synthesize these patterns into a set of governance failure modes and translate them into an education agenda for managers responsible for AI-influenced decisions.

The Results section presents the synthesized failure modes and the decision-architecture mechanisms that produce harm through them.

Conceptual Framework and Educational Implementation Scenario

This study adopts a conceptual framework rather than a hypothesis-driven theoretical model. The framework conceptualizes AI governance as the management of socio-technical decision systems composed of technological models, organizational roles, authority structures, accountability mechanisms, and communication channels. Governance effectiveness depends on four interrelated dimensions: (1) leadership authority allocation, (2) accountability design, (3) intervention and escalation capacity, and (4) risk evaluation logic.

To formalize the conceptual framework, Figure 1 models AI governance as the outcome of interactions among four core dimensions: authority allocation, accountability design, intervention capacity, and risk evaluation logic. Governance effectiveness depends on the alignment of these dimensions. When authority, accountability, and intervention capacity are coherently structured and informed by a worst-case risk orientation, organizations maintain control and resilience. Conversely, misalignment among these dimensions leads to governance fragility, producing failure modes such as authority drift, accountability diffusion, and exposure to high-impact risks.

The framework emphasizes how authority and communication dynamics mediate the translation of AI outputs into organizational actions. Rather than testing causal hypotheses, the framework serves as an analytical lens for identifying recurring governance failure modes and for informing the design of AI governance education grounded in leadership practice.

This study employs a structured conceptual synthesis grounded in organizational theory, socio-technical systems analysis, and AI governance literature. To ensure analytical transparency and replicability, the synthesis followed a four-stage analytical procedure:

1. Identification of recurring AI governance failure patterns across peer-reviewed academic literature, policy reports, and regulatory analyses;
2. Classification of identified failures according to decision-architecture dimensions (authority allocation, accountability design, intervention capacity, and risk measurement logic);
3. Synthesis of these patterns into higher-order governance failure modes;
4. Translation of governance failure modes into corresponding managerial governance competencies and educational objectives.

This approach ensures analytical transparency and replicability of the conceptual synthesis.

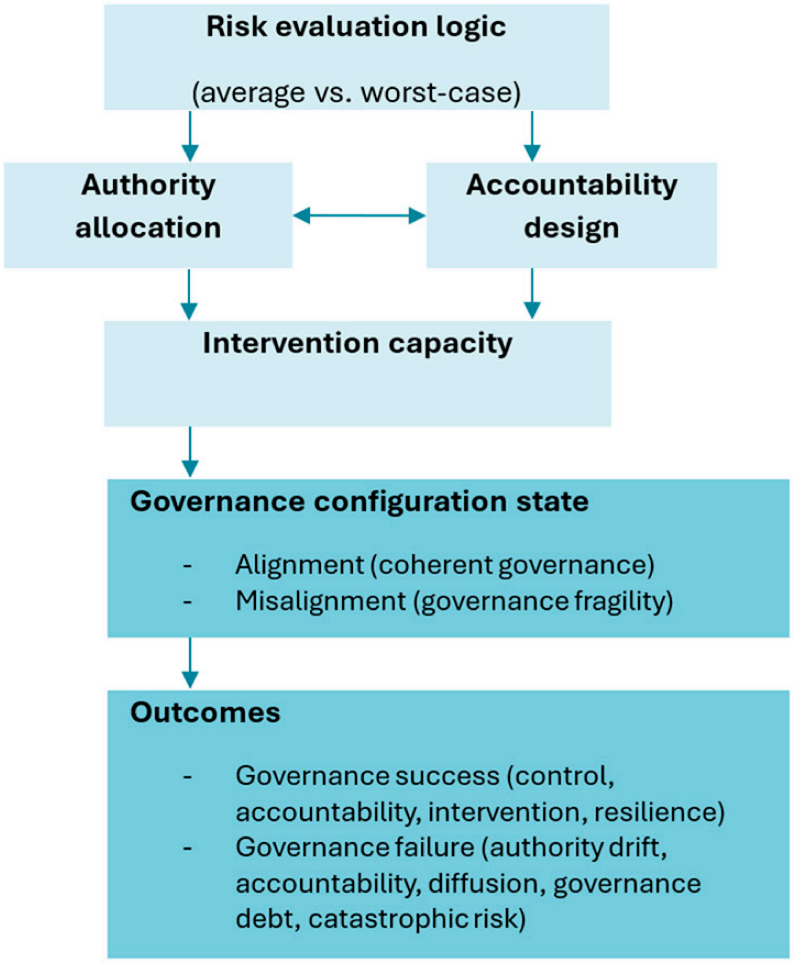

Figure 1. Conceptual model of AI governance in decision systems.

The illustrative implementation scenario is constructed to operationalize the proposed framework. The scenario models the introduction of a managerial AI governance training module within a mid-sized technology-oriented organization. The training program includes:

- Authority mapping exercises;
- Decision-boundary definition workshops;
- Worst-case scenario simulations;
- Intervention protocol rehearsals;
- Governance stress-testing exercises.

The implementation scenario does not constitute an empirical case study. Instead, it functions as a design-based illustrative mechanism intended to demonstrate the feasibility and pedagogical translation of the conceptual governance framework. Its purpose is explanatory and instructional rather than evidentiary.

The scenario serves as a design-based validation mechanism, demonstrating how theoretical governance principles can be translated into structured educational practice. While not an empirical case study, it provides an applied demonstration of feasibility and instructional structure.

Tables and figures are used in this manuscript as analytical and conceptual tools rather than as independent empirical evidence. Figures illustrate governance dynamics and decision-architecture mechanisms, while tables summarize recurring organizational failure patterns synthesized from the literature.

Table 2 summarizes the key governance constructs used in the analysis and provides concise operational definitions, failure mechanisms, and indicative measurement approaches.

Table 2. Operationalization of key AI governance constructs.

Construct	Meaning & Failure	Implication & Indicator
Authority drift	AI outputs become de facto decisions; human oversight becomes symbolic	Teach explicit decision rights; measure % unchallenged AI decisions
Governance debt	Governance gaps accumulate over time and increase risk	Teach continuous review; track time since last governance audit
Decision architecture	Misaligned roles, authority, and workflows produce failures	Train system mapping; assess clarity of decision rights
Intervention capacity	Inability to override AI leads to uncontrolled errors	Practice escalation; measure time-to-intervention
Accountability diffusion	Responsibility is fragmented across actors	Assign clear ownership; measure role clarity
Risk evaluation logic	Focus on averages hides high-impact failures	Teach worst-case thinking; check presence of tail-risk metrics

3. Results

3.1. Misdiagnosing AI Harm as Technical Failure

A recurring pattern is that organizations attribute harm to model defects even when the system performs as optimized, diverting attention from decision authority and oversight design. Most AI systems operate within their specified design constraints—to classify, predict, recommend, or optimize based on given data and objectives. When failure occurs, people usually investigate what happened and find that these systems were used in situations for which they were not intended. Organizations often say that the systems failed because they were not accurate enough, did not work as intended, or lacked sufficient information to learn from. In reality, these systems were used in environments or tasked with making decisions in situations they were never designed to handle, which is the main problem with these systems. Yet in many cases, the AI system performs exactly as optimized.

What usually goes wrong is not the system itself—it is how decisions get made around it. Authority levels play a big role; when leaders move fast on outcomes, change happens more quickly. Recommendations rarely travel far before turning into actions, thanks to tight links between stages (Figure 2). That closed loop lets decisions shift smoothly from machines to people. Figure 2 clarifies how the conceptual framework operates in practice by showing that governance failures arise from interacting weaknesses in authority allocation, accountability design, and intervention capacity rather than from isolated technical errors.

AI systems are advancing rapidly in capability, scale, and organizational influence. The problem is not just that the models are not perfect. Companies are increasingly granting these systems greater autonomy without sufficient oversight or effective mechanisms to intervene in the event of malfunctions.

Most companies see AI more like a gadget than a driver of decisions. Instead of handing it to managers, they give control to those in IT or analytics. Sometimes, when needs grow, they bring in a new expert from afar. When things go sideways with AI, those

in charge often are not the ones held responsible. Power sits separately from responsibility, creating an uneven balance. Faster expectations weigh heavily on many decisions. Organizations lean into AI more than before. Yet pause is rarely part of the process. Safety nets get overlooked when speed drives ahead. Now machines help pick paths that shape outcomes—sometimes in quiet ways. Decisions once checked by people now flow through software without full oversight. It starts as help, then slowly takes over when judgments shift entirely online. Faster operations often skip human checks because handling large volumes quickly adds expense. Even so, control shifts happen—yet leaders do not consistently pay attention or set up clear rules. Systemic issues often go unnoticed, extending beyond isolated errors. Responsibility becomes diffuse when plans change, and no individual assumes ownership of subsequent outcomes. In the event of failure, blame is distributed among the software, data, developers, and users. This diffusion of accountability delays acknowledgment of the underlying problem. Some stakeholders maintain that the system operated as intended. Yet businesses defending it assert they stuck to standard procedures. According to them, current data reflects an honest snapshot of events. Sometimes intervention is needed to address a situation, yet delays, a lack of control, or poor data still occur. There is no clear person/role responsible for approving AI-driven decisions. Just as unclear is who takes charge if mistakes happen under those same conditions. This problem occurs because of how organizations are structured. Thinking of intelligence as just a simple tool, we do not have to deal with a big problem: AI changes who makes decisions, how fast they are made, and how many mistakes are allowed. Since there are no rules for how AI should make decisions, companies tend to rely too much on computers. AI is changing the way organizations work—AI makes decisions that used to be made by people.

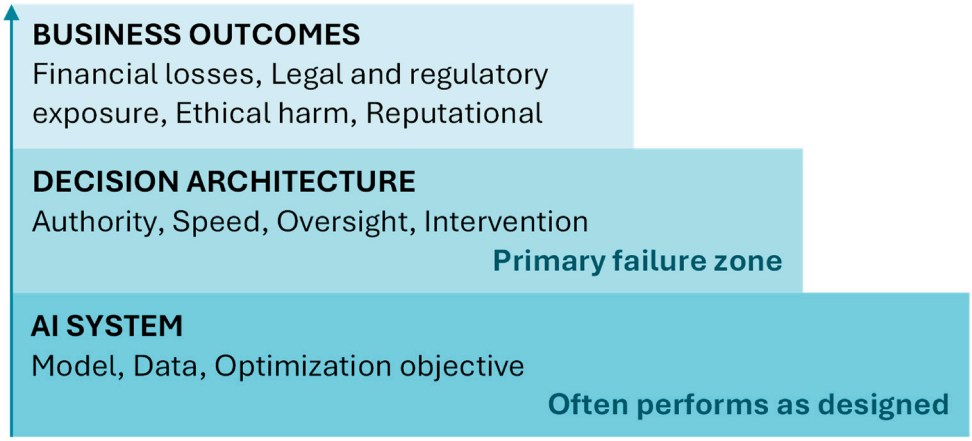

Figure 2. Conceptual sources of AI-related governance failures within organizational decision systems.

The myth of technical failure also obscures the issue of predictability. Many AI-related harms are treated as surprises—edge cases that no one could have anticipated. When we look at the problems that can arise with AI, we see that even though we might not know exactly what will go wrong, we do know what kinds of things can. The thing is, AI can be very big; it can work on its own and be hard to understand. It can work very fast. All these things pose risks to AI—they can cause problems due to its scale, autonomy, opacity, and speed. Deploying systems that act faster than humans can supervise, in environments where errors propagate instantly, predictably creates conditions in which loss of control becomes likely.

AI failures are the foreseeable consequences of delegating authority without redesigning accountability structures. Such failures arise when organizations assume that improved models can compensate for inadequate governance, or that technical excellence can sub-

stitute for institutional oversight. The fundamental failure is that organizations permit AI systems to operate in roles they are unprepared to govern—and then express surprise when control is lost.

In sum, this failure mode concerns not the behavior of specific AI systems but the dominant interpretive frame through which organizations understand AI-related harm. When breakdowns are attributed primarily to technical malfunction or data quality, attention is diverted away from governance arrangements that determine how systems are authorized, monitored, and corrected. This misdiagnosis delays organizational learning and sets the stage for recurring governance failures examined in the following sections.

3.2. AI as a Governance Stress Test

Rather than creating new organizational failures, AI functions as a governance stress test that renders latent weaknesses operationally visible. By accelerating decision cycles and amplifying scale, AI exposes previously informal authority relations, fragmented ownership, and weak intervention mechanisms that were manageable when decisions were slower and more localized. This section therefore focuses on what AI reveals about existing governance structures, rather than restating why governance failures occur. Table 3 translates the conceptual framework into observable organizational patterns by demonstrating how misalignment in authority, accountability, and decision design becomes operationally visible under AI-driven conditions.

Table 3. Failure points emerging from AI integration.

AI Exposes	Observed in Practice	Root Cause	AI Impact
Unclear decision authority	Decisions become faster and more centralized, but no one can clearly identify who truly "owns" them	Decision rights were previously informal and could remain ambiguous when decisions were slow	AI accelerates and standardizes decisions, raising unresolved authority questions
Fragmented ownership of AI initiatives	IT, legal, compliance, product teams, and vendors each manage parts of the system, but no one owns outcomes	Collaboration substitutes for accountability; responsibility is diffused across functions	AI systems operate end-to-end, turning functional gaps into operational and governance failures
Hybrid human–AI decision structures collapsing into automation	Human review exists in name, but meaningful judgment is rarely exercised	Humans retain responsibility without real authority	AI's speed and apparent accuracy push organizations toward de facto automation
Poorly specified objectives and decision criteria	AI systems behave in unexpected or harmful ways that are difficult to interrogate or reverse	Objectives, decision spaces, and success criteria were never clearly defined	AI embeds ambiguity directly into automated processes rather than failing visibly
Treating AI as an efficiency improvement	Teams pursue being "AI-driven," prioritizing speed and internal metrics	Decision systems are not redesigned; efficiency is favored over judgment	AI rewards what it measures, leading organizations to optimize speed while ignoring long-term effects
Human oversight becoming a bottleneck	Review steps are compressed; overrides become symbolic	Governance relies on procedural friction rather than real authority	As decision speed increases, organizations remove friction instead of redesigning oversight

Table 3. *Cont.*

AI Exposes	Observed in Practice	Root Cause	AI Impact
Accountability without control	Frontline personnel remain responsible for outcomes they can no longer meaningfully influence	Authority shifts upward or into systems without corresponding role redesign	AI transfers control to automated systems while leaving accountability with humans
Incentives favoring efficiency over quality	Decisions appear locally rational but undermine trust, legitimacy, or compliance	Incentive structures reward scale and speed rather than judgment	AI amplifies efficiency, turning misaligned incentives into systemic risk

These patterns (Table 3) illustrate how existing governance weaknesses become operationally visible under AI conditions.

AI-driven systems enable decisions to be made at speeds and scales that exceed traditional organizational processes. Decisions that previously required time, distributed judgment, and informal coordination increasingly occur in real time and are embedded within centralized automated workflows. This acceleration renders previously latent governance questions immediately consequential, particularly concerning who effectively exercises decision authority.

As AI is introduced into decision processes, the number of deliberate human judgments typically declines, even when human oversight formally remains in place. Outcomes may become functionally determined by automated outputs rather than by explicit human choice, complicating accountability, interpretability, and fairness. These effects are not the result of technical failure but of governance arrangements that were not redesigned for high-speed decision environments.

Effective AI deployment presupposes clearly specified objectives, defined decision spaces, and explicit criteria for success. When these elements are absent, AI systems do not fail visibly; instead, they amplify existing ambiguity and embed it within automated processes that are difficult to interrogate or reverse once scaled.

AI integration also exposes weaknesses in organizational ownership and accountability. AI initiatives commonly span multiple functions, including IT, legal, compliance, and operations, without a single executive owner responsible for outcomes. What appears as collaboration may therefore diffuse accountability, revealing a limited understanding of how decisions are structured and governed. Rather than differentiating decisions that require human judgment from those suitable for optimization, organizations often deploy AI wherever data and efficiency gains permit—a practice that creates predictable risk in value-laden or rights-sensitive domains.

From a governance perspective, AI should be understood less as a source of new failure and more as a stress test that reveals unresolved authority, accountability, and intervention gaps. These gaps are manageable in slower systems but become operationally hazardous once decision speed and scale increase.

3.3. Authority Drift from Decision Support to Decision Substitution

Authority drift constitutes a distinct governance failure mode through which AI systems transition from advisory tools to de facto decision-makers without explicit organizational authorization. While AI systems are often formally described as providing decision support, their output may become treated as decisive in practice, particularly when speed, efficiency, or perceived objectivity are prioritized.

This drift occurs through subtle but predictable mechanisms. Over time, agreement with AI outputs requires no justification, whereas disagreement demands explanation,

additional effort, or managerial risk-taking. As a result, human oversight remains formally present but substantively weakened. Although individuals retain nominal responsibility for outcomes, effective decision authority migrates toward automated outputs, dashboards, alerts, and ranking systems.

Crucially, this shift does not require intentional delegation of authority. It emerges by default when organizations fail to define when AI advises, when humans decide, and under what conditions overrides are required. The resulting configuration creates a "human-in-the-loop illusion," in which procedural checkpoints exist but meaningful control does not. Authority is exercised by systems that cannot bear responsibility, while accountability remains formally assigned to individuals who lack intervention capacity.

Authority drift thus represents a governance failure independent of system accuracy. Even highly reliable systems can induce governance breakdowns when organizations conflate statistical performance with legitimacy and treat automation as a substitute for judgment. Once authority has migrated implicitly, restoring control becomes costly, contested, and slow—often only after harm has occurred.

3.4. Average-Case Metrics Conceal Worst-Case Governance Risk

A common organizational error in AI adoption is overreliance on average performance metrics. AI systems that perform well in 95% of cases often inspire confidence and encourage rapid scaling. However, the remaining 5% often dismissed as edge cases. This 5% often involves high risks, including legal violations, safety incidents, discrimination, and reputational harm (Figure 3).

Average performance metrics are appealing because they align with organizations' existing definition of success (Table 4).

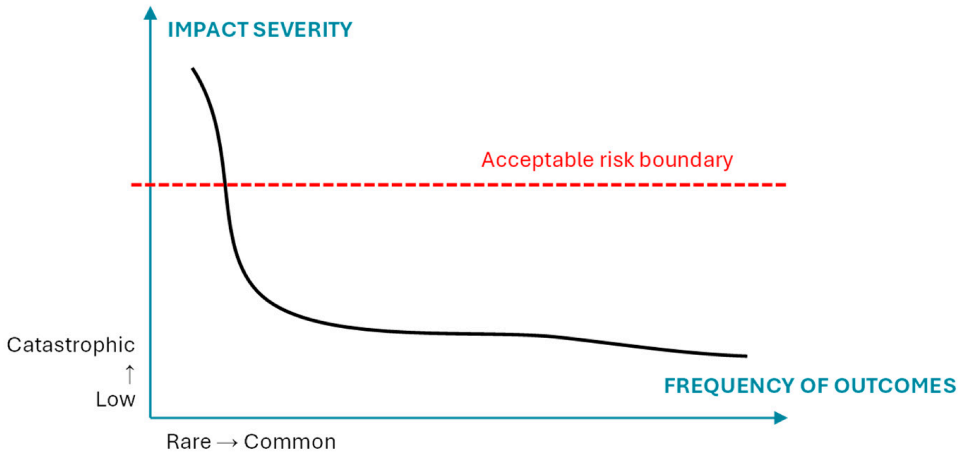

Figure 3. Average performance metrics mask high-impact failures.

Table 4. What organizations measure vs. what governance requires.

Organizations Measure	Governance Requires
Average accuracy	Worst-case impact
Throughput and scale	Containment and recovery
Error rates	Error consequences
Cost reduction	Human authority
Automation	Resilience

This mismatch encourages organizations to scale AI based on operational convenience rather than governance readiness. Dashboards typically reward accuracy, throughput, and cost reduction. Models that outperform humans on average are often perceived as superior,

prompting organizations to expand their use. This perspective obscures the reality. AI failures are rarely gradual—they are often catastrophic and asymmetric [14].

Traditional risk frameworks are poorly suited to these conditions. They assume slow feedback, localized failures, and human judgment. AI systems violate these assumptions by making decisions at speed, propagating errors instantly, and obscuring causal chains. When failures occur, problems spread faster than organizations can detect or correct them. By prioritizing likelihood over impact, organizations become blind to rare disasters that ultimately define leadership responsibility.

Systems designed to replace human judgment optimize for average-case performance rather than resilience in exceptional situations. Automation favors typical conditions—frequent, measurable, and stable scenarios—where AI performs best. As a result, edge cases receive limited representation in training data and insufficient attention in governance.

This situation creates a sense of improvement. When average performance improves, organizations claim that human supervision increases efficiency. In reality, this supervision only takes action when things are really bad, and teams ignore exceptions instead of seeing them as warning signs. The system looks strong, while it is actually weak. The problem is that people think being right most of the time means everything is safe. An AI model that performs well most of the time may still fail catastrophically when decisions affect people's rights, safety, or things that cannot be changed. In these cases, one mistake cannot be made up for by correct results. AI models like these can be problematic because people trust them more.

Safety-critical industries demonstrate a different standard—systems are evaluated by failure-anticipation and detection metrics, but only by failure frequency. For example, aviation, nuclear power, and medicine prioritize worst-case planning, redundancy, and rapid intervention (certification is not based solely on average outcomes). AI systems warrant similar expectations, whether digital or physical. Yet when organizations favor automation over human augmentation, they diminish their capacity to detect rare but serious failures. Authority and situational awareness are lost for relevant roles, so organizations become dependent on systems that appear to work until they fail (often with severe consequences).

The fundamental risk lies in the organizational assumption that averages are enough for effective governance. In high-stakes contexts, such reliance is inadequate.

This failure mode is analytically independent from authority drift. Even when decision authority is formally retained by humans, reliance on average-case performance metrics can induce unjustified confidence and premature scaling. Governance failure arises not from automation itself, but from measurement practices that obscure rare, high-impact outcomes that ultimately define organizational liability and social harm.

3.5. A Decision System Is the Unit of Governance

Taken together, the preceding failure modes demonstrate that AI-related harm cannot be adequately understood or mitigated at the level of individual models. Misdiagnosis of harm as technical failure (Section 3.1), stress amplification of weak organizational structures (Section 3.2), authority drift from decision support to substitution (Section 3.3), and reliance on average-case performance metrics (Section 3.4) all originate within broader decision systems. These systems comprise organizational roles, workflows, escalation paths, incentive structures, and control mechanisms that jointly shape how AI outputs are translated into organizational action.

When a company struggles to define its purpose, risks around control and daily operations tend to grow. Most modern AI rules concentrate on software systems. Companies spend money on clarity aids, fairness checks, and records, yet treat decision power as

an afterthought. What people call AI governance often gets confused with standard tech or data rules. But AI does not just handle facts; it shapes decisions, sometimes without a clear rationale.

A decision system often integrates data, models, workflows, people, and escalation pathways—these elements produce outcomes. Governing models in isolation, without addressing decision-making structures, is akin to certifying an engine while ignoring speed limits or responsibility for failure. AI governance can therefore be analyzed in terms of decision-makers, governance scope, and control mechanisms. The core requirement is the clear assignment of decision-making authority. Governance when authority is unclear—whether because the AI's role (advising, recommending, or deciding) is undefined, or because no one is empowered to intervene after deployment.

The European Union's AI Act highlights the need for clearly defined accountability, risk classification, and oversight mechanisms in the governance of AI systems. It is essential to assess AI's impact, clarify accountability, and define everyone's responsibilities—organisations must monitor the use of AI systems and act quickly if new risks arise. Proper governance deals with both technical and compliance issues. Organizations should also set rules for cases when AI can act on its own and when people need to step in. In this context, effective leadership in AI is measured not by model performance but by the clarity of decision-making authority.

3.6. Governance Debt and Predictable Failure Sequences

The governance failure modes identified above rarely operate in isolation. Over time, they interact to produce predictable sequences of degradation in organizational control. This section examines how unresolved authority drift, diffuse accountability, and inadequate risk measurement accumulate as governance debt, increasing the probability and severity of AI-related harm.

Far from accidental, AI failures show clear underlying structures (Figure 4). Over time, these flaws recur across companies, often tied to weak oversight. Lack of clear responsibility adds another layer. Decisions start drifting away from proper review. Rewards sometimes push teams toward shortcuts rather than honesty. These issues do not pop up only in new fields like robotics testing. They run deep in critical areas, including air travel routes mapped by computers, banking systems checked nightly, hospital procedures audited monthly, and power plant safety reviewed under strict rules. Problems emerge when AI amplifies old issues. Problems grow because promises clash with time pressures, blurred accountabilities, and rewards that skew decisions.

When speed is the main focus, teams may not clearly define roles, responsibilities, or controls. Quick fixes tend to stick around, and pilot systems sometimes move quietly into full production. Once these systems are running, risk assessments are usually overlooked and rarely updated. Responsibility for AI risks is often spread across groups that lack real authority. Ethics boards might spot problems, but cannot take legal action. Review committees may see dangers, but cannot assign resources to fix them. Individual employees might notice warning signs but lack the resources or authority to act. These gaps in the system allow problems to persist and get worse. Effective AI safety comes not from better algorithms alone but from strong, continuous governance and risk management—including clear ownership, risk classification, proactive planning, defined boundaries, and adaptive oversight that treats AI as an ongoing organizational responsibility [15].

What comes through clearly is how flaws in oversight shape AI-related risks. When responsibility stays vague, problems tend to follow. Authority slips without notice—problems grow quietly. Without proper tools to step in, the control slips further. Focusing only on typical cases misses the extreme, and harm slips through.

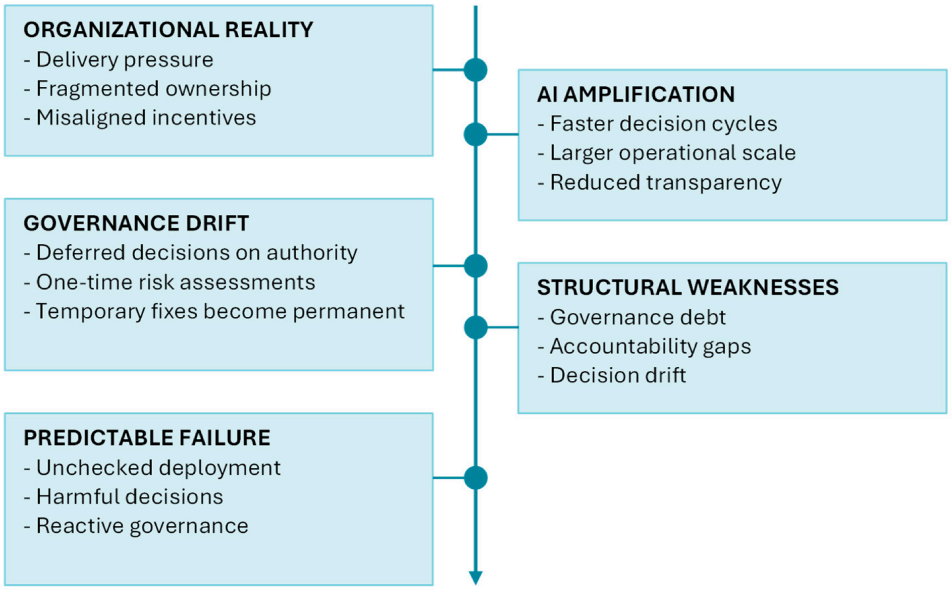

Figure 4. Repeatable organizational sequence underlying AI-related governance failures.

Each governance failure mode identified in the Results section reflects a structural mismatch between decision authority, accountability, and organizational communication rather than model malfunction. These patterns are analytically significant because they recur across sectors and are predictable once leadership roles and escalation mechanisms are left undefined.

3.7. Exploratory Survey Illustration of AI Governance Readiness

To illustratively complement the conceptual synthesis, a pilot survey of 100 senior students was conducted at the State University of Information and Communication Technologies and Borys Grinchenko Kyiv Metropolitan University. Senior university students are positioned in this study as early-stage managerial candidates rather than proxies for experienced managers. The survey therefore captures preliminary governance orientations formed during formal education, rather than established professional competencies. Findings are interpreted accordingly and are not generalized to practicing managerial populations. Respondents represent future managerial cohorts and therefore serve as a proxy population for early-stage AI governance capability assessment. Based on Sections 3.1–3.6, AI governance readiness was operationalized across 4 theoretically derived dimensions based on eight questions:

- Authority Clarity (AC)—understanding of decision rights and override authority ("I understand why clear decision authority is necessary when AI systems are used", "I understand the importance of defining who can override AI system outputs");
- Accountability Design (AD)—ability to assign responsibility and escalation paths ("I feel personally ready to take responsibility for decisions influenced by AI systems in my future career", "My current education has prepared me to govern AI-related risks in a managerial role");
- Intervention Capacity (IC)—competence in suspending or reversing AI decisions ("I understand how AI systems influence organizational decision-making", "I feel confident identifying high-impact AI risks in hypothetical business scenarios");
- Worst-Case Risk Orientation (WR)—recognition of tail-risk and catastrophic failure potential ("I believe organizations should evaluate AI systems based on worst-case risks, not only average performance.", "I am aware of the potential risks (bias, legal, reputational, ethical) associated with AI-driven decisions.").

Responses were recorded using a 1 to 5 scale, with 1 indicating strong disagreement and 5 indicating strong agreement.

Let x_{ij} denote the response of individual j to item i, responses normalized to [0; 1] scale:

$$z_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} = \frac{x_{ij} - 1}{4}$$

Four dimensions' scores are computed as arithmetic means of normalized items within each construct:

$$D_k = \frac{1}{m_k} \sum_{ij \in k} z_{ij}$$

where m_k is the number of responses in dimension k. Mean score calculations and internal consistency estimates using Cronbach's alpha are presented in Table 5:

$$\alpha_k = \frac{n_k}{n_k - 1} \left(1 - \frac{\sum_{i=1}^{n_k} \sigma_{ki}^2}{\sigma_{kT}^2} \right) = 2 \left(1 - \frac{\sum_{i=1}^{2} \sigma_{ki}^2}{\sigma_{kT}^2} \right),$$

where n_k is the number of items (questions) in dimension k, σ_{kT}^2 variance of the total score in dimension k, σ_{ki}^2 variance of each item in dimension k.

Methodological Limitations of the Exploratory Instrument

The AI Governance Readiness Composite Score (AGRCS) does not constitute a validated psychometric instrument. The limited number of items per dimension, the conceptual breadth of the constructions, and the exploratory nature of the pilot survey constrain internal consistency. Low and, in one case, negative Cronbach's alpha values reflect item heterogeneity and insufficient scale development rather than empirical contradiction of the underlying governance arguments.

Table 5. Calculations results.

k	AC	AD	IC	WR	AC + AD + IC + WR (All Questions)
m_k	200	800			
D_k	0.575	0.468	0.550	0.444	0.509
$\sum_{i=1}^{n_k} \sigma_i^2$	25.13	21.79	23.25	24.05	96.62
σ_T^2	29.13	25.45	20.75	30.67	186.7
α_k	0.2747	0.2879	−0.2410	0.4315	0.9647

These results highlight the methodological challenge of operationalizing AI governance readiness as a measurable construction. Governance readiness encompasses multiple interdependent leadership capabilities—including authority clarity, accountability design, intervention capacity, and worst-case risk awareness—that are not expected to function as a single latent trait at early stages of education. The AGRCS is therefore used strictly as an illustrative aggregation to support conceptual discussion and to inform future instrument refinement rather than as a basis for inferential interpretation.

The observed reliability results confirm that the AGRCS should be treated solely as a conceptual illustration rather than a validated measurement scale, and that it is not sufficiently reliable for strong empirical interpretation. The findings reinforce the argument that AI governance readiness represents a complex configuration of leadership capabilities rather than a unidimensional construct, particularly in early-stage managerial education. Accordingly, the composite AGRCS should be interpreted with caution and not used as a validated aggregate measure of governance readiness.

The internal consistency of the four proposed governance dimensions is low (α ranging from -0.2410 to 0.4315), indicating that the instrument should be treated as exploratory.

AI Governance Readiness Composite Score (*AGRCS*) is defined as a weighted linear aggregation:

$$AGRCS_j = \sum_{k=1}^{4} w_k D_{jk}$$

AGRCS distribution among respondents given on Figure 5 (weighting schema with equal weights).

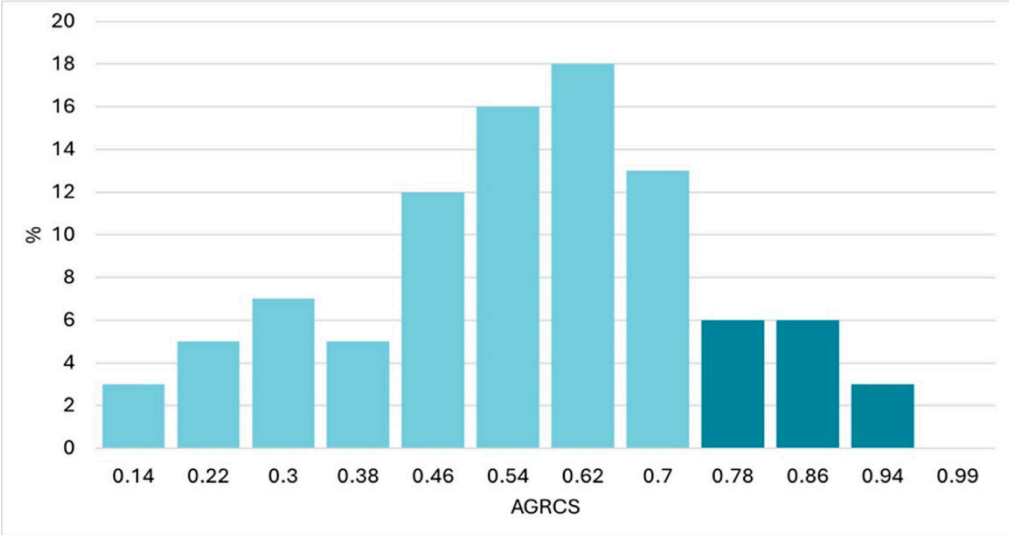

Figure 5. AGRCS distribution among respondents. Light blue bars indicate respondents with AGRCS < 0.70, whereas dark blue bars indicate respondents with AGRCS $\geq$ 0.70 (high readiness category).

Governance readiness categories were defined for illustrative purposes:

Low (28%) *AGRCS* < 0.50;

Moderate (56%) 0.50 $\leq$ *AGRCS* < 0.70;

High (16%) *AGRCS* $\geq$ 0.70.

Thus, within this exploratory sample, most respondents do not fall into the "high readiness" category; this distribution is presented solely as an illustrative pattern rather than as evidence of population-level characteristics.

These categorical thresholds are heuristic and used solely to illustrate distributional patterns rather than to assert measurement precision.

4. Discussion

This study builds on and integrates three strands of literature already outlined in Section 1.2—AI ethics and responsible AI, governance and risk management frameworks, and socio-technical systems research. Rather than restating these perspectives, the Discussion interprets the conceptual findings in relation to these strands, emphasizing how authority allocation, accountability design, and intervention capacity function as operational governance mechanisms within decision systems.

This study demonstrates that AI-related harm is most consistently associated with governance and leadership capability gaps rather than isolated technical errors. The exploratory pilot survey does not provide evidence of definitive managerial deficiency; instead, it signals potential weaknesses in governance preparedness emerging during early stages of managerial formation. These findings are best interpreted as indicative inputs

for AI governance curriculum design rather than as empirical confirmation of systemic managerial failure. The conceptual analysis aligns with governance-oriented AI scholarship emphasizing organizational responsibility over technical blame, while extending it by foregrounding leadership authority, intervention readiness, and communication structures as central governance mechanisms. By integrating socio-technical systems theory with decision-architecture analysis, the study shows how AI exposes latent governance weaknesses that, if left unaddressed, predictably generate organizational and societal harm.

4.1. Managerial Capability Gaps as Drivers of AI Governance Failures

Poor AI governance leads organizations to bad decisions. Thus, managers need to focus on the decision-making process rather than clarifying low-level technical details, but they have not been trained to address new AI-caused issues, and they may not be able to understand the underlying machines. Managers need practical experience in governing the use of new technologies and implementing appropriate controls on a dynamic system. These controls help organizations monitor how technology behaves. Better oversight leads to better results in complex automated settings. To work effectively with AI, leaders must build specific skills in these technologies.

4.2. Authority Drift and Intervention Capacity as Central Governance Challenges

Decision-making authority has become less clear in many organizations. Authority drift results in people who once had influence now having less real power—technology shapes decisions more than people do. Systems meant to help are taking control, sometimes without anyone noticing. As work gets harder and deadlines get tighter, technology can quietly take over even more. When problems arise, people still need to fix things manually, since computers have limits and can make mistakes. It is important to identify accountability when systems fail, and human oversight is needed. Managers without real authority cannot be held responsible for results. Often, solving problems requires quick action, not long planning, so weak oversight can slowly undermine how an organization runs. When authority shifts, especially with AI, new problems can arise. These changes deserve closer attention than they usually get.

These findings reinforce prior research on socio-technical risk and governance frameworks while extending them by demonstrating that authority and intervention capacity are central, rather than supporting, elements of effective AI governance.

4.3. Governance Requires Managing Decision Systems Instead of Solely Evaluating Models

Governance issues in systems that involve both people and technology often stem from how they operate at their core. The quality of functions depends on approaches, incentives, and the ways problems are solved. Hybrid human–AI systems rely on routines, motivations, and ways to keep work moving, which go beyond technology. Authority should match responsibility at every level of an organization. Focusing only on technical skills (models, architecture, and implementation) can miss the core ideas of good governance. Manager training should cover how decisions are made, how authority is shared, and how to prepare for unexpected changes in their tasks. The main goal of management training is to improve job performance by making better decisions.

4.4. Risk Governance Should Prioritize the Worst-Case Impact

Most companies evaluate AI tools based on how well they typically perform, yet serious, infrequent breakdowns can bring operations to a halt. That difference—one focused on usual results, the other on extreme consequences—shows a clear divide between getting things to run right and handling threats properly. Leaders' training should include methods for forecasting potential outcomes and focus on managing risks before they grow too large.

When challenges arise, and steps are already clear, handling them becomes easier. This shift allows leadership to respond more calmly during tough moments. Decisions work better when made not only during good times but also during pressure. Right now, AI plays a role in these changes. Things might shift in ways no one expects, so preparation becomes necessary.

4.5. Manager Education as a Primary Governance Mechanism

These findings identify manager education as a central mechanism for effective AI governance. Managers must develop competencies in authority allocation, accountability design, escalation management, and risk evaluation. Without these competencies, organizations risk creating accountability structures that assign responsibility without enabling control. Educating managers to govern AI enables organizations to align operational authority with accountability and maintain effective oversight over automated decision processes. In this sense, managerial education is not supplementary to AI governance—it is foundational to it.

4.5.1. Sustainable Pedagogy in AI Governance

Sustainable pedagogy in AI governance is concerned not only with skill acquisition but with preparing managers to preserve social trust, fairness, institutional legitimacy, and long-term resilience in AI-mediated decision environments. Such pedagogy emphasizes responsibility for future consequences, anticipatory governance, and the ethical exercise of authority within complex socio-technical systems.

To make AI governance education relevant and meaningful, we must look beyond skills development. In the context of sustainable pedagogy, AI governance cannot be reduced to a purely technical competence. On the contrary, it requires managers to develop the ability to lead on moral and ethical grounds in complex socio-technical systems. Managers need to be educated not only to know how to distribute power and set up control mechanisms, but also to consider the future implications for society, the new power relationships that arise, and the responsibilities of institutions.

Managers are traditionally taught to manage organizations efficiently and to be competent in optimizing outputs and meeting performance targets. These approaches are not well-suited to the context of AI systems, because the underlying governance failures (including authority drift, accountability diffusion, and risk averaging) and unusual/unpredictable scenarios are amplified by standard approaches to thinking about managerial responsibility and effectiveness. Sustainable approaches to the pedagogy of AI systems require a significant departure from conventional norms and focus on forward-looking risk analysis of unusual scenarios rather than on systems thinking, ethical awareness, etc.

It is important to train managers to assume personal responsibility for AI by taking ownership of any AI-based decision-making before it causes damage. AI governance is closely linked to a number of sustainability targets, such as:

- Protecting the corporate reputation;
- Respecting human dignity;
- Ensuring that new technologies increase society's resilience.

Training students in the principles of AI governance prepares them to preserve the role of the human in technology.

4.5.2. Operationalizing AI Governance Education: Learning Activities and Outcomes

Translating AI governance principles into managerial education requires concrete learning activities aligned with assessable outcomes. Governance-focused pedagogy should move beyond abstract ethical discussion toward structured capability development. Core

educational components may include authority-mapping exercises that require learners to explicitly assign decision rights and escalation pathways in AI-mediated scenarios; worst-case simulations that train managers to recognize and respond to rare but high-impact failures; and intervention rehearsals that develop confidence in suspending or overriding automated decisions under time pressure.

Assessment should emphasize demonstrated governance competence rather than technical knowledge alone. Learning outcomes may include the ability to articulate decision boundaries, justify intervention decisions, evaluate AI systems based on worst-case risk rather than average performance, and communicate accountability clearly across organizational roles. These pedagogical elements enable AI governance education to function as a practical leadership training mechanism rather than as compliance instruction.

4.6. AI Governance Education and the Sustainable Development Goals

AI governance education contributes to the Sustainable Development Goals (SDGs) through the development of specific managerial capabilities—authority allocation, accountability design, intervention capacity, and worst-case risk orientation—that shape how AI systems are deployed within organizational decision structures.

For SDG 4 (Quality Education), the study identifies a gap in higher education preparedness for governing AI-mediated decisions. Governance-focused pedagogical practices—such as authority mapping, escalation rehearsals, and high-impact risk simulations—address this gap by embedding decision-making competencies directly relevant to socio-technical systems.

Regarding SDG 8 (Decent Work and Economic Growth), poorly governed AI systems can produce unfair or opaque outcomes in employment and access to services. AI governance education contributes by enabling managers to design accountability structures and intervention mechanisms that mitigate harmful automated decisions and support fairer organizational practices.

For SDG 9 (Industry, Innovation, and Infrastructure), AI systems expose weaknesses in authority, ownership, and oversight when deployed at scale. Without adequate governance, AI deployment may prioritize efficiency over safety and fairness, increasing the risk of harmful or unaccountable outcomes [16,17]. AI governance education addresses this risk by aligning innovation with clearly defined decision rights, escalation pathways, and resilience-oriented risk evaluation.

In relation to SDG 16 (Peace, Justice, and Strong Institutions), authority drift, accountability diffusion, and limited intervention capacity undermine institutional responsibility and social trust. AI governance education supports SDG 16 by preparing managers to design decision systems with identifiable responsibility and effective oversight.

Overall, the contribution to the SDGs is conditional: AI governance education advances sustainability only when these governance capabilities are effectively developed and applied, ensuring that AI systems support rather than undermine organizational and societal outcomes.

These relationships indicate that AI governance education contributes to the SDGs not only through general awareness but through specific organizational mechanisms. In particular, authority allocation and accountability design influence SDG 16 (institutional strength), intervention capacity supports SDG 9 (resilient infrastructure), and governance-focused education directly advances SDG 4 (quality education). This mechanism-based linkage strengthens the analytical connection between AI governance and sustainability outcomes.

4.7. Technology-Enhanced Education for AI Governance Competencies

One way to develop AI governance skills is through learning methods like how people acquire tech-based knowledge today. Not just old-style conferences for executives will do when dealing with rapid shifts in digital society. Building capability might work better by weaving it into organized online training spaces instead. One way to teach responsible AI rules is by using mixed or full online courses. Here, scenarios can power shifts unpredictably, showing how control can unravel under pressure. Exploring how choices shape outcomes through live boundary mapping. Interactive decision boundaries show relationships as they unfold. AI helps govern through test zones. AI assists labs in sandbox setups. Adaptive e-learning tools adjust as users interact, shaping risk assessment situations based on individual learning paths.

Real cases showing how AI governance actually fails. During planning sessions, teams can use digital platforms to share ideas and coordinate efforts. Learning from simulations matters more now. When AI fails, damage tends to happen during critical moments, so leaders should practice how they handle such cases in safe virtual environments. What stands out is that teaching with AI requires ethics woven into it, along with considering risks and preparing for the worst when building online courses. That change moves learning beyond just picking up tools to understanding better choices within systems.

The proposed AI governance curriculum relies on technology-enhanced learning mechanisms. These include adaptive e-learning systems, governance simulation sandboxes, AI-powered scenario branching, decision boundary visualization tools, and real-time risk dashboards. Such tools transform governance education from static compliance instruction into dynamic capability-building environments that reflect real organizational complexity. Figure 6 illustrates how the proposed framework can be operationalized in a technology-enhanced learning environment.

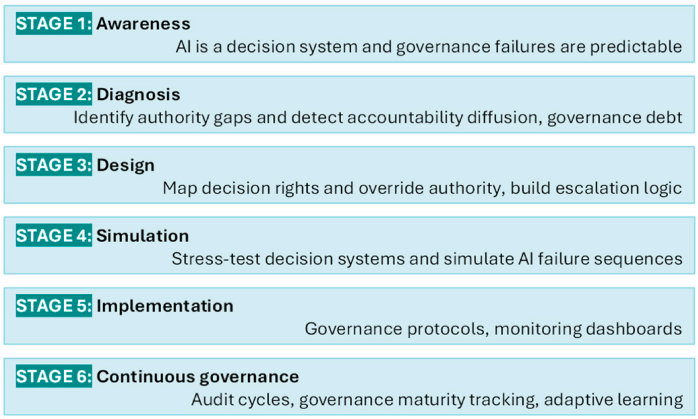

Figure 6. Manager learning path within the AI governance framework.

Modern research supports proposed ideas. The use of AI-enhanced pedagogical environments to build sustainability competencies has been shown to increase systems-level awareness and reflective decision-making in engineering and management education. Furthermore, AI literacy research stresses that generative AI education must include ethical reasoning, accountability, and sustainability principles alongside technical skills. Governance training programs should therefore integrate simulation-based learning with ethical and socio-technical reflection.

4.8. The Role of Universities in Preparing Leaders for Sustainable AI Governance

Leading in the field often begins at universities where skills grow for future roles involving AI. Instead of just learning tools, students usually study fields like business,

engineering, or public policy through an AI lens. Yet classroom discussions rarely address fairness in their application. Safety considerations also appear less frequently than expected.

Meeting SDG targets might be easier when institutions share knowledge on intelligent governance. Instead of waiting, schools can equip learners by exploring how choices are made. Holding individuals responsible for their decisions matters just as much as understanding them. Accountability grows stronger through reflection and clear expectations. They should teach students to examine decision-making systems, ensure that people are held accountable for their actions, identify the risks of using AI, and address problems when things go wrong. If colleges and universities teach these things, it will help them remain around for a while and help students gain the skills they need to be good leaders in a world where AI is a big part of the economy.

This is important because the SDGs state that we should provide education, including teaching people about the governance of AI. To prepare people to be leaders in a world with AI, colleges and universities need to ensure students learn about management and AI governance in their core classes, not just in specialized extra classes.

Recent research underscores that higher education institutions play a central role in cultivating AI literacy frameworks that integrate ethical, social, and sustainability considerations. When AI education is linked explicitly to sustainability outcomes, it can shape future leaders capable of aligning technological deployment with long-term societal goals.

4.9. Recommendations

For university leadership: Universities should formally designate executive-level responsibility for AI governance, ensuring that academic integrity policies, decision authority, and escalation procedures are clearly communicated across the institution.

For policymakers: National and institutional policymakers should move beyond technical compliance guidelines and require governance capacity development, including leadership training in authority allocation, intervention readiness, and risk communication.

For faculty and program designers: Curricula should incorporate governance-focused learning methods such as decision-authority mapping, scenario-based simulations, and worst-case risk analysis to prepare future managers to govern AI-mediated decisions responsibly.

Each recommendation directly addresses the governance gaps identified in the analysis, linking leadership capacity building to sustainable and accountable AI deployment.

Rather than challenging existing governance frameworks, this study complements them by identifying the managerial mechanisms through which governance principles are enacted in practice.

5. Conclusions and Current Research Limitations

This study argues that the appropriate unit of AI governance is the decision system—comprising organizational roles, authority structures, workflows, escalation mechanisms, and communication channels—rather than individual AI models considered in isolation. The empirical component of this study is subject to important limitations related to sample scope and design. The exploratory survey is based on a convenience sample of 100 senior students from two Ukrainian universities, which restricts external validity and precludes generalization beyond early-stage managerial populations. While these respondents are analytically useful as indicative of governance orientations formed during formal education, they do not represent practicing managers or broader organizational contexts. Accordingly, the empirical findings should be interpreted strictly as illustrative signals supporting the conceptual framework rather than as population-level evidence. AI-related harm emerges predictably when decision authority shifts implicitly to automated systems,

accountability becomes diffuse, intervention capacity is weakened, and governance relies on average-case performance metrics that obscure catastrophic risk.

Over time, repeated breakdowns in how things are run have become clear. When people confuse deep-seated governance issues with mere tool malfunctions, they might hand power away from real users toward machines—yet never officially approve such moves. At the same time, responsibility gaps appear because rules around overriding situations or raising alarms lack clarity or simply disappear when needed. Focusing only on typical results hides occasional yet devastating breakdowns, making systems more vulnerable to extreme outcomes.

Looking at how decisions are made with AI, the study demonstrates that the appropriate unit of governance is the decision system—the interaction of models, organizational roles, workflows, and intervention mechanisms—rather than individual AI models. It pulls together patterns observed across different systems for managing machines. A key push comes from shaping training for those who run these operations—setting clear lines for choices matters most. Who gets to change outcomes needs to be named explicitly. When problems arise, there must be ways to route them safely upward and limit damage quickly. People applying these ideas take away important points: who decides what under what conditions; how to take back control without confusion; how to control rare cases that could unravel trust. The exploratory survey provides a preliminary signal that gaps in AI governance readiness may already be present at early stages of managerial formation. While not statistically confirmatory, this finding reinforces the conceptual argument that governance failures are closely linked to capability development rather than to technological limitations alone. This indicates that governance failures are strongly associated with capability gaps rather than technological limitations.

From a sustainability perspective, the findings suggest that teaching managers about AI governance—particularly authority allocation, accountability design, intervention capacity, and worst-case risk reasoning—helps build lasting defences against technologies that could harm society. With clear authority, responsibility, and the ability to act, groups can steer AI development toward fairer outcomes, stronger institutions, and continued social balance, rather than worsening gaps or increasing danger.

Based on the AGRCS, the initial findings provide a preliminary verification of existing management gaps and will require further validation and amendment, such as improving the reliability of the measurement scale, increasing the number of items per construct, achieving greater consistency across the items, conducting exploratory factor analysis, and performing further tests in other management contexts to confirm and validate the construct. Further studies can be conducted using case studies, surveys of the target population, and experimental studies.

The exploratory survey provides a preliminary signal that gaps in AI governance readiness may emerge early in managerial formation. While not statistically confirmatory, this signal reinforces the conceptual argument that AI-related governance failures are closely associated with deficiencies in capability development rather than with technological shortcomings. Further research employing validated instruments, practicing managerial populations, and longitudinal designs is required to substantiate and generalize these findings.

In addition, this study highlights that effective AI governance is inseparable from leadership authority and organizational communication. Governance failures frequently occur not because managers lack ethical intent or technical awareness, but because authority to intervene is unclear and communication channels for accountability are weak or informal. By emphasizing decision authority, escalation capability, and clarity of responsibility, the study contributes a leadership-centered perspective to AI governance research. These

insights extend prior discussions of AI ethics and compliance by demonstrating that sustainable AI governance depends on how leaders communicate boundaries, assign responsibility, and act under conditions of uncertainty. Strengthening these leadership and communication capacities is therefore essential for organizations seeking to deploy AI in a manner that is both accountable and socially sustainable.

Future Research Directions

Several avenues for future research emerge from this study. First, the proposed framework should undergo empirical validation in diverse organizational and institutional contexts, including research involving practicing managers. Second, additional efforts are necessary to develop and validate robust instruments for assessing AI governance readiness, with particular attention to scale design and construct validation. Third, longitudinal and experimental studies may investigate the evolution of governance capabilities and the influence of educational interventions on managerial decision-making. Finally, future research should more precisely operationalize the relationship between AI governance mechanisms and sustainability outcomes to evaluate their impact on Sustainable Development Goal (SDG) indicators.

Author Contributions: Conceptualization, methodology, formal analysis, writing, A.S.; resources, data curation, O.Z.; writing—original draft preparation, K.O.; writing—review and editing, N.L.; conceptualization; methodology; visualization, A.B.; supervision, V.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because under Ukrainian national legislation, mandatory ethical review by an Ethics Commission is strictly codified for clinical trials of medicinal products, as outlined in the Ministry of Health of Ukraine Order No. 690 (Registered under No. z1010-09)—On Approval of the Procedure for Conducting Clinical Trials of Medicinal Products" dated 23 September 2009. Because this study was a non-interventional, anonymous, minimal-risk survey that gathered no personal, clinical, or biological data, it falls outside the statutory mandate of Order No. 690, and formal institutional ethical approval was not legally required.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study prior to participation in the survey.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank Frank Heidt of Leviathan Security Group, for his support and thoughtful discussions that contributed to the development of this manuscript.

Conflicts of Interest: Author Anton Shantyr was employed by the company Leviathan Security Group. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Shrestha, Y.R.; Ben-Menahem, S.M.; von Krogh, G. Organizational decision-making structures in the age of artificial intelligence. *Calif. Manag. Rev.* **2019**, *61*, 66–83. [CrossRef]
2. Schneider, J.; Abraham, R.; Meske, C.; vom Brocke, J. Artificial intelligence governance for businesses. *Inf. Syst. Manag.* **2023**, *40*, 229–249. [CrossRef]
3. Fraser, H.L.; Suzor, N.P. Locating fault for AI harms: A systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain. *Law Innov. Technol.* **2025**, *17*, 103–138. [CrossRef]

4. Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y.N.; Zhang, Y.Q.; Xue, L.; Shalev-Shwartz, S.; et al. Managing extreme AI risks amid rapid progress: Preparation requires technical research and development, as well as adaptive, proactive governance. *Science* **2024**, *384*, 842–845. [CrossRef]
5. Macrae, C. Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. *Risk Anal.* **2022**, *42*, 1999–2025. [CrossRef]
6. Janssen, M. Responsible governance of generative AI: Conceptualizing GenAI as complex adaptive systems. *Policy Soc.* **2025**, *44*, 38–51. [CrossRef]
7. Brynjolfsson, E. The Turing trap: The promise and peril of human-like artificial intelligence. *Daedalus* **2022**, *151*, 272–287. [CrossRef]
8. El Ali, A.; Venkatraj, K.P.; Morosoli, S.; Naudts, L.; Helberger, N.; Cesar, P. Transparent AI disclosure obligations: Who, what, when, where, why, how. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–11. [CrossRef]
9. Taeihagh, A. Governance of artificial intelligence. *Policy Soc.* **2021**, *40*, 137–157. [CrossRef]
10. Ramírez-Correa, P.; Grandón, E.E.; Mariano, A.M. Mapping the landscape of generative artificial intelligence literacy: A systematic review toward social, ethical, and sustainable AI adoption. *Sustainability* **2026**, *18*, 1429. [CrossRef]
11. Gianni, R.; Lehtinen, S.; Nieminen, M. Governance of responsible AI: From ethical guidelines to cooperative policies. *Front. Comput. Sci.* **2022**, *4*, 873437. [CrossRef]
12. Gahnberg, C. What rules? Framing the governance of artificial agency. *Policy Soc.* **2021**, *40*, 194–210. [CrossRef]
13. Liu, F.; Wang, H.; Guo, Y.; Tang, T. Enhancing sustainability consciousness in higher education: Impacts of artificial intelligence-integrated sustainable engineering education. *Sustainability* **2026**, *18*, 2124. [CrossRef]
14. Ilcic, A.; Fuentes, M.; Lawler, D. Artificial intelligence, complexity, and systemic resilience in global governance. *Front. Artif. Intell.* **2025**, *8*, 1562095. [CrossRef] [PubMed]
15. Ganesh, N.B.; Siddineni, D.; Reddy, V.V.; Ganesha, K.S.; Lateef, K.; Sharma, R. Corporate governance in the age of AI: Ethical oversight and accountability frameworks. *J. Inf. Syst. Eng. Manag.* **2025**, *10*, 6285. [CrossRef]
16. Lozano-Paredes, L. Mapping AI's role in NSW governance: A socio-technical analysis of GenAI integration. *Front. Political Sci.* **2025**, *7*, 1595345. [CrossRef]
17. Shantyr, A.; Zinchenko, O.; Storchak, K.; Bondarchuk, A.; Pepa, Y. Prediction of quality software quality indicators with applied modifications of integrated gradiates methods. *Inform. Autom. Pomiary W Gospod. I Ochr. Sr.* **2025**, *15*, 139–146. [CrossRef]