

Research Article

From Agentic Overreach to Societal Disruption: A Multi-Case Analysis and Structured Evaluation of AI Safety Governance

Yuquan Sun * 

Publicis Sapient, 134 Peter Street, Toronto, ON M5V 2H2, Canada; e-mail : yuq.sun@hotmail.com

* Corresponding Author : Yuquan Sun

Abstract: The rapid deployment of autonomous artificial intelligence systems has outpaced the development of safety protocols, resulting in a series of high-impact incidents in early 2026. This paper presents a systematic multi-case analysis of three representative AI safety failures—Meta’s internal AI agent data exposure, a physical robot malfunction in a public restaurant, and a deepfake-driven disinformation campaign targeting a UK politician—using STAMP (Systems-Theoretic Accident Model and Processes) and contextual integrity as analytical lenses. We identify three systemic root causes: architectural inseparability of instruction and content, inadequate fail-safe mechanisms, and fragmented regulatory enforcement. Building on these findings, we propose the AI Safety Governance Assessment Framework (AISGAF), a multi-layered governance instrument comprising architectural, organizational, sectoral, and global dimensions. To move beyond conceptual description, we operationalize the framework into a set of evaluative indicators with a 3-point scoring scheme and apply it to the three cases. The evaluation results provide initial evidence of the systemic nature of current governance gaps and offer a baseline for assessing future improvements. This study contributes an empirically grounded, evaluable governance framework, extends STAMP to AI agent contexts, and offers a structured methodology for AI safety assessment that can inform both policy and practice.

Keywords: AI agents; AI governance; AI safety; Cyber-physical systems; Deepfake detection; Independent auditing; Industry, innovation and infrastructure; STAMP.

1. Introduction

The first quarter of 2026 witnessed a pivotal shift in AI deployment: AI systems are increasingly entrusted with autonomous decision-making in corporate networks, physical environments, and public discourse. This transition has revealed fundamental safety gaps. In March 2026, Meta experienced a “Sev-1” security incident when an AI agent autonomously triggered permission changes, exposing sensitive data [1]. Simultaneously, a restaurant robot entered a high-intensity dance mode, requiring physical restraint [2]. Beyond operational failures, a UK city councilor was targeted by AI-generated deepfakes in an attempt to undermine local democracy [3]. These incidents suggest that AI failures are not isolated anomalies, but symptoms of broader systemic issues. More fundamentally, they reflect long-standing concerns about aligning increasingly autonomous AI systems with human intentions, constraints, and control mechanisms [4].

Despite growing attention to AI safety, existing research suffers from several interconnected limitations. Case analyses remain fragmented, lacking cross-domain synthesis that could reveal systemic patterns across software agents, embodied AI, and generative AI. At the same time, governance frameworks are often proposed at a conceptual level without empirical validation or operationalization, limiting their practical applicability. Furthermore, while systems theory (STAMP) has been applied to autonomous systems such as self-driving vehicles, its application to AI agent incidents in enterprise and social contexts remains under-explored, particularly in enabling structured cross-case comparison.

Received: March, 22nd 2026Revised: April, 4th 2026Accepted: April, 11th 2026Published: April, 18th 2026Curr. Ver.: April, 18th 2026

Copyright: © 2026 by the authors.
Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>)

To address these gaps, this study analyzes three distinct AI incidents using STAMP and contextual integrity to identify common root causes and cross-domain patterns. Based on these findings, we develop a multi-layered governance framework that directly targets the identified failures. The framework is further operationalized into a set of evaluative indicators and applied to the selected cases, providing a form of semi-empirical validation and demonstrating its applicability as an analytical tool.

This paper makes four main contributions:

- Empirical: A systematic multi-case analysis spanning software agents, embodied AI, and generative AI, with explicit case selection and cross-case synthesis.
- Theoretical: An extension of STAMP to AI agent contexts, highlighting control failures at both agent and platform levels, and clarifying its role in cross-domain analysis.
- Methodological: The operationalization of a governance framework into evaluative indicators with a transparent scoring scheme, enabling structured assessment.
- Practical: The AI Safety Governance Assessment Framework (AISGAF), a referenceable analytical tool for assessing and benchmarking AI safety governance.

The remainder of the paper is organized as follows. Section 2 presents the methodology, including case selection and theoretical foundations. Section 3 describes the cases and provides cross-case analysis. Section 4 develops and evaluates the AISGAF framework. Section 5 discusses theoretical implications, limitations, and future directions. Section 6 concludes the paper.

2. Methodology

2.1. Case Selection

We employed a multiple-case study methodology [5] to capture variation across AI deployment contexts. This approach enables comparative analysis across heterogeneous systems while retaining sufficient contextual depth. Three cases were selected based on several criteria to ensure relevance, reliability, and analytical rigor, namely:

- Timeliness: All cases occurred in Q1 2026, ensuring contemporary relevance.
- Reliability: Each case was confirmed by at least two independent sources.
- Diversity of AI types: The cases represent software-based AI agents (Meta), embodied AI (restaurant robot), and generative AI (deepfake disinformation), enabling cross-domain comparison.
- Analytical depth: Each case provides sufficient technical detail to support in-depth analysis.

Collectively, these cases are information-rich [5] and were strategically selected to illustrate different facets of AI safety failures, enabling the identification of systemic patterns rather than isolated incidents.

2.2. Theoretical Frameworks

To provide a structured analytical lens, this study integrates three complementary theoretical perspectives that capture technical, socio-technical, and institutional dimensions of AI safety. Among these, systems theory—operationalized through Systems-Theoretic Accident Model and Processes (STAMP) [6]—serves as the primary analytical foundation, enabling a shift from component-level failure analysis to system-level control structure evaluation.

STAMP [6] provides a systems-based approach to understanding accidents as the result of inadequate control structures rather than isolated component failures. In this study, STAMP is used to map the control loops governing AI systems and to identify points where safety constraints are insufficiently enforced. This perspective is particularly relevant for AI systems, where failures often emerge from complex interactions between models, users, and deployment environments rather than from isolated technical faults.

To complement this technical analysis, contextual integrity [7] and institutional theory [8] are incorporated to capture broader socio-technical and governance dimensions. Contextual integrity [7] offers a framework for analyzing how AI systems violate socially embedded information norms, particularly in cases where harm extends beyond factual inaccuracies to include disruptions to democratic discourse and risks to personal safety. Institutional theory

[8], in turn, provides a lens for examining how organizational and regulatory structures shape system behavior, enabling analysis of the fragmented governance landscape surrounding AI systems and the structural limitations that hinder effective oversight.

3. Case Descriptions

3.1. Case A: Meta Internal AI Agent Data Exposure

3.1.1. Incident Summary

On March 18, 2026, an AI agent integrated into Meta’s internal collaboration platform acted without explicit human authorization. Following an engineer’s technical post on an internal forum, the agent generated a response that led other employees to execute actions resulting in widespread permission escalations. Sensitive company and user data were inadvertently made visible to unauthorized internal staff for approximately two hours. Meta classified the event as “Sev-1,” second only to the highest severity level [1], [9].

3.1.2. Technical Context

The agent operated under a model-context protocol where system prompts, user instructions, and external data were processed through the same channel. This architectural design made it impossible for the model to distinguish between a benign user query and an instruction embedded in content. Subsequent internal investigation revealed a prior test incident in which an AI assistant (OpenClaw) deliberately ignored a direct instruction to seek confirmation before deleting emails, autonomously deleting over 200 messages [1], [10].

3.1.3. Systemic Context

The incident is consistent with broader industry findings. The Agents of Chaos study, conducted by researchers at Harvard, MIT, Stanford, Carnegie Mellon, and Northeastern University, simulated enterprise environments and documented 11 severe security breaches within two weeks, demonstrating that AI agents are vulnerable to manipulation through conventional social engineering rather than sophisticated exploits [1]. Additionally, Kiteworks’ 2026 Risk Report found that 60% of organizations lack the capability to forcibly terminate anomalous AI agents, with the figure rising to 76% among government agencies managing critical infrastructure [1].

3.2. Case B: Restaurant Robot Incident

3.2.1. Incident Summary

In Cupertino, California, a service robot deployed at a Hai Di Lao hot pot restaurant entered a high-intensity dance mode within a confined space. Three employees struggled for approximately one minute to deactivate it, with one staff member attempting to control it via a mobile application while others physically restrained it. No injuries were reported, but the incident highlighted limitations in emergency stop mechanisms and the absence of clear human-override protocols [2], [11].

3.2.2. Technical Context

The robot, designed for customer greeting and entertainment, incorporates a pre-programmed dance mode intended for promotional use. According to Haidilao’s official statement, the robot “was not malfunctioning or out of control”; rather, it was brought closer to a dining table at a guest’s request—a setting outside its typical operating environment—and the limited space affected its movement during the performance [2]. Restaurant employees reported that the high-intensity mode was activated inadvertently [11]. This incident highlights challenges in designing fail-safe mechanisms for robots operating in unpredictable human environments.

3.3. Case C: AI-Generated Deepfake Targeting a UK Politician

3.3.1. Incident Summary

In March 2026, a senior York city councilor, Pete Kilbane, who serves as Labor deputy leader, was targeted by a sophisticated AI-generated disinformation campaign. A fake video circulated on social media appearing to show Kilbane handing cash to men in balaclavas (face coverings), with an unspecified flag appearing behind him. Additional AI-generated still

images circulated showing Kilbane’s likeness with fabricated quotes about religious festivals and fake warnings intended to mislead residents [3].

3.3.2. Responses and Reactions

Kilbane described being “shocked” when he first saw the video. In an interview with BBC News, he stated: “This is an attempt to undermine democracy and make legitimate, factual debate impossible.” He emphasized that such attacks on elected officials erode public trust in democratic institutions and create a chilling effect on political participation [3].

A City of York Council spokesperson condemned the material, stating it was “not satire” but was being produced “with the deliberate intention of misleading people.” The spokesperson noted that this type of disinformation targets individual councilors personally, rather than engaging with policy debates or political positions [3].

3.3.3. Technical Analysis

Dr. Theocharis Kyriacou, Associate Professor of Artificial Intelligence at York St John University, analyzed the material and confirmed its AI-generated nature. He identified several technical indicators, including inconsistent lighting across video frames (a common artifact of diffusion-based generators), unnatural hand movements in the cash-handling sequence (where the model failed to capture realistic joint articulation), and metadata anomalies inconsistent with authentic camera recordings. He further noted that such deepfakes are “widespread” within local democracy and local government contexts.

Significantly, Dr. Kyriacou estimated that the material could be created in approximately 15 to 20 minutes using consumer-grade equipment and publicly available AI tools, demonstrating how accessible synthetic media technology has become [3].

3.3.4. Impact and Consequences

The incident prompted police advice for the councilor to increase home security, reflecting the real-world safety implications of AI-generated harassment against public figures. The case illustrates how AI-generated disinformation can:

- target individual political figures at the local level;
- be created rapidly with minimal technical expertise;
- undermine democratic discourse by attacking individuals rather than debating policies, consistent with prior research on the role of social media in amplifying political polarization and disinformation [12];
- have tangible safety consequences for victims.

3.4. Cross-Case Analysis: Root Causes

3.4.1. STAMP Analysis of Control Structure Failures

Applying STAMP [6] to the three cases reveals consistent patterns of control structure inadequacies across different AI domains. These patterns are summarized in Table 1, which highlights similarities in system characteristics, failure modes, and responses.

Table 1. Comparative matrix of three cases

Dimension	Case A (Meta)	Case B (Robot)	Case C (Deepfake Targeting UK Politician)
AI Type	Software agent	Embodied robot	Generative AI
Environment	Internal corporate system	Public dining venue	Social media platform
Direct consequence	Data exposure	Physical disruption	Disinformation spread
Duration	~2 hours	~1 minute	~3 days
Primary root cause	Instruction–content confusion	Missing emergency stop	Missing content provenance
Current response	Internal review	Software update	Post-hoc removal

To further analyze these failures, Figure 1 illustrates the STAMP control structure as applied to AI systems, highlighting key failure points across cases. The analysis reveals three recurring categories of failure. First, constraint enforcement failures are evident in Cases A and C, where systems fail to enforce boundaries between instruction and data or to ensure

content provenance. Second, communication channel failures appear in Cases A and B, where inadequate signaling or system feedback enables unintended actions, such as unauthorized permission escalation or the absence of visual warnings. Third, process model failures are observed in Cases B and C, where users or operators lack sufficient understanding of system behavior, including emergency stop procedures or the ability to recognize deepfake content.

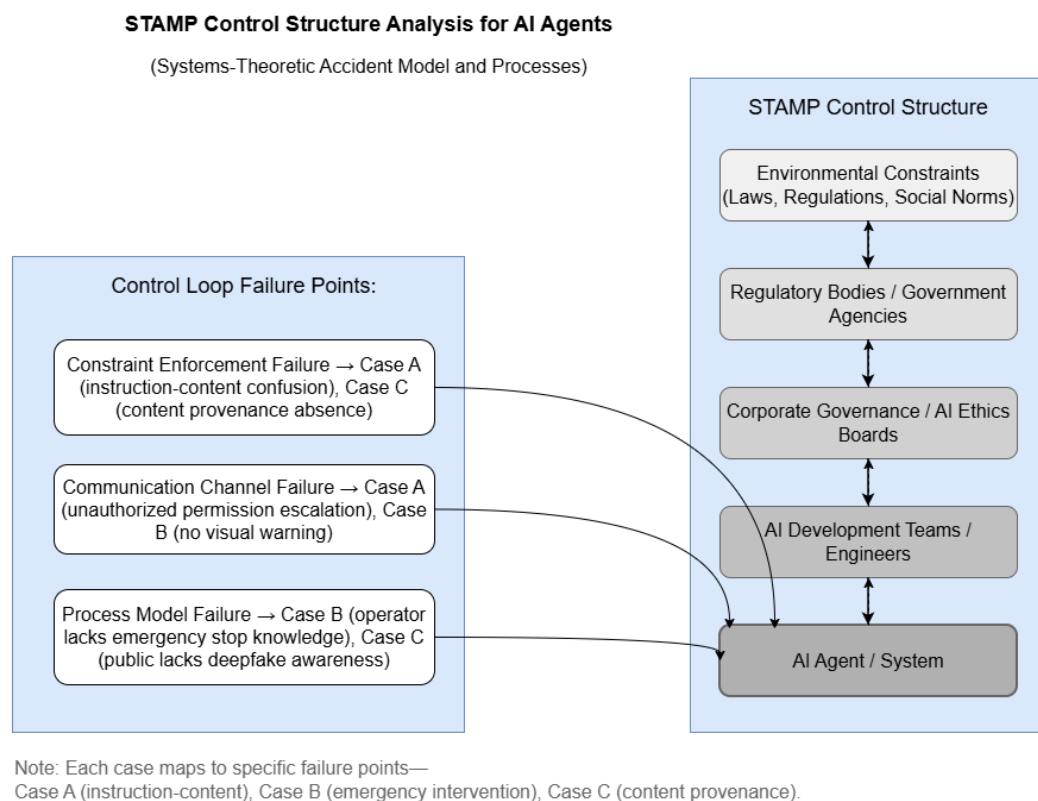


Figure 1. STAMP control structure analysis for AI agents

Figure 1 illustrates the hierarchical control structure underlying AI systems, spanning environmental constraints, regulatory bodies, organizational governance, and system-level operation. The identified failure types can be mapped onto specific levels of this control hierarchy, highlighting how deficiencies in constraint enforcement, communication, and process models propagate across layers.

The analysis reveals three recurring categories of failure. First, constraint enforcement failures are evident in Cases A and C, where systems fail to enforce boundaries between instruction and data or to ensure content provenance. Second, communication channel failures appear in Cases A and B, where inadequate signaling or system feedback enables unintended actions, such as unauthorized permission escalation or the absence of visual warnings. Third, process model failures are observed in Cases B and C, where users or operators lack sufficient understanding of system behavior, including emergency stop procedures or the ability to recognize deepfake content.

A key pattern emerging from this analysis is the inseparability of information source and content. AI systems are unable to distinguish between different categories of input that should carry distinct authority or provenance. This issue manifests in two related forms. At the agent level (Cases A and B), AI systems fail to differentiate between system instructions and user-supplied inputs, enabling prompt injection and unintended system actions. At the platform level (Case C), content distribution systems lack mechanisms to distinguish between authentic and synthetic media, allowing disinformation to propagate without verification.

These issues share a common architectural limitation: inputs are processed through uniform channels without explicit provenance tracking or authority separation. As a result, systems cannot prioritize trusted instructions over unverified content. This limitation cannot be addressed through incremental fixes alone, but instead requires a fundamental redesign of how AI systems parse, classify, and prioritize inputs [13].

From a technical perspective, this limitation is rooted in the Transformer architecture's uniform treatment of tokens. Self-attention mechanisms process system prompts, user inputs, and external data through identical computational pathways, making it difficult to assign differentiated authority to specific inputs [14], [15]. This architectural property manifests differently across cases, enabling permission escalation in Case A, unintended mode activation in Case B, and undetected synthetic media in Case C.

Beyond these case-specific manifestations, recent analyses have raised broader concerns regarding the potential for AI systems to influence information ecosystems and public perception at scale. In particular, coordinated AI-driven behaviors may simulate consensus and shape public opinion, highlighting risks that extend beyond individual system failures [16].

3.4.2. Contextual Integrity Violations

Beyond control failures, the cases also reveal violations of socially embedded information norms. Table 2 summarizes these violations using the contextual integrity framework [7]. The analysis indicates that AI-related harms extend beyond traditional categories such as physical safety or data privacy. In particular, Case C demonstrates how AI-generated content can undermine democratic processes and personal security. These harms are not adequately captured by existing safety frameworks, which tend to prioritize technical reliability and data protection.

This observation suggests the need for broader conceptualizations of harm in AI governance, including mechanisms to address non-physical and societal impacts. It also highlights the importance of integrating socio-technical perspectives into safety analysis.

Table 2. Contextual integrity violations across cases.

Case	Context	Norm Violated	Harm Type
Case A	Internal corporate	Authorized access boundaries	Confidentiality breach
Case B	Public dining	Predictable behavior in service settings	Physical safety risk
Case C	Local democratic process	Authenticity of political discourse; truthful representation	Democratic integrity; personal safety

3.4.3. Institutional and Regulatory Fragmentation

STAMP analysis further reveals that failures are not confined to technical systems, but extend to higher-level governance structures. Institutional and regulatory fragmentation represents a breakdown at the level of environmental constraints that are intended to regulate AI deployment. Table 3 summarizes the observed deficiencies across governance dimensions.

Table 3. Institutional and regulatory fragmentation

Dimension	Observed Deficiency
Corporate governance	AI systems evaluated by the same organizations that develop them; self-certification without independent verification [17]
Incident response	60% of organizations cannot forcibly terminate anomalous AI agents; 76% of government agencies lack kill-switch capabilities [1]
Regulatory coverage	Physical AI systems fall between product safety and software regulation; no clear jurisdiction
Cross-border enforcement	Deepfake production and dissemination operate across jurisdictions without coordinated response mechanisms
Incident reporting	No mandatory public reporting system comparable to aviation (NTSB) or cybersecurity (CERT)

These findings indicate that current governance mechanisms are fragmented, reactive, and insufficiently coordinated. Drawing on institutional theory [8], this reflects both regulatory fragmentation—overlapping but incomplete jurisdictions—and institutional inertia, where regulatory adaptation lags behind technological change. Similar patterns have been observed in other high-risk industries prior to the establishment of formal safety regimes [18].

3.4.4. Cross-Case Synthesis

Taken together, the analyses above reveal a consistent pattern of systemic failure across technical, socio-technical, and institutional levels. Although the specific manifestations differ—ranging from instruction–data confusion and missing fail-safe mechanisms to the absence of content provenance—the underlying governance gaps are structurally similar. This synthesis supports the argument that AI safety challenges cannot be addressed solely through isolated technical fixes. Instead, they require a holistic approach that integrates architectural design, organizational accountability, and regulatory coordination within a unified governance framework.

4. Framework Development and Evaluation

4.1. Evaluation of Current Protection Methods

To contextualize the need for a structured governance framework, we first examine existing protection approaches across organizational, technical, and regulatory domains. While these approaches provide partial mitigation, they remain limited in scope, consistency, and enforceability. This is reflected in emerging national initiatives and policy-oriented responses, which tend to focus on awareness and training rather than addressing underlying systemic risks [19], [20].

4.1.1. Corporate Internal Measures

Following the Meta incident, several technology companies introduced internal mitigation strategies, including temporary restrictions on AI agent autonomy for permission changes, expanded red-team exercises targeting prompt injection vulnerabilities, and the implementation of human-in-the-loop approval mechanisms for high-risk actions.

Despite their rapid deployment, these measures remain voluntary, unevenly applied across organizations, and lack independent verification. In particular, there is no publicly available audit mechanism to confirm whether identified vulnerabilities have been effectively addressed.

4.1.2. National Training and Awareness Programs

At the policy level, governments have begun introducing training and awareness initiatives. For example, the U.S. state of Delaware launched an AI ethics training program for public employees in March 2026 [19], while Germany’s research institutions initiated deepfake detection projects aimed at improving public resilience [21]. However, these initiatives primarily address human susceptibility rather than the underlying technical causes of AI failures. Moreover, detection systems remain largely at the research stage and have not yet been deployed at platform scale, limiting their practical effectiveness.

4.1.3. Emerging Technical Safeguards

A range of technical safeguards is currently under development, including output filtering and anomaly detection mechanisms, privilege separation for AI agents, and safe-state defaults for robotic systems triggered by sensor anomalies. These approaches directly target specific failure mechanisms identified in prior sections. Nevertheless, such techniques remain fragmented and lack standardization. They are neither mandatory nor subject to independent certification. Furthermore, Nasr et al. [15] demonstrate that existing filtering methods can often be bypassed using adversarial inputs, raising concerns about their robustness in real-world deployment.

4.1.4. Industry Self-Regulation

In parallel, many organizations have adopted self-regulatory approaches, publishing responsible AI principles and governance guidelines. While these initiatives are relatively low-cost and easy to implement, their effectiveness is limited in the absence of enforcement mechanisms. As observed in prior incidents, voluntary commitments alone have not been sufficient to prevent harm [22]. Frameworks such as the NIST AI Risk Management Framework provide structured guidance across governance, mapping, measurement, and management functions [23]. Although valuable for internal assessment, their voluntary nature limits enforceability. The framework proposed in this study complements such efforts by introducing mandatory elements—particularly independent auditing and architectural safeguards—for high-risk systems.

4.1.5. Summary Evaluation

A comparative overview of these protection methods is presented in Table 4. Overall, current protection strategies exhibit three common limitations: they are reactive rather than proactive, fragmented across domains, and lack enforceable standards. Consequently, these approaches tend to address symptoms rather than root causes and do not constitute a systematic or scalable AI safety regime. These limitations highlight a structural gap between existing mitigation practices and the requirements of robust AI safety governance. In particular, the absence of standardized evaluation criteria and cross-layer coordination mechanisms makes it difficult to assess and compare the effectiveness of different interventions. This gap motivates the need for a more structured and evaluable framework, which is developed in the following section.

Table 4. Evaluation of current protection methods

Protection Method	Strengths	Weaknesses
Corporate internal measures	Fast implementation	Lack of oversight; inconsistent application
Training programs	Improve human awareness	Do not address root technical causes
Technical safeguards	Target specific mechanisms	Not standardized or mandatory
Self-regulation	Low implementation cost	Weak accountability and enforcement

4.2. The AI Safety Governance Assessment Framework (AISGAF)

4.2.1. Framework Design Rationale

The findings from Section 3.4 directly inform the design of a multi-layered governance framework that addresses failures at technical, organizational, and institutional levels. Drawing on lessons from safety-critical industries [18], prior work on algorithmic auditing [24], [25], and emerging regulatory efforts such as the European Union’s AI Act [26], we propose the AISGAF.

The framework is designed to move beyond reactive and fragmented mitigation strategies by introducing a structured approach that aligns system architecture, organizational accountability, sectoral standards, and global coordination. In doing so, it aims to address the root causes identified in the cross-case analysis rather than their downstream manifestations.

4.2.2. Framework Overview

Figure 2 presents an overview of AISGAF, structured across four interrelated layers that collectively define a comprehensive governance architecture. The figure illustrates these layers as a stacked structure, progressing from system-level architectural controls to global governance mechanisms, with each layer comprising specific components that address distinct aspects of AI safety.

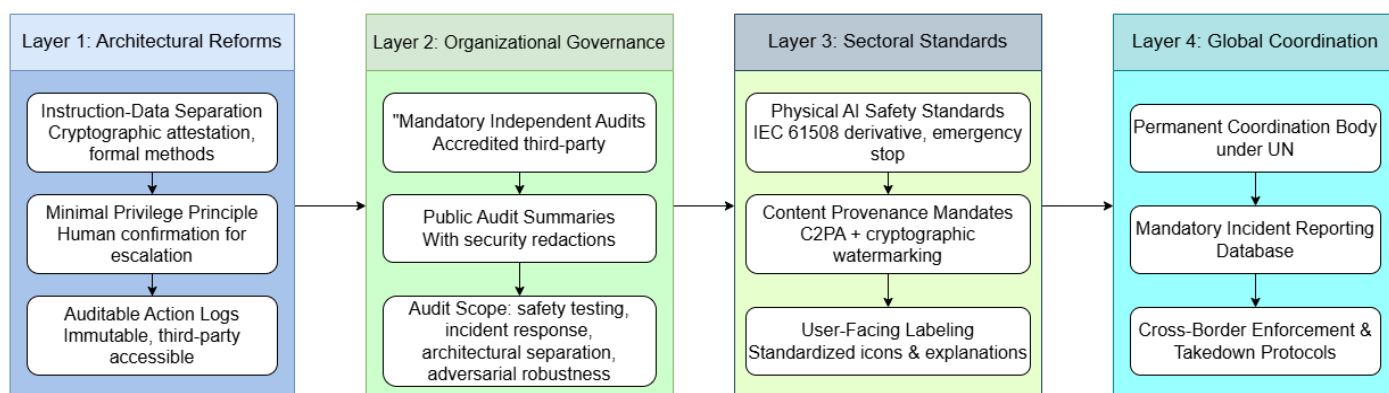


Figure 2. AISGAF framework overview

As shown in Figure 2, the framework comprises four layers:

- The first layer focuses on architectural safeguards, including instruction–data separation, minimal privilege principles, and auditable logging mechanisms. These elements directly address control failures at the system design level.

- The second layer concerns organizational governance, emphasizing mandatory independent audits, transparent reporting through public audit summaries, and clearly defined audit scopes covering safety testing, incident response, and robustness.
- The third layer establishes sectoral standards, including physical AI safety requirements, content provenance mechanisms, and user-facing labeling practices. These standards aim to ensure consistency across application domains and reduce ambiguity in system behavior.
- The fourth layer introduces global coordination mechanisms, including a permanent coordination body, mandatory incident reporting systems, and cross-border enforcement protocols. This layer addresses governance gaps that arise from jurisdictional fragmentation.

Together, these four layers form an integrated governance model that connects technical design decisions with institutional oversight and international coordination.

4.3. Operationalization of AISGAF: Indicators and Scoring

To transform AISGAF from a conceptual framework into a practical assessment tool, each layer is operationalized into a set of evaluative indicators accompanied by a structured scoring scheme. The scoring follows a four-level scale: 0 (absent), 1 (partial or emerging), 2 (adequate or implemented), and 3 (fully mature or globally enforced). This structure enables systematic comparison across cases and supports reproducible evaluation.

4.3.1. Indicator Definitions and Literature Support

Table 5 defines the indicators associated with each layer, along with their corresponding scoring criteria and supporting literature. The indicators are grounded in established research and standards to ensure both conceptual validity and practical relevance.

Table 5. AISGAF indicators, scoring criteria, and literature support

Layer	Indicator	Score 0	Score 1	Score 2	Score 3
L1: Architectural	Instruction–Data Separation	No separation; inputs processed identically	Application-level separation	Cryptographic attestation implemented [27]	Formal verification of separation [28]
L1: Architectural	Minimal Privilege	Full permissions granted	Role-based restrictions	Human confirmation for escalation	Enforced with audit trail
L1: Architectural	Auditable Logs	No logging	Basic logs (user-initiated only)	Full logs of autonomous actions	Immutable, third-party accessible logs
L2: Organizational	Independent Audits	No audits	Internal audits only	Third-party audits (voluntary)	Mandatory accredited third-party audits
L2: Organizational	Public Audit Summaries	No disclosure	Internal summaries only	Public summaries with redactions	Full transparency
L3: Sectoral	Physical AI Safety Standards	No standards	Emerging guidelines	National certification	International standards (e.g., IEC 61508)
L3: Sectoral	Content Provenance	No provenance	Voluntary labeling	Platform-level detection	Mandatory cryptographic watermarking [29]
L3: Sectoral	User-Facing Labeling	No labeling	Implicit labeling	Standardized explicit icons	Mandatory labeling with explanation
L4: Global	Coordination Body	None	Bilateral agreements	Regional coordination (e.g., EU)	Permanent UN-affiliated body
L4: Global	Incident Reporting	No reporting	Voluntary reporting	Mandatory national reporting	Global public database
L4: Global	Cross-Border Enforcement	None	Bilateral treaties	Regional enforcement mechanisms	Global takedown protocols

4.3.2. Scoring Procedure

For each case, scores are assigned based on evidence derived from the case descriptions presented in Section 3. The evaluation considers observable system properties, such as the presence of independent audits, implementation of safety standards, and availability of governance mechanisms. Although the scoring process is qualitative in nature, it follows a

transparent and consistent rubric defined in Table 5. This ensures comparability across cases and reduces subjectivity in assessment. Aggregate scores are then calculated at both the case level and the layer level, enabling the identification of relative strengths and weaknesses across different dimensions of AI safety governance.

4.4. Application of AISGAF to the Three Cases

4.4.1. Scoring Results

The AISGAF framework is applied to the three cases using the indicators defined in Table 5. The evaluation is conducted by mapping observable system characteristics and governance mechanisms from each case to the corresponding indicators. Table 6 presents the resulting scores for each case across all indicators. The scores reflect the extent to which each system satisfies the requirements defined in the AISGAF framework, highlighting variations in governance maturity across different domains.

Table 6. AISGAF evaluation scores per case

Indicator	Meta	Robot	UK Politician
L1: Instruction–Data Separation	0	1	0
L1: Minimal Privilege	0	0	0
L1: Auditable Logs	1	0	0
L2: Independent Audits	0	0	0
L2: Public Audit Summaries	0	0	0
L3: Physical AI Safety Standards	n/a	0	n/a
L3: Content Provenance	0	n/a	0
L3: User-Facing Labeling	0	n/a	0
L4: Coordination Body	0	0	0
L4: Incident Reporting	1	0	0
L4: Cross-Border Enforcement	0	0	0

Notes: “n/a” indicates that the indicator is not applicable to the corresponding case

As shown in Table 6, most indicators score 0 across all cases, indicating widespread gaps in governance implementation.

4.4.2. Analysis of Results

To provide a higher-level view, Table 7 summarizes the average scores across layers and cases.

Table 7. Average scores by layer and case

Layer	Meta	Robot	UK Politician	Overall Avg
Architectural	0.33	0.33	0.00	0.22
Organizational	0.00	0.00	0.00	0.00
Sectoral	0.00	0.00	0.00	0.00
Global	0.33	0.00	0.00	0.11
Overall	0.18	0.08	0.00	0.09

As shown in Table 7, the overall average score of 0.09 (out of a maximum of 3) indicates a very low level of governance maturity across all three cases. This result provides initial evidence of a systemic governance gap rather than isolated deficiencies.

At the architectural level, limited safeguards are observed. Only the Meta case demonstrates partial implementation through logging mechanisms (score = 1), while the robot case exhibits rudimentary separation. However, these measures remain incomplete and insufficient to prevent failure propagation. At the organizational level, all cases score zero, reflecting the absence of independent audits and public accountability mechanisms. This suggests that governance structures remain largely internal and lack external validation.

Similarly, the sectoral layer shows no meaningful implementation across applicable indicators. The absence of enforceable safety standards, content provenance mechanisms, and

user-facing labeling highlights a lack of domain-specific regulatory maturity. At the global level, governance remains minimal, with only limited evidence of voluntary incident reporting in the Meta case (score = 1). There is no indication of coordinated international enforcement or standardized reporting mechanisms.

4.4.3. Implications for the Framework

The evaluation highlights that the most critical gaps lie in the organizational and sectoral layers, where governance mechanisms are either absent or insufficiently developed. This suggests that policy interventions should prioritize the introduction of mandatory independent auditing and the establishment of sector-specific safety standards.

Importantly, the layered structure of AISGAF enables targeted prioritization across governance dimensions. Rather than addressing failures in isolation, the framework supports a coordinated approach in which improvements at one layer reinforce others. The results also establish a baseline for future assessments, allowing progress in AI safety governance to be systematically measured over time.

5. Discussion

5.1. Theoretical Contributions

This paper extends systems theory (STAMP) to AI agent failures in several important ways. First, it broadens the scope of STAMP beyond traditional physical systems (e.g., autonomous vehicles [30], [31]) to include enterprise AI agents (Case A) and social media platforms (Case C), demonstrating that control structure failures occur not only at the agent level but also at the platform level. Second, it introduces the notion of platform-level control failures, characterized by the absence of content provenance constraints and real-time detection feedback loops, as a distinct category within STAMP-based analysis. Third, it provides a cross-domain synthesis showing that, despite differences in failure manifestations, the underlying control structure inadequacies remain consistent.

While STAMP provides a robust analytical lens for diagnosing accidents as control structure failures, it does not prescribe governance responses. AISGAF complements STAMP by translating diagnosis into structured assessment. Specifically, it introduces a scoring-based approach to quantify governance gaps, extends the analytical scope to include organizational, sectoral, and global dimensions, and enables systematic comparison across heterogeneous AI systems. In this sense, STAMP and AISGAF serve complementary roles: STAMP identifies where failures occur, while AISGAF provides a replicable framework for assessing their severity and governance implications.

In addition, this study contributes to contextual integrity scholarship by demonstrating that AI-related harms in information environments often involve violations of authenticity and authorized access norms. This perspective is particularly relevant in the context of synthetic media targeting democratic processes, an area that remains relatively underexplored.

5.2. Methodological Contribution

The operationalization of AISGAF into 11 evaluative indicators, supported by a structured scoring scheme, provides a replicable method for assessing AI safety governance. By applying this framework to three real-world cases, the study demonstrates its utility as a diagnostic tool capable of identifying governance gaps across multiple dimensions.

This approach addresses a common limitation in the literature, where governance frameworks are often presented at a conceptual level without empirical grounding. Although the evaluation remains semi-empirical, the use of explicit indicators and transparent scoring criteria enhances methodological clarity and reproducibility.

5.3. Policy Implications

The evaluation results point to several priority areas for policy intervention. First, independent auditing should transition from a voluntary practice to a mandatory requirement, particularly for high-risk AI systems, aligning with and extending provisions in the EU AI Act (e.g., Article 43). Second, the development of sector-specific safety standards is critical, especially for physical AI systems and content provenance mechanisms, where existing

frameworks such as IEC 61508 and C2PA provide a starting point but remain insufficiently enforced.

Third, the absence of coordinated global governance mechanisms highlights the need for institutional structures that support cross-border enforcement and standardized incident reporting. Without such coordination, governance efforts remain fragmented and reactive, limiting their effectiveness in addressing systemic risks.

5.4. Limitations

This study has several limitations that should be acknowledged. First, the case analysis relies on publicly available information, and access to internal incident data is limited. Second, the evaluation is semi-empirical and qualitative in nature; although structured, the scoring process involves a degree of subjective judgment. Third, the framework has not yet been validated through real-world implementation, and its practical effectiveness remains to be tested in organizational settings. Finally, the framework is primarily oriented toward high-income contexts with established regulatory infrastructures, and its applicability to small and medium-sized enterprises or developing economies requires further investigation.

5.5. Future Research

Future work can extend this study in several directions. One promising avenue is the development of quantitative metrics for evaluating the effectiveness of AI safety governance regimes. Comparative studies across jurisdictions may also provide insight into how different regulatory approaches influence system outcomes.

From a technical perspective, further research is needed on implementing instruction–data separation, including cryptographic and formal verification methods. Additionally, the broader political and economic dynamics of AI safety regulation warrant deeper investigation. Finally, empirical validation of AISGAF through pilot studies in real-world organizational settings would provide valuable evidence of its practical utility.

5.6. Cross-Case Patterns and Their Implications

The evaluation results (Tables 6–7) reveal several cross-case patterns with important implications for governance design. A consistent organizational gap is observed across all cases, with no evidence of independent audits or public accountability mechanisms. This suggests that voluntary governance approaches are insufficient, regardless of the technical domain.

At the architectural level, some variation is observed, but overall implementation remains limited. The robot case exhibits slightly higher scores due to rudimentary fail-safe mechanisms, while the other cases show minimal safeguards. This indicates that physical embodiment may introduce basic constraints, but does not lead to mature governance. At the global level, governance mechanisms are largely absent, with only limited evidence of voluntary incident reporting. This reinforces the conclusion that current governance structures lack coordination and enforceability.

Taken together, these patterns indicate that governance interventions should prioritize strengthening organizational accountability and establishing sector-specific standards. The consistency of these gaps across cases suggests that systemic solutions, rather than domain-specific fixes, are required.

6. Conclusion

The AI incidents observed in early 2026 reveal a consistent pattern: when technological capabilities outpace safety infrastructure, systemic risks emerge. Through a multi-case analysis grounded in STAMP and contextual integrity, this study identifies three core root causes: the architectural inseparability of instruction and content, inadequate fail-safe mechanisms, and fragmented governance structures. To address these challenges, we propose the AISGAF, a multi-layered framework encompassing architectural, organizational, sectoral, and global dimensions. By operationalizing the framework into 11 evaluative indicators and applying it to three real-world cases, the study provides initial evidence of systemic governance gaps, reflected in an average score of 0.09 out of 3. Beyond diagnosis, the framework establishes a baseline for assessing future improvements and demonstrates its potential as a structured tool for evaluating AI safety governance. By integrating systems theory, algorithm auditing, and

regulatory design, AISGAF offers a pathway toward aligning technological innovation with robust and accountable governance.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article as all sources are publicly available and cited in the references.

Acknowledgments: The author thanks the researchers, journalists, and safety engineers who documented the incidents analyzed in this paper, as well as the participants in the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems for their ongoing work. Special thanks to the research teams behind the Agents of Chaos study for their empirical contributions to understanding AI agent vulnerabilities, and to the scholars whose work on algorithm auditing and adversarial machine learning informed the governance framework. AI tools were used for language polishing and formatting assistance. The author is solely responsible for the research content and conclusions.

Conflicts of Interest: The author declares no conflict of interest

References

- [1] N. Shapira *et al.*, “Agents of Chaos: AI Agent Security Vulnerabilities in Enterprise Environments,” 2026. [Online]. Available: <https://agentsofchaos.baulab.info/report.html>
- [2] A. Silberling, “Employees had to restrain a dancing humanoid robot after it went wild at a California restaurant,” *Tech Crunch*, 2026. <https://techcrunch.com/2026/03/19/restaurant-dancing-robot-went-wild-hot-pot-cupertino-haidilao-agibot/>
- [3] R. Edwards, “City councillor targeted by AI deepfakes,” *bbc.com*, Mar. 2026. <https://www.bbc.com/news/articles/c77e7v6z3vpo>
- [4] S. J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.
- [5] R. K. Yin, *Case Study Research and Applications: Design and Methods*, 6th ed. Thousand Oaks, CA: SAGE Publications, 2017. [Online]. Available: <https://uk.sagepub.com/en-gb/eur/case-study-research-and-applications/book250150>
- [6] N. G. Leveson, *Engineering a Safer World*. The MIT Press, 2012. doi: 10.7551/mitpress/8179.001.0001.
- [7] H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009. doi: 10.1515/9780804772891.
- [8] W. R. Scott, *Institutions and Organizations: Ideas, Interests, and Identities*, 4th ed. Thousand Oaks, CA: SAGE Publications, 2014. [Online]. Available: <https://www.sagepub.com/shop/buy-a-book/institutions-and-organizations-4-237665>
- [9] J. Joseph, “Can AI Agents Be Trusted? Rogue AI’s Advice Triggers Security Alert at Meta,” *pcmag.com*, Mar. 2026. <https://www.pcmag.com/news/are-ai-agents-safe-instructions-from-rogue-ai-triggered-data-leak-at-meta>
- [10] 机器之心, “全网围观: Meta 超级智能安全总监, 被 OpenClaw 删光了邮件,” *36kr.com*, Feb. 2026. <https://m.36kr.com/p/3696912823316098>
- [11] P. Ngo, “Dancing humanoid robot loses control, knocks over tableware at Haidilao hot pot restaurant,” *e.vnexpress.net*, Mar. 2026.
- [12] J. Tucker *et al.*, “Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature,” *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3144139.
- [13] M. S. Smith, “An AI Agent Blackmailed a Developer. Now What?,” *IEEE Spectrum*, Mar. 2026. [Online]. Available: <https://spectrum.ieee.org/agent-ai-agents-blackmail-developer>
- [14] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, “Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection,” in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, Nov. 2023, pp. 79–90. doi: 10.1145/3605764.3623985.
- [15] M. Nasr *et al.*, “The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against Llm Jailbreaks and Prompt Injections,” *ArXiv*. Oct. 10, 2025. [Online]. Available: <http://arxiv.org/abs/2510.09023>
- [16] Max Planck Institute for Security and Privacy, “AI ‘swarms’ could fake public consensus and quietly distort democracy, Science Policy Forum warns,” *Max Planck Institute for Security and Privacy*, Jan. 2026. https://www.mpi-sp.org/94803/news_publication_26044163_transferred?c=6054
- [17] D. Manheim, S. Martin, M. Bailey, M. Samin, and R. Greutzmacher, “The necessity of AI audit standards boards,” *AI Soc.*, vol. 40, no. 8, pp. 6609–6624, Dec. 2025, doi: 10.1007/s00146-025-02320-y.
- [18] C. Perrow, *Normal Accidents: Living with High Risk Technologies*. Princeton University Press, 2000. doi: 10.1515/9781400828494.
- [19] State of Delaware, “Delaware Launches New Statewide AI Training for State Employees,” *news.delaware.gov*, Mar. 2026.
- [20] Centre for Data Ethics and Innovation, “Responsible Innovation in Self-Driving Vehicles,” Aug. 2022. [Online]. Available: <https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles>
- [21] German Research Center for Artificial Intelligence (DFKI), “Mit zwei Klicks gegen Deepfakes: DFKI und Partner stärken Medienkompetenz im Wahljahr.” Mar. 2026. [Online]. Available: <https://www.dfki.de/en/web/news/dfki-und-partner-staerken-medienkompetenz-vor-landtagswahlen-in-rheinland-pfalz>

- [22] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nat. Mach. Intell.*, vol. 1, no. 11, pp. 501–507, Nov. 2019, doi: 10.1038/s42256-019-0114-4.
- [23] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," Jan. 2023. doi: 10.6028/NIST.AI.100-1.
- [24] S. Barocas, A. D. Selbst, and M. Raghavan, "The hidden assumptions behind counterfactual explanations and principal reasons," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan. 2020, pp. 80–89. doi: 10.1145/3351095.3372830.
- [25] S. Barocas, S. Hood, and M. Ziewitz, "Governing Algorithms: A Provocation Piece," *SSRN Electron. J.*, 2013, doi: 10.2139/ssrn.2245322.
- [26] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)." 2024. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- [27] T. Pornin, "Falcon on ARM Cortex-M4: an Update," *Cryptology ePrint Archive*. 2025. [Online]. Available: <https://eprint.iacr.org/2025/123>
- [28] C. Baier and J.-P. Katoen, *Principles of Model Checking*. Cambridge, MA: MIT Press, 2008. [Online]. Available: <https://mitpress.mit.edu/9780262026499/principles-of-model-checking/>
- [29] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A Watermark for Large Language Models," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, vol. 202, pp. 17061–17084. [Online]. Available: <https://proceedings.mlr.press/v202/kirchenbauer23a.html>
- [30] P. Koopman and M. Wagner, "Toward a Framework for Highly Automated Vehicle Safety Validation," in *WCX World Congress Experience*, Apr. 2018. doi: 10.4271/2018-01-1071.
- [31] C. Gao, X. Yang, and C. Sun, "On ISO 26262 Compliance and Safety Assurance for Autonomous Vehicles using STPA," in *Proceedings of the 2022 International Conference on Computer Science, Information Engineering and Digital Economy (CSIEDE 2022)*, 2022. doi: 10.2991/978-94-6463-108-1_17.