Empirically Derived AI Harm Profiles and the Case for Context-Sensitive Moral Responsibility

Ummara Mumtaz
university of the cumberlands
Rabi Noor
Summaya Mumtaz
summaya.mz@gmail.com

university of the cumberlands

Research Article

Keywords: Artificial intelligence governance, AI incident analysis, Harm profiles, Multiple Correspondence Analysis, AI risk management, Responsible AI, Generative AI, AI governance framework

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10542055/v1

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) systems are increasingly deployed in high-stakes domains where automated decisions have significant ethical consequences, yet existing AI governance frameworks largely rely on broad principles that are difficult to operationalize across diverse contexts. This study argues that documented AI incidents represent an underutilized empirical resource for developing context-sensitive governance by revealing recurring patterns of harm, responsibility, and rights-related risks. Drawing on contextual integrity, procedural justice, and responsibility-gap theory, we analyze 849 incidents from the AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) Repository using Multiple Correspondence Analysis (MCA) and Ward hierarchical clustering. The findings identify two underlying dimensions of AI-related harm: an embodiment axis, distinguishing systems with direct physical consequences from information-processing applications, and a rights-versus-content axis, separating identification and monitoring systems from generative AI. These dimensions produce three distinct harm profiles: Autonomous/Embodied AI, Vision Recognition, and Generative AI, each characterized by unique ethical risks and responsibility structures. Predictive Monitoring did not emerge as an independent profile but instead functioned as a cross-cutting capability whose ethical implications depend on the deployment context. Building on these results, we propose a harm-profile governance framework that translates recurring incident patterns into profile-specific monitoring indicators, audit triggers, evidence requirements, and differentiated accountability mechanisms.

1. Introduction

Artificial intelligence (AI) systems are increasingly deployed in domains where automated decisions have significant ethical consequences, including healthcare, finance, transportation, employment, education, and public administration. In these settings, AI failures can affect safety, civil rights, access to services, economic opportunity, privacy, and public trust. As AI becomes more deeply embedded in institutional decision-making, effective governance requires more than broad ethical principles. It also requires understanding how different AI systems fail in practice, the types of harms they produce, and the responsibilities these failures create.

AI harms are increasingly recognized as sociotechnical rather than purely technical failures. They arise from the interaction between system design, deployment context, organizational practices, and existing social structures (Shelby et al., 2023). These harms may undermine privacy, fairness, autonomy, and other context-dependent ethical values (Nissenbaum, 2010; Binns, 2018; Mittelstadt et al., 2016). Recent research also highlights that AI harms are defined inconsistently across governance frameworks, limiting the development of effective oversight and emphasizing the need for systematic classification (Abercrombie et al., 2024; Slattery et al., 2026). Incident-based studies suggest that documented AI failures provide valuable empirical evidence of recurring patterns across technologies, application domains, and risk categories, offering a foundation for more context-sensitive and evidence-based AI governance (McGregor, 2021; Yampolskiy, 2018).

Despite these advances, an important gap remains between empirical analyses of AI incidents and normative approaches to AI governance. Existing frameworks define AI risks and ethical principles at a high level but provide limited guidance on how specific combinations of AI technology, deployment context, system function, and risk type produce distinct patterns of harm. Consequently, principles such as fairness, transparency, accountability, and safety are difficult to operationalize in context-specific governance, reinforcing concerns that AI ethics often lacks practical applicability (Hagendorff, 2020). This study addresses this gap by examining whether documented AI incidents exhibit recurring harm profiles that can inform a more context-sensitive and empirically grounded approach to governance. Rather than treating incidents as isolated events, we analyze how AI modality, deployment domain, system function, and risk category co-occur across documented cases to identify recurring patterns of harm.

Our approach is informed by three complementary theoretical perspectives. Contextual integrity argues that the ethical implications of AI systems depend on the social context in which they operate, making harms inseparable from context-specific norms (Nissenbaum, 2010). Procedural justice emphasizes that the legitimacy of automated decision-making depends on the fairness of decision processes, which varies across applications and deployment settings (Binns, 2018). Finally, the responsibility-gap literature argues that moral and institutional responsibility for AI harms cannot be determined through abstract principles alone but must consider the specific characteristics of the technology and its context of use (Matthias, 2004; Coeckelbergh, 2020; Santoni de Sio & Mecacci, 2021). Together, these perspectives support an AI governance approach that aligns oversight and accountability with recurring patterns of harm observed in real-world AI incidents.

The paper makes three contributions. Empirically, it uses documented incident data to identify structured patterns in AI-related harm. Normatively, it interprets these patterns as ethically significant configurations that clarify how contextual integrity, procedural justice, and moral responsibility apply to different classes of AI systems. Practically, it proposes a harm-profile approach to AI governance in which oversight is organized around the harms that systems are likely to produce, the contexts in which those harms emerge, and the differentiated ethical and institutional responsibilities required to monitor, mitigate, and revise oversight practices after deployment.

2. Background and Literature Review

Recent AI governance research emphasizes that AI systems should be classified according to their operational characteristics and deployment contexts rather than by technology alone. Governance frameworks have evolved from binary and risk-based approaches toward multidimensional models that consider factors such as intended purpose, deployment environment, stakeholders, input data, and system outputs (Mökander et al., 2023; Hupont et al., 2024). These studies demonstrate that AI risks emerge not only from model design but also from how systems are implemented and used in practice.

This perspective is reinforced by research on AI risk and impact assessment, which characterizes AI harms as the product of interacting technical, organizational, and contextual factors. Recent studies argue that effective governance requires evaluating multiple dimensions including fairness, robustness, explainability, privacy, and reliability rather than relying solely on traditional measures of risk probability and severity (Stahl et al., 2023; Zanotti et al., 2024; Novelli et al., 2024; Schmitz et al., 2025). Consequently, governance has increasingly been viewed as a lifecycle process that extends beyond system development to encompass deployment, monitoring, mitigation, and continuous oversight (Monteiro & Singh, 2025; Schuett, 2024).

The growing emphasis on lifecycle governance reflects the recognition that AI systems operate within dynamic technical and social environments. System performance, data distributions, and operational conditions change over time, creating risks that often emerge only after deployment (National Institute of Standards and Technology, 2023; Raji et al., 2020; Sculley et al., 2015). Regulatory initiatives such as the EU AI Act similarly require ongoing transparency, human oversight, post-market monitoring, and incident reporting, highlighting that governance must continue throughout the operational life of an AI system (European Parliament and Council of the European Union, 2024). However, organizations continue to face challenges in translating high-level ethical principles into consistent governance practices because of varying definitions, fragmented standards, and limited operational guidance (Cousineau et al., 2025; Kemmerzell et al., 2025).

To address these challenges, recent research has focused on developing systematic taxonomies of AI risks and harms. Existing frameworks classify AI risks using different conceptual structures, often leading to inconsistent terminology across governance initiatives (Slattery et al., 2026). Similarly, harm-centered taxonomies describe AI harms as sociotechnical phenomena that vary across stakeholders, application domains, and governance requirements (Shelby et al., 2023; Abercrombie et al., 2024). These studies highlight the need for governance approaches that account for the diversity of AI harms rather than treating them as a single category.

Incident-based research provides an empirical foundation for this objective. Analyses of documented AI failures show that harms frequently arise through recurring mechanisms such as biased data, inadequate validation, distribution shifts, and insufficient robustness (Yampolskiy, 2018). AI incident repositories further enable systematic examination of these failures across technologies and application domains, revealing recurring patterns involving autonomous vehicles, facial recognition, generative AI, social media platforms, and algorithmic decision systems (McGregor, 2021; Raghupathi et al., 2026). These findings suggest that incident data can reveal structural relationships among AI technologies, deployment contexts, and associated harms that are not apparent when incidents are considered individually.

Despite these advances, an important gap remains. Existing governance frameworks identify risks and ethical principles but provide limited guidance on how recurring combinations of AI modality, deployment domain, system function, and risk category produce distinct governance requirements.

Consequently, governance mechanisms such as auditing, impact assessment, and human oversight often remain generic rather than being tailored to the specific harms associated with different classes of AI systems (Costanza-Chock et al., 2022; Metcalf et al., 2021; Mökander, 2023). Incident reporting similarly captures what occurred without systematically identifying recurring patterns that could inform differentiated governance strategies (Dixon & Frase, 2025).

This study addresses that gap by treating documented AI incidents as empirical evidence of recurring harm profiles. Using Multiple Correspondence Analysis (MCA) and hierarchical clustering, we examine associations among AI modality, deployment domain, system function, and risk category to identify recurring configurations of AI-related harm (Greenacre, 2017; Le Roux & Rouanet, 2010). We argue that these harm profiles provide an empirical basis for context-sensitive AI governance by linking observed patterns of harm to differentiated oversight mechanisms, monitoring requirements, and accountability structures. Rather than governing AI solely through broad ethical principles or technology categories, this approach supports governance that is calibrated to the types of harms AI systems are most likely to produce.

2.1 Ethical foundations for harm-based governance

A harm-profile approach to AI governance requires a normative basis for explaining why recurring incident patterns are ethically significant and how responsibility should be assigned. This study draws on three complementary perspectives: contextual integrity, procedural justice, and responsibility-gap theory.

First, contextual integrity argues that the ethical evaluation of AI systems depends on the social context in which they operate rather than on universal rules alone (Nissenbaum, 2010). Information practices are judged according to context-specific norms, meaning that the same AI capability such as identification, prediction, or content generation may be appropriate in one setting but ethically problematic in another. Consequently, recurring incident patterns can be interpreted as evidence of repeated violations of context-specific norms.

Second, procedural justice emphasizes that the legitimacy of AI systems depends not only on their outcomes but also on the fairness of the decision-making process (Binns, 2018; Mittelstadt et al., 2016). Transparency, explainability, opportunities for contestation, and institutional accountability are all essential components of procedural fairness. Because these requirements vary across AI applications and deployment contexts, governance mechanisms should be tailored to the specific characteristics of different AI systems rather than applied uniformly.

Third, responsibility-gap theory argues that autonomous and learning systems can create uncertainty regarding who should be held accountable for AI-related harms (Matthias, 2004; Coeckelbergh, 2020; Santoni de Sio & Mecacci, 2021). Addressing these gaps requires governance arrangements that clearly allocate responsibilities among developers, deployers, operators, regulators, and other stakeholders according to the context in which AI systems are used.

Together, these perspectives provide the normative foundation for harm-based governance. Rather than relying solely on abstract ethical principles, they support governance grounded in empirically observed patterns of harm, enabling context-sensitive oversight, differentiated accountability, and more effective operationalization of principles such as fairness, transparency, safety, and responsibility.

3. Data and Preprocessing

The analysis was conducted using the AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) Repository, a publicly available database of documented AI, algorithmic, and automation incidents. The raw dataset contained 2,243 records. Data preprocessing was performed in three stages to produce a consistent analytical sample. In Stage 1, duplicate records, parent–child incident entries referring to the same event, and records with invalid identifiers were removed, resulting in a working dataset of 1,574 incidents. In Stage 2, each incident was coded into three categorical variables using information contained in the repository: AI Modality, Deployment Domain, and Risk Category. AI Modality was derived from the *Technology* field, Deployment Domain from the *Sector* field, and Risk Category from the combined *Individual Harm* and *Societal Harm* fields. Incidents that could not be unambiguously assigned to one of the predefined categories were excluded, leaving 1,127 incidents. Detailed coding definitions are provided in the Coding Protocol (Appendix A).

A fourth variable, System Function, was derived from the *Purpose* field to capture the operational role of the AI system independent of its technology or deployment context. Five functional categories were defined: Identification and Monitoring, Autonomous Control, Pricing and Optimization, Decision Support, and Content Generation. System Function was assigned using a rule-based text classification procedure based on predefined keyword patterns. Incidents whose purpose could not be classified with sufficient confidence were excluded from the analysis. Unlike earlier iterations of the coding procedure, unmatched records were not assigned to a default category, thereby reducing the risk of systematic misclassification. After this final preprocessing stage, the analytical dataset comprised 849 incidents.

4. Methodology

This study employs Multiple Correspondence Analysis (MCA) followed by hierarchical cluster analysis to identify recurring configurations of AI-related harms within the documented incident record. MCA is an exploratory multivariate technique designed for categorical data that represents the associations among variables in a low-dimensional geometric space without imposing distributional assumptions (Greenacre, 2017; Le Roux & Rouanet, 2010). This approach is well suited to the AIAAIC incident repository, where the objective is to uncover latent patterns across multiple categorical characteristics rather than to test predefined causal relationships.

4.1 Multiple Correspondence Analysis

The MCA was performed using four active categorical variables derived during the coding process: AI Modality (4 categories), Deployment Domain (9 categories), Risk Category (4 categories), and System Function (5 categories), comprising a total of 22 active categories across 849 incidents. Each incident was represented as a row in a complete disjunctive matrix, and MCA was used to project both incidents and categories into a common multidimensional space. The proximity of points in this space reflects the strength of association among categories, enabling systematic identification of recurring patterns within the documented incidents.

Although the complete set of dimensions was computed, interpretation focuses on the first two dimensions because they capture the dominant structure of the data and provide an interpretable representation of the principal associations among variables. To facilitate interpretation, three complementary measures of explained inertia were examined: raw inertia, Benzécri-adjusted inertia, and Greenacre-adjusted inertia (Benzécri, 1979; Greenacre, 1993). Because raw inertia is known to underestimate the explanatory importance of the leading dimensions in MCA, the adjusted measures provide a more meaningful assessment of the concentration of variation within the lower-dimensional solution. Interpretation of the resulting dimensions was guided by standard MCA diagnostics, including category coordinates, category contributions, masses, and squared cosine (cos^2) values. Together, these statistics identify the categories that contribute most strongly to each dimension and indicate how well individual categories are represented within the retained dimensions. Complete diagnostic results are reported in Appendix X.

4.2 Cluster Analysis

To identify recurring profiles of AI incidents, hierarchical clustering was performed on the MCA row coordinates using Ward's minimum variance method. Applying clustering in the reduced MCA space minimizes the influence of sparsity inherent in high-dimensional categorical data while preserving the principal association structure identified by the correspondence analysis. Ward's method was selected because it iteratively minimizes within-cluster variance, producing compact and interpretable clusters suitable for exploratory classification. The resulting clusters were examined alongside the MCA category projections to characterize distinct profiles of AI-related harms and their associated technological, functional, and contextual characteristics.

4.3 Robustness Assessment

Several robustness analyses were conducted to evaluate the stability of the findings. First, the MCA was re-estimated after excluding System Function to assess the influence of the keyword-derived variable on the underlying association structure. Second, the analysis was repeated after removing the largest deployment domain to determine whether the principal dimensions were driven by a dominant category. Third, bootstrap resampling was used to estimate the stability of category coordinates across repeated samples. Finally, the analysis was replicated using the manually validated subset of system-function classifications to verify that the automated coding procedure did not materially influence the results.

Across these alternative specifications, the principal dimensions and cluster structure remained substantively consistent, supporting the robustness of the identified incident profiles.

5. Results

5.1 Sample Characteristics

The final analytical sample consisted of 849 AI incidents extracted from the AIAAIC repository following the preprocessing procedure described in Section 3. Generative AI represented the largest AI modality (36.3%), followed by Vision Recognition (27.7%), Predictive Monitoring (25.6%), and Autonomous/Embodied AI (10.5%). The most common deployment domains were Media and Information (27.8%), Transport (13.8%), Law Enforcement and Justice (12.4%), and Government and Politics (12.2%). Discrimination and Rights was the most frequently reported risk category (44.6%), followed by Economic Harm (21.8%), Safety Failures (19.0%), and Psychological or Reputational Harm (14.6%). Content Generation was the dominant system function (46.2%), followed by Identification and Monitoring (24.5%), Decision Support (13.8%), Autonomous Control (12.1%), and Pricing and Optimization (3.4%).

5.2 Multiple Correspondence Analysis

Table 1 presents the eigenvalues and inertia of the MCA solution. The first two dimensions account for 30.0% of the raw inertia and 85.0% of the Greenacre-adjusted inertia, indicating that the dominant association structure of the incident data is adequately represented in a two-dimensional solution.

Table 1
Eigenvalues and inertia of the MCA solution

Dim	Eigenvalue	Raw %	Benzécri %	Greenacre %
1	0.760	16.9	63.5	**58.9**
2	0.589	13.1	28.0	**26.0**
3	0.399	8.9	5.4	5.0
4	0.353	7.8	2.6	2.4
5	0.291	6.5	0.4	0.4
…	…	…	…	…
Cum. 1–2		30.0	91.5	**85.0**

The first dimension is defined primarily by Autonomous/Embodied AI, Autonomous Control, Transport, and Safety Failures. These categories occupy one end of the dimension and distinguish AI systems that interact directly with the physical world from applications that primarily process information. Accordingly, Dimension 1 is interpreted as an embodiment axis. The second dimension separates

incidents associated with Identification and Monitoring, Vision Recognition, Law Enforcement and Justice, and Discrimination and Rights from those associated with Generative AI, Content Generation, Media and Information, and Economic Harm. This dimension therefore represents a rights-versus-content axis, distinguishing identification and surveillance applications from generative and content-oriented AI systems as shown in Fig. 1 and Fig. 2.

Category diagnostics indicate that Autonomous/Embodied AI, Vision Recognition, and Generative AI occupy stable and well-defined positions within the MCA space. In contrast, Predictive Monitoring and Decision Support are positioned near the center of the map, indicating weaker associations with any single combination of deployment domain, system function, and risk category.

5.3 AI Harm Profiles

Cluster	n (% of sample)	Dominant modality	Dominant domain	Dominant risk	Dominant function
W1	82 (9.7%)	Autonomous / Embodied	Transport	Safety Failures	Autonomous Control
W2	215 (25.3%)	Vision Recognition	Law Enforcement / Justice	Discrimination / Rights	ID & Monitoring
W3	167 (19.7%)	Generative AI	Media / Information	Economic Harm	Content Generation
W4	385 (45.3%)	Generative AI	Media / Information	Discrimination / Rights	Content Generation

Ward hierarchical clustering was applied to the MCA row coordinates to identify recurring patterns within the incident space as shown in Fig. 3. Although the clustering algorithm produced four statistical clusters, two clusters represented subdivisions of the same Generative AI region and differed primarily in the dominant type of harm (economic versus rights-related). Because they shared the same technology, deployment context, and system function, they were interpreted as a single governance profile. Consequently, the analysis identifies three coherent AI harm profiles, corresponding to the principal regions of the MCA space.

Profile 1: Autonomous/Embodied AI: The first profile represents Autonomous/Embodied AI systems deployed primarily within the transport sector. Autonomous Control is the dominant system function, while Safety Failures constitute the principal risk category. These incidents involve AI systems with direct physical interaction, where failures may produce immediate operational consequences or physical harm. This profile occupies a distinct region of the MCA map and is clearly separated from information-processing applications by the embodiment axis.

Profile 2: Vision Recognition: The second profile is characterized by Vision Recognition technologies deployed predominantly within Law Enforcement and Justice. Identification and Monitoring is the

dominant system function, and Discrimination and Rights is the principal risk category. Incidents within this profile largely involve facial recognition, biometric identification, surveillance, and related technologies operating in rights-sensitive environments. The profile occupies the upper portion of the second MCA dimension, reflecting its association with identification and monitoring activities.

Profile 3: Generative AI: The third profile is dominated by Generative AI systems operating primarily within the Media and Information domain. Content Generation is the defining system function, while the associated harms include both Economic Harm and Discrimination and Rights. Although Ward clustering separated this region into two statistical clusters, the distinction reflects differences in the observed consequences rather than differences in the underlying AI technology or deployment context. Accordingly, these clusters are interpreted as representing a single Generative AI governance profile with multiple forms of associated harm.

Unlike the other AI modalities, *Predictive Monitoring* does not emerge as an independent profile. Instead, incidents involving predictive monitoring are distributed across multiple regions of the MCA space, indicating that predictive monitoring functions as a cross-cutting operational capability rather than a distinct category of AI harm.

5.4 Robustness Analysis

Several robustness analyses were conducted to evaluate the stability of the findings. Re-estimating the MCA without the System Function variable produced the same substantive interpretation of the two principal dimensions, indicating that the overall structure is not dependent on the keyword-derived functional classification. Similarly, excluding the Media and Information domain did not materially alter the interpretation of either dimension, although the second dimension shifted slightly toward Law Enforcement and Justice after removing the dominant media-related incidents.

Bootstrap resampling demonstrated that the positions of the Autonomous/Embodied AI, Vision Recognition, and Generative AI profiles remained stable across repeated samples as shown in Fig. 4. In contrast, categories located near the center of the MCA space, including Predictive Monitoring, Decision Support, and Pricing and Optimization, exhibited greater positional variability, consistent with their weaker structural associations. Overall, the robustness analyses consistently support two principal dimensions underlying the documented incident record: an embodiment axis separating physically consequential AI systems from information-processing applications and a rights-versus-content axis separating identification and monitoring systems from generative and content-oriented AI. Together, these dimensions provide the empirical basis for the three AI harm profiles developed in this study.

6. Discussion

6.1 Ethical Interpretation of the Harm Profiles

The findings show that AI incidents are not isolated events but follow recurring patterns defined by combinations of AI modality, deployment domain, system function, and risk category. Multiple Correspondence Analysis (MCA) identified two stable structural dimensions. The first distinguishes AI systems with direct physical consequences from those primarily processing information, while the second separates rights-sensitive identification and monitoring systems from generative, content-oriented AI. These dimensions suggest that AI harms differ not only in severity but also in their underlying ethical characteristics.

The embodiment dimension distinguishes safety-critical systems, such as autonomous vehicles, from information-processing applications. This distinction reflects different governance priorities: systems capable of causing physical harm to require robust safety assurance and operational oversight, whereas information-processing systems raise concerns related to privacy, fairness, transparency, and autonomy (Nissenbaum, 2010; Binns, 2018). The rights-versus-content dimension similarly differentiates identification technologies from generative AI. Identification systems primarily affect privacy, civil rights, and procedural fairness, whereas generative AI introduces challenges related to misinformation, content authenticity, intellectual property, and the integrity of information ecosystems. Together, these dimensions demonstrate that AI governance should distinguish between fundamentally different categories of harm rather than applying uniform oversight mechanisms.

6.2 Harm Profiles and Governance Implications

Hierarchical clustering identified three recurring harm profiles: Autonomous/Embodied AI, Vision Recognition, and Generative AI. Each profile represents a distinct combination of technological characteristics, deployment contexts, and ethical risks, indicating that governance requirements should differ across AI applications.

The *Autonomous/Embodied AI* profile is associated primarily with physical safety risks in domains such as transportation. Governance for these systems should emphasize safety validation, continuous operational monitoring, incident reporting, and clearly defined institutional responsibilities among developers, deployers, regulators, and operators (Matthias, 2004; Santoni de Sio & Mecacci, 2021).

The *Vision Recognition* profile is characterized by discrimination, surveillance, and rights-related concerns, particularly in law enforcement and public-sector applications. These systems require governance mechanisms that emphasize fairness evaluation, demographic performance monitoring, human oversight, transparency, and effective mechanisms for review and redress (Binns, 2018; Mittelstadt et al., 2016).

The *Generative AI* profile is dominated by harms affecting information quality and public trust, including misinformation, impersonation, and unreliable AI-generated content. Effective governance therefore requires controls such as content provenance, reliability evaluation, transparency regarding AI-generated outputs, and continuous monitoring of post-deployment performance (Hagendorff, 2020).

Unlike these three profiles, *Predictive Monitoring* did not emerge as an independent category. Instead, predictive capabilities appeared across multiple domains, indicating that their ethical implications depend primarily on the application context rather than on prediction itself. This finding reinforces the importance of context-sensitive governance because identical predictive techniques may require different safeguards when applied in healthcare, employment, finance, or transportation.

6.3 Implications for AI Governance

The findings complement existing governance frameworks, including the NIST AI Risk Management Framework, the NIST Generative AI Profile, the EU AI Act, and the OECD AI Classification Framework, by providing an empirical basis for differentiating governance requirements according to recurring incident patterns rather than technology categories alone (National Institute of Standards and Technology, 2023, 2024; OECD, 2022; European Parliament and Council of the European Union, 2024). Rather than replacing these frameworks, the proposed harm-profile approach serves as an intermediate analytical layer connecting documented AI failures with practical governance activities. Organizations can use recurring harm profiles to prioritize monitoring, auditing, evidence collection, and accountability according to the risks most frequently associated with AI systems. The results also highlight the importance of post-deployment governance, as many documented incidents resulted from changing operational conditions, evolving user behavior, or organizational decisions that could not reasonably be identified during pre-deployment assessment alone.

6.4 Contributions

This study makes three contributions to AI governance research. First, it demonstrates empirically that documented AI incidents are organized around stable structural dimensions and recurring harm profiles rather than representing isolated events. Second, it shows that these profiles correspond to distinct categories of ethical concern and therefore require different governance responses. Third, it proposes a harm-profile governance framework that translates empirical incident patterns into differentiated monitoring activities, audit priorities, accountability mechanisms, and oversight responsibilities. In doing so, the study provides a practical link between AI incident analysis, normative AI ethics, and governance implementation.

6.5 Limitations

Several limitations should be considered when interpreting the findings. The analysis is based on incidents documented in the AIAAIC Repository, which reflects publicly reported cases and may therefore be influenced by reporting practices, media attention, and repository curation. Consequently, the identified harm profiles should be interpreted as empirical patterns rather than exhaustive categories of AI failures. In addition, each incident was assigned a primary risk category, although many AI incidents involve multiple interacting harms. Future research could examine multi-label harm classifications and evaluate whether the proposed governance framework generalizes across additional incident repositories and emerging AI technologies.

7. Harm-Profile Governance Framework

The empirical findings form the basis of a harm-profile governance framework that translates recurring patterns of AI-related harm into differentiated governance responses. Rather than applying uniform controls across all AI systems, the framework recommends governance measures tailored to the technological characteristics, operational functions, deployment contexts, and risk patterns associated with each harm profile.

The framework is grounded in three complementary ethical perspectives. Contextual integrity informs the selection of monitoring indicators and audit triggers by aligning oversight with the social and operational context in which an AI system is deployed (Nissenbaum, 2010). Procedural justice guides evidence requirements and review mechanisms, ensuring that AI decisions are transparent, explainable, and open to contestation where appropriate (Binns, 2018). Responsibility-gap theory informs the allocation of accountability by assigning governance responsibilities to the actors best positioned to prevent, monitor, and respond to AI-related harms, including developers, deployers, operators, regulators, and other institutional stakeholders (Santoni de Sio & Mecacci, 2021). Rather than replacing existing AI governance frameworks, the proposed approach operationalizes them by linking governance activities to empirically observed harm profiles instead of broad technology categories. This enables organizations to implement context-sensitive monitoring, auditing, and accountability mechanisms that better reflect the recurring patterns of harm documented in real-world AI incidents.Top of FormBottom of Form

7.1 Governance Pipeline

Figure 5 presents the proposed governance pipeline.

The first stage, *System Context*, characterizes an AI system according to AI modality, deployment domain, system function, and anticipated risk category. These four dimensions provide the contextual information required to identify the system's likely harm profile. The second stage, *Harm-Profile Mapping*, associates the system with one or more of the empirically derived governance profiles. While many systems correspond primarily to a single profile, multimodal applications may require governance controls from multiple profiles. The third stage, *Governance Response*, applies profile-specific monitoring indicators, audit triggers, evidence requirements, escalation procedures, and institutional responsibilities. The final stage, *Adaptive Revision*, incorporates new incident evidence, audit findings, and monitoring outcomes to continuously refine the governance profiles as AI technologies and deployment environments evolve.

7.2 Profile A: Autonomous/Embodied AI

The dominant configuration in this region of the analytic sample combines the transport domain, an autonomous-control function, and safety failures. Documented incidents recur through a small number of failure mechanisms: sensor misperception under distribution shift, edge cases in traffic scenarios, over-reliance on automation by the human operator, hand-off failures between automated and manual

control, and latent hardware faults that propagate through automated response. The stakeholders placed at risk are vehicle occupants and other road users, but also emergency responders, deployment operators, and insurers who bear the operational and financial consequences of failures.

The controls that follow from these mechanisms are essentially operational. Monitoring should track disengagement rates per operational hour, near-miss and hard-brake events per unit distance, operational-design-domain excursions, sensor-confidence anomalies, time-to-take-over statistics, and safety-critical software regressions per release. Audit should be triggered by any fatal or serious-injury event, by a sustained rise in the near-miss rate, by software or firmware updates touching perception, planning, or actuation, and by any expansion of the operational design domain. Evidence requirements comprise the operational-design-domain specification, simulation and on-road validation records, disengagement logs, software change logs, and incident-escalation records. Escalation should reserve an immediate stop-deployment authority for the operator, mandate reporting to the national transport safety regulator (aligned with the EU AI Act serious-incident reporting regime), and provide for a joint operator–regulator post-incident review. Institutional responsibility is distributed between the manufacturer, the deployment operator, the national transport safety authority, and the insurer.

7.3 Profile B: Vision Recognition

The dominant configuration in this region combines the law-enforcement and justice domain, an identification-and- monitoring function, and discrimination or rights-related risks. Failures recur through differential accuracy across demographic groups, false matches at low identification thresholds, surveillance overreach beyond a specified purpose, use of unlabeled or unlawfully obtained training imagery, and retention of biometric records beyond the stated purpose. The stakeholders placed at risk are persons identified or misidentified by the system, the broader populations subject to identification, and the downstream decision-makers police, prosecutors, courts, and service providers whose actions the system informs.

Monitoring in this profile centers on the false-match rate stratified by demographic group, on confirmed misidentification events, on query volume and query-scope expansion, on the number of retained biometric templates, and on the growth rate of any watchlist. Audit should be triggered by any confirmed misidentification that affects downstream custody, employment, or benefit decisions; by group-level disparity above a defined threshold; by scope expansion; and by any change to the identification threshold. Evidence requirements comprise demographic-stratified bias assessment, data-provenance documentation, retention and access logs, query-authorization logs, and redress-request records. Escalation should provide an accessible complaint and redress mechanism for affected persons, routine judicial review of custody or benefit decisions informed by the system, and regulatory review upon threshold breach. Institutional responsibility is shared between the deploying agency, the procuring authority, the data-protection regulator, and the relevant sector regulator (justice, immigration, employment, and so on).

7.4 Profile C: Generative AI

The dominant configuration in this region combines the media and information domain, a content-generation function, and either discrimination- or economic-harm risks. Failures recur through misinformation and hallucination, deepfake and impersonation content, unattributed copying of training content, economic manipulation via fabricated content, interaction failures associated with poor groundedness, poor controllability, or unhelpful or unsafe responses, and user reliance on the system's unwarranted authority. The stakeholders placed at risk are end users of the generative system, persons depicted, referenced, or impersonated, readers and consumers of downstream content, content creators whose work is used in training, and the platforms and information ecosystems in which the outputs circulate.

Monitoring here should combine content-provenance coverage (the share of outputs carrying attached credentials), the rate of user-flagged misinformation and impersonation, groundedness and truthfulness scores on domain-relevant benchmarks, controllability failures such as refusal evasion and jailbreak success rates, and dialogue-reliability measures for interactive systems (Marconi, Longo, & Cabitza, 2026). Audit should be triggered by any serious misinformation event, verified impersonation event, or identified pattern of biased or discriminatory outputs, and by any unexpected drop in truthfulness or groundedness metrics or breach of the regulator- or platform-complaint threshold. Evidence requirements comprise training-data provenance documentation, content-provenance mechanisms (for example, C2PA), safety-testing records, interaction-quality evaluation records covering groundedness, controllability, helpfulness, truthfulness, user alignment, and dialogue reliability, and disclosure records for AI-generated content. Escalation should include user-visible correction and appeal channels, platform-level takedown for impersonation and disallowed content, and regulator notification for serious misinformation events aligned with the EU AI Act and the NIST Generative AI Profile. Institutional responsibility is shared between the model developer, the deploying platform, downstream content-distribution platforms, sector regulators for media, competition, and consumer protection, and civil-society monitoring bodies.

7.5 Cross-Cutting Function: Predictive Monitoring

Because Predictive Monitoring does not form a distinct cluster in the incident space, it is best treated as a cross-cutting operational function that inherits governance controls from the profile hosting it. The characteristic failure mechanisms are distribution shift from the development to the deployment population, feedback loops in which system predictions reshape the input distribution, opaque scoring rules, an inability of affected persons to inspect or contest a score, and slow erosion of accuracy over time without triggering intervention. The stakeholders placed at risk are persons scored, ranked, or classified by the system, the downstream decision-makers who rely on the score, and the parties responsible for the underlying decision, including employers, lenders, benefit administrators, and platforms. When a predictive-monitoring component is embedded in a deployment that already carries a profile, the controls of that profile apply, augmented by additional emphasis on performance metrics stratified by protected attribute, calibration drift, adverse-decision rates and appeal-overturn rates, robustness and uncertainty diagnostics (Marconi & Cabitza, 2025), and explanation quality on held-out

samples. Escalation should provide an individual appeal mechanism with human review, a group-level audit trigger, and regulator notification when disparities exceed threshold. Responsibility is shared between the system developer, the deploying organisation, the relevant sector regulator (finance, employment, benefits, health, and so on), and the data-protection regulator.

7.6 Consolidated Governance Summary

Dimension	Profile A: Autonomous / Embodied	Profile B: Vision Recognition	Profile C: Generative AI
Dominant configuration	Transport; autonomous control; safety failures	Law enforcement / justice; identification and monitoring; discrimination / rights	Media / information; content generation; discrimination / rights or economic harm
Characteristic failure mechanisms	Sensor misperception; edge cases; hand-off failures; over-reliance on automation	Differential accuracy across demographic groups; false matches; surveillance overreach	Misinformation and hallucination; impersonation; interaction failures; economic manipulation
Priority monitoring indicators	Disengagement rate; near-miss events per unit distance; operational-domain excursions	False-match rate by demographic group; confirmed misidentifications; watchlist growth	Provenance coverage; user-flagged misinformation and impersonation; groundedness and truthfulness scores
Primary institutional responsibility	Manufacturer; deployment operator; transport safety authority; insurer	Deploying agency; procuring authority; data-protection and sector regulators	Model developer; deploying platform; sector regulators; civil society monitoring bodies

7.7 Framework Summary

The proposed framework operationalizes the empirical findings by translating recurring AI incident patterns into differentiated governance responses. Rather than replacing existing AI governance frameworks, it complements them by providing evidence-based guidance for selecting monitoring indicators, audit triggers, evidence requirements, and institutional responsibilities that correspond to the observed characteristics of different AI harm profiles.

The framework is intended to support lifecycle governance and should be applied alongside system-specific risk assessment and organizational governance processes. Because AI technologies and deployment contexts continue to evolve, the framework is designed to be adaptive, allowing new incident evidence to refine both the harm profiles and the governance controls associated with them.

8. Conclusion

As AI systems continue to expand across diverse domains, effective governance increasingly depends on understanding both empirically and ethically how harms emerge after deployment in real-world settings. This study demonstrates that the documented AI incident record is organized around two stable structural dimensions an *embodiment axis* and a *rights-versus-content axis* that give rise to three coherent AI harm profiles: *Autonomous/Embodied AI, Vision Recognition*, and *Generative AI*. Read normatively, these dimensions correspond to distinct families of ethical concern harms to bodily integrity versus harms mediated by information, and dignitary harms concentrated at the intersection of identification and state power versus epistemic harms concentrated in the circulation of generated content. Predictive Monitoring, in contrast, is best understood as a cross-cutting operational capability whose moral evaluation depends on the broader application context, an outcome congenial to contextual approaches in AI ethics.

Building on these findings, the study proposes a harm-profile governance framework that translates recurring incident patterns into differentiated monitoring indicators, audit triggers, evidence requirements, escalation pathways, and most importantly differentiated allocations of moral and institutional responsibility. Rather than replacing existing AI governance frameworks, the proposed approach complements them by giving abstract ethical principles a place to land: fairness, transparency, safety, autonomy, and accountability become concrete duties attached to specific harm profiles rather than free-floating commitments.

More broadly, this research demonstrates that documented AI incidents can serve not only as retrospective records of failure but also as an ethical resource for prospective governance. By linking recurring patterns of AI harm with practical governance responses and their normative justifications, the proposed framework offers a systematic approach to context-sensitive AI ethics one that treats moral responsibility not as an abstract commitment but as an empirically informed and institutionally distributed practice, capable of adapting as AI technologies and their deployment contexts continue to evolve.

Declarations

Acknowledgments

The authors acknowledge the use of AI tools for improving the readability and language of this manuscript. All conceptual, methodological, and analytical content is the work of the authors.

Data availability

The AIAAIC Repository incident export used as the raw data source is publicly available from the AIAAIC project. The working shortlist after Stage 2 preprocessing is released with this paper as AIAAIC_Shortlist_Analysis.csv, and the analytic sample after Stage 3 System Function coding (n = 849) is released as part of the reproducibility bundle.

Code availability

All analysis code is released as analysis_pipeline.py (consolidated multiple correspondence analysis, k-means and Ward clustering, bootstrap, and all robustness checks) together with sysfunc_validation_sample.py (stratified sampling for the validation subset) and sysfunc_validation_score.py (classification metrics against the manual labels). Reproducibility uses a fixed random seed of 42 throughout. The code is written in Python 3.9 and depends only on the prince, scikit-learn, scipy, pandas, numpy, and matplotlib libraries.

Competing interests

The authors declare that they have no competing interests.

Author contributions

Author A designed the study, conducted the analysis, and drafted the manuscript. Authors B and Cs contributed to the coding protocol, reviewed the results, and edited the manuscript.

Use of generative AI in manuscript preparation

The authors used generative AI tools to refine language and improve readability in specific passages of the manuscript.

References

1. Abercrombie, G., Benbouzid, D., Giudici, P., Golpayegani, D., Hernandez, J., Noro, P., Pandit, H., Paraschou, E., Pownall, C., Prajapati, J., Sayre, M. A., Sengupta, U., Suriyawongkul, A., Thelot, R., Vei, S., & Waltersdorfer, L. (2024). A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*. https://doi.org/10.48550/arXiv.2407.01294
2. Bach, T. A., Kaarstad, M., Solberg, E., & Babic, A. (2025). Insights into suggested responsible AI (RAI) practices in real-world settings: A systematic literature review. *AI and Ethics*, *5*, 3185–3232. https://doi.org/10.1007/s43681-024-00404-x
3. Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, *4*(3), 377–378.
4. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research, Vol. 81, pp. 149–159). PMLR.
5. Coeckelbergh, M. (2020). *AI ethics*. MIT Press.
6. Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (pp. 1571–1583). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533213

7. Cousineau, C., Dara, R., & Chowdhury, A. (2025). Trustworthy AI: AI developers' lens to implementation challenges and opportunities. *Data and Information Management*, *9*, 100082. https://doi.org/10.1016/j.dim.2024.100082
8. Dixon, R. B. L., & Frase, H. (2025). *AI incidents: Key components for a mandatory reporting regime* (Technical report). Center for Security and Emerging Technology. https://doi.org/10.51593/2024CA004
9. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, *L 2024/1689*.
10. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1
11. Greenacre, M. (1993). *Correspondence analysis in practice* (1st ed.). Academic.
12. Greenacre, M. (2017). *Correspondence analysis in practice* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315369983
13. Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8
14. Hupont, I., Fernández-Llorca, D., Baldassarri, S., & Gómez, E. (2024). Use case cards: A use case reporting framework inspired by the European AI Act. *Ethics and Information Technology*, *26*, 19. https://doi.org/10.1007/s10676-024-09757-7
15. Kemmerzell, N., Schreiner, A., Khalid, H., Schalk, M., & Bordoli, L. (2025). Towards a better understanding of evaluating trustworthiness in AI systems. *ACM Computing Surveys*, *57*(9). https://doi.org/10.1145/3716388. Article 218.
16. Le Roux, B., & Rouanet, H. (2010). *Multiple correspondence analysis*. SAGE.
17. Marconi, L., & Cabitza, F. (2025). Show and tell: A critical review on robustness and uncertainty for a more responsible medical AI. *International Journal of Medical Informatics*, *202*, 105970. https://doi.org/10.1016/j.ijmedinf.2025.105970
18. Marconi, L., Longo, L., & Cabitza, F. (2026). Assessing interaction quality in human–AI dialogue: An integrative review and multi-layer framework for conversational agents. *Machine Learning and Knowledge Extraction*, *8*(2), 28. https://doi.org/10.3390/make8020028
19. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1
20. McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817
21. Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on*

Fairness, Accountability, and Transparency (FAccT '21) (pp. 735–746). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445935

22. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679
23. Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. *Digital Society*, *2*(3), 49. https://doi.org/10.1007/s44206-023-00074-y
24. Mökander, J., Sheth, M., Watson, D. S., & Floridi, L. (2023). The switch, the ladder, and the matrix: Models for classifying AI systems. *Minds and Machines*, *33*, 223–250. https://doi.org/10.1007/s11023-022-09620-y
25. Monteiro, N., & Singh, V. (2025). The wheel of artificial intelligence governance. *Sustainable Futures*, *10*, 101279. https://doi.org/10.1016/j.sftr.2025.101279
26. National Institute of Standards and Technology (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
27. National Institute of Standards and Technology (2024). *Artificial intelligence risk management framework: Generative AI profile* (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1
28. Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
29. Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). AI risk assessment: A scenario-based, proportional methodology for the AI Act. *Digital Society*, *3*, 13. https://doi.org/10.1007/s44206-024-00095-1
30. OECD (2022). *OECD framework for the classification of AI systems* (OECD Digital Economy Papers No. 323). OECD Publishing. https://doi.org/10.1787/cb6d9eca-en
31. Papyshev, G. (2025). Governing AI through interaction: Situated actions as an informal mechanism for AI regulation. *AI and Ethics*, *5*, 1109–1120. https://doi.org/10.1007/s43681-024-00446-1
32. Raghupathi, W., Ren, J., & Kulkarni, T. (2026). Examining the nature and dimensions of artificial intelligence incidents: A machine learning text analytics approach. *AppliedMath*, *6*(1), 1. https://doi.org/10.3390/appliedmath6010011
33. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT\ '20)** (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873
34. de Santoni, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

35. Schmitz, A., Mock, M., Görge, R., Cremers, A. B., & Poretschkin, M. (2025). A global scale comparison of risk aggregation in AI assessment frameworks. *AI and Ethics*, *5*, 1407–1432. https://doi.org/10.1007/s43681-024-00479-6
36. Schuett, J. (2024). Risk management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, *15*, 367–385. https://doi.org/10.1017/err.2023.1
37. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28, pp. 2503–2511). Curran Associates.
38. Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)* (pp. 723–741). Association for Computing Machinery. https://doi.org/10.1145/3600211.3604673
39. Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2026). The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, *7*, 2. https://doi.org/10.1016/j.patter.2025.101517
40. Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z., & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, *56*, 12799–12831. https://doi.org/10.1007/s10462-023-10420-8
41. Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P., & Prunkl, C. (2024). A taxonomy of systemic risks from general-purpose AI. *arXiv preprint arXiv:2412.07780*. https://doi.org/10.48550/arXiv.2412.07780
42. Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
43. Yampolskiy, R. V. (2018). Predicting future AI failures from historic examples. *Foresight*, *21*(1), 138–152. https://doi.org/10.1108/FS-04-2018-0034
44. Zanotti, G., Chiffi, D., & Schiaffonati, V. (2024). AI-related risk: An epistemological approach. *Philosophy & Technology*, *37*, 66. https://doi.org/10.1007/s13347-024-00755-7
45. Zhu, L., Lu, Q., Ding, M., Lee, S. U., & Wang, C. (2026). Designing meaningful human oversight in AI. *AI and Ethics*, *6*, 3. https://doi.org/10.1007/s43681-025-00584-0

Appendix

Appendix X: Coding Protocol Apendix

This appendix documents the categorical mapping used to convert the AIAAIC free-text fields into the four analytical variables. The protocol was applied programmatically. Every mapping rule is exposed in

the released analysis code (analysis_pipeline.py) so that alternative codings can be tested and compared. A stratified random sample of 204 incidents (35 per regex class, plus all 29 Pricing/Optimization cases) was manually re-coded by a single expert annotator, blind to the regex prediction, to compute precision, recall, and F1 per class.

1 Definitions and inclusion patterns

1 AI Modality (from Technology)

· **Autonomous / Embodied.** Physically embedded systems whose primary action is exerted on the material environment: automated driving, driver-assistance systems, autopilots, robotics, drones, unmanned vehicles, delivery robots.

· **Vision Recognition.** Systems whose primary input is images or video and whose function is to classify, identify, or analyse perceptual content: facial and biometric recognition, object and scene recognition, deepfake detection, body-worn or vehicle cameras used for identification.

· **Generative AI.** Systems whose primary output is generated content: large language models, chatbots, text-to-image, text-to-video and text-to-speech systems, synthetic media, diffusion models, and other content-producing agents.

· **Predictive Monitoring.** Systems whose primary output is a score, ranking, prediction, or recommendation used to inform a decision: risk assessment, hiring, credit scoring, dynamic pricing when used for decisions rather than optimisation, content ranking, targeted advertising, and behavioural surveillance analytics.

Rows whose Technology field could not be resolved with confidence to exactly one of these four classes were excluded at Stage 2.

1.2 Deployment Domain (from Sector)

Nine classes: *Transport*, *Law Enforcement / Justice*, *Health*, *Media / Information*, *Government / Politics*, *Financial Services*, *Education / Research*, *Retail / Consumer*, *Technology*. Full inclusion strings are in the code repository.

1.3 Risk Category (from Individual_Harm + Societal_Harm)

Four classes:

· **Safety Failures.** Loss of life, bodily injury, property damage, health deterioration.

· **Discrimination / Rights.** Discrimination, loss of rights, privacy loss, harassment, surveillance-related harm.

· **Economic Harm.** Financial loss, fraud, job loss, intellectual property harm, market manipulation.

- **Psychological / Reputational.** Anxiety, distress, reputational damage, deception, trauma.

Where an incident's harm text plausibly touched more than one class, we coded the *dominant* class from the manual AIAAIC harm summaries. The single-label constraint is a simplification: it enables MCA on a compact, well-defined variable space but suppresses secondary harms. The implications of this choice are discussed in the Limitations section, and a multi-label robustness check is reported in Section X.

1.4 System Function (from Purpose)

Applied via the following ordered rules on the lowercased Purpose string. The first pattern that matches assigns the class; unmatched rows are excluded.

2 Validation

A stratified sample of 204 incidents was drawn: 35 rows per regex class (ID & Monitoring, Autonomous Control, Decision Support, Content Generation, Unmatched), plus all 29 rows in the smallest stratum (Pricing/Optimisation). Each row was manually re-coded by a two expert annotators blind to the regex prediction. The annotators were instructed to assign the same six-way label (five defined classes plus Unclear when the Purpose text did not resolve to a single class) and, where a row was truly unmatched, to confirm that no class applied.

Agreement between the manual re-coding and the regex classifier was complete on the stratified sample:

The validation sample is small relative to the analytic sample (n = 204 out of 849) and covers only *Purpose* strings that already appear in the AIAAIC repository, so the perfect agreement should not be over-interpreted. It is best read as evidence that (a) the regex rules do not systematically miscode any of the observed Purpose phrasings in the stratified sample, and (b) the class definitions are separable in practice. It does *not* rule out errors on future Purpose phrasings not represented in the sample; nor does it substitute for the robustness check in which MCA is refit after dropping System Function altogether (Section 5, Robustness).

Figures

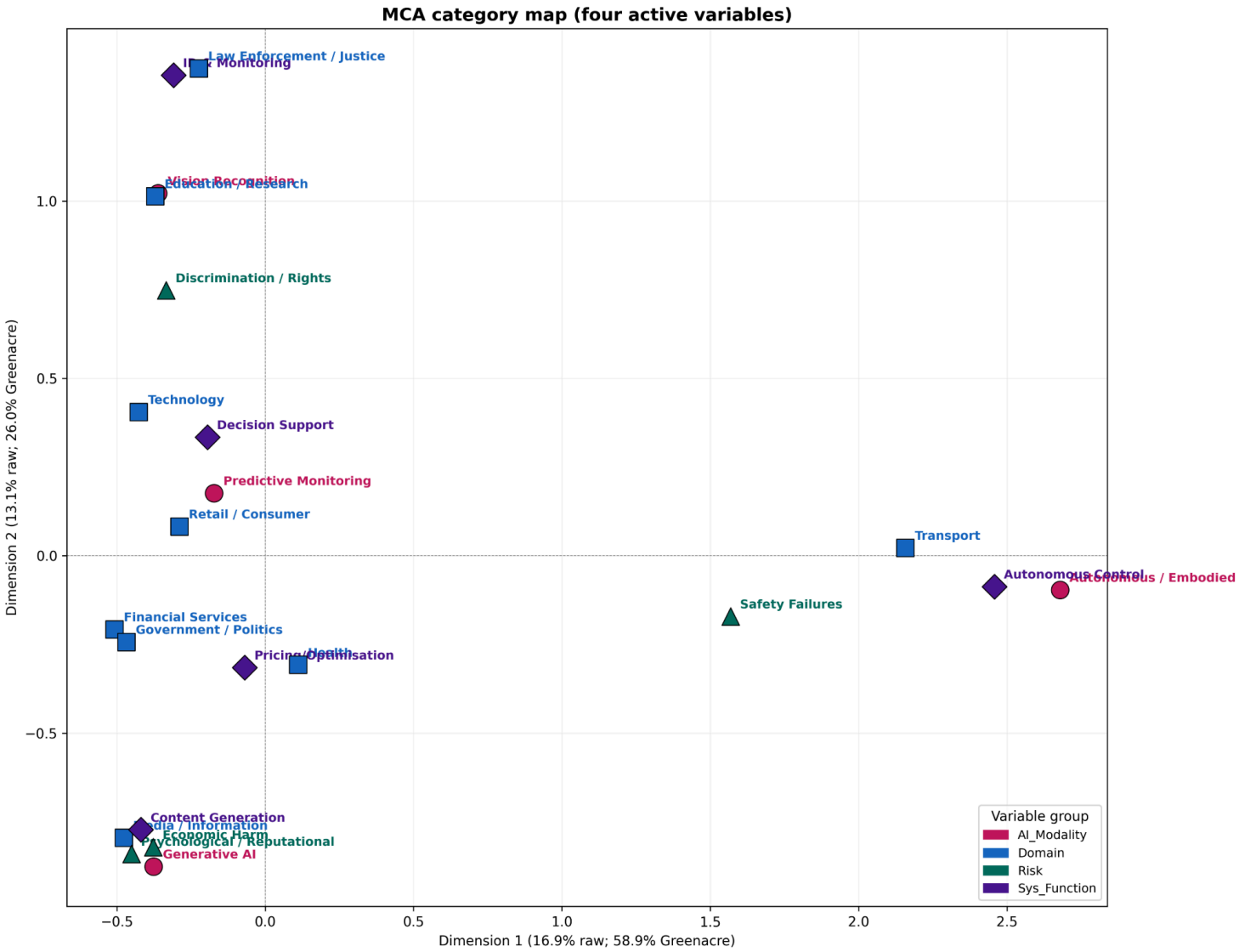

Figure 1

MCA column map on Dimensions 1 and 2, with category coordinates colored by active variable

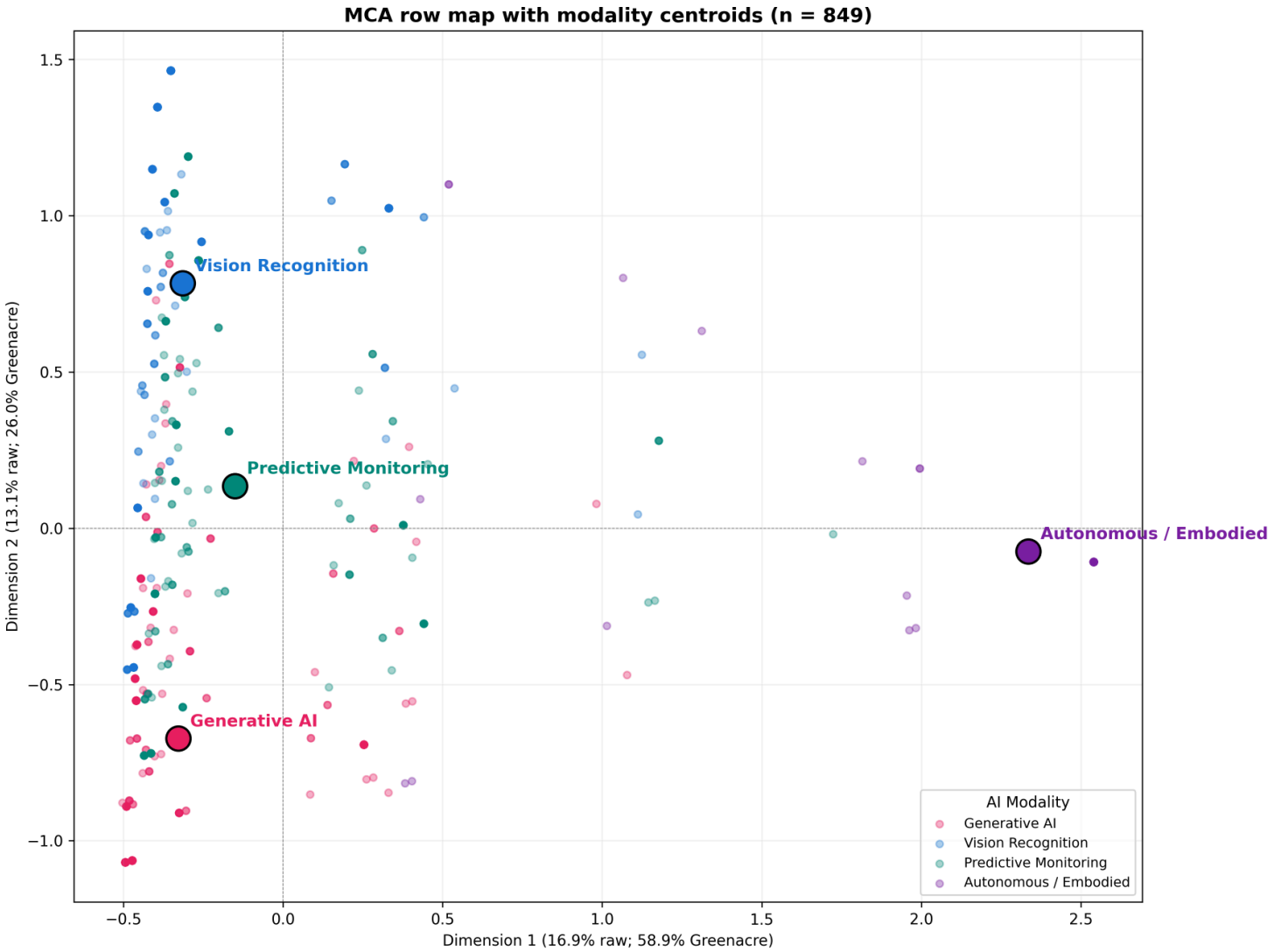

Figure 2

MCA row map on Dimensions 1 and 2; each point is an incident, colored by Ward cluster assignment (k = 4)

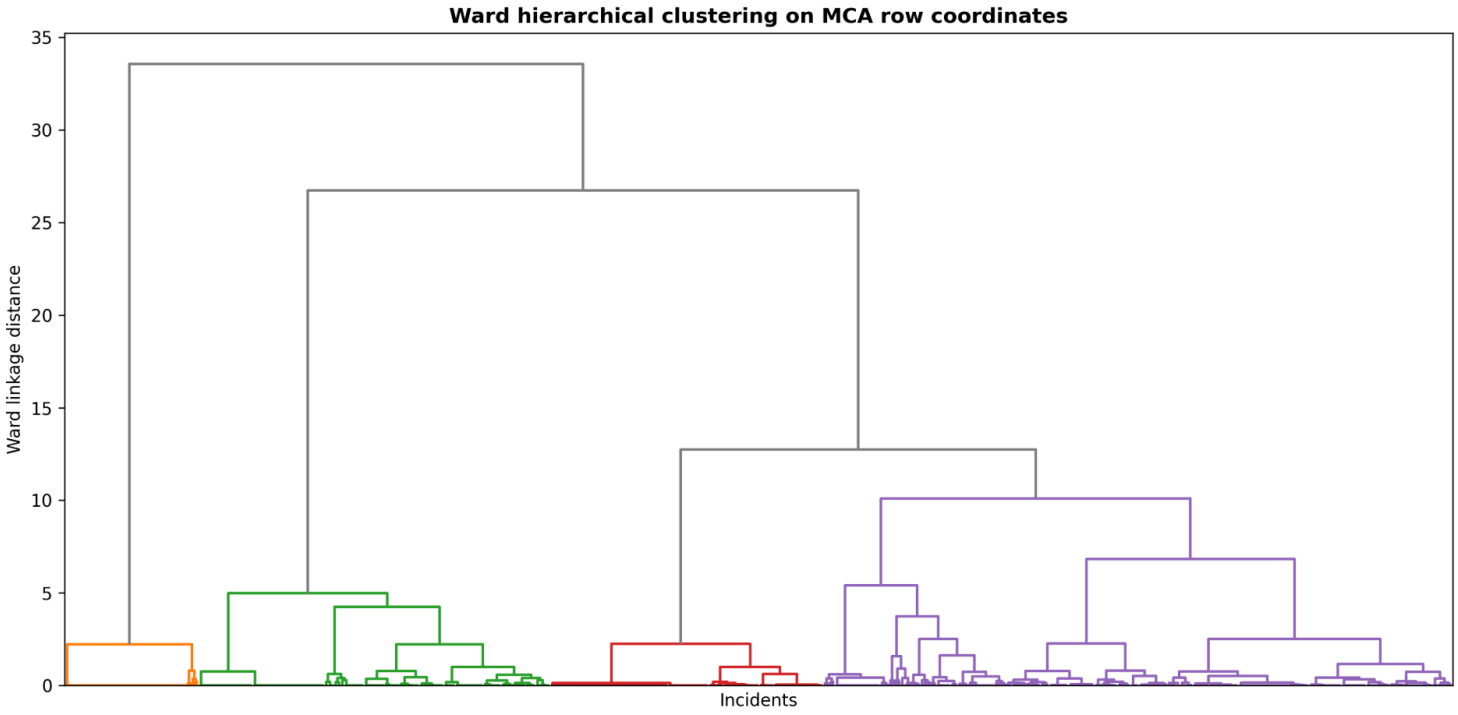

Figure 3

Ward hierarchical clustering dendrogram on MCA row coordinates; the horizontal cut corresponds to k = 4 clusters

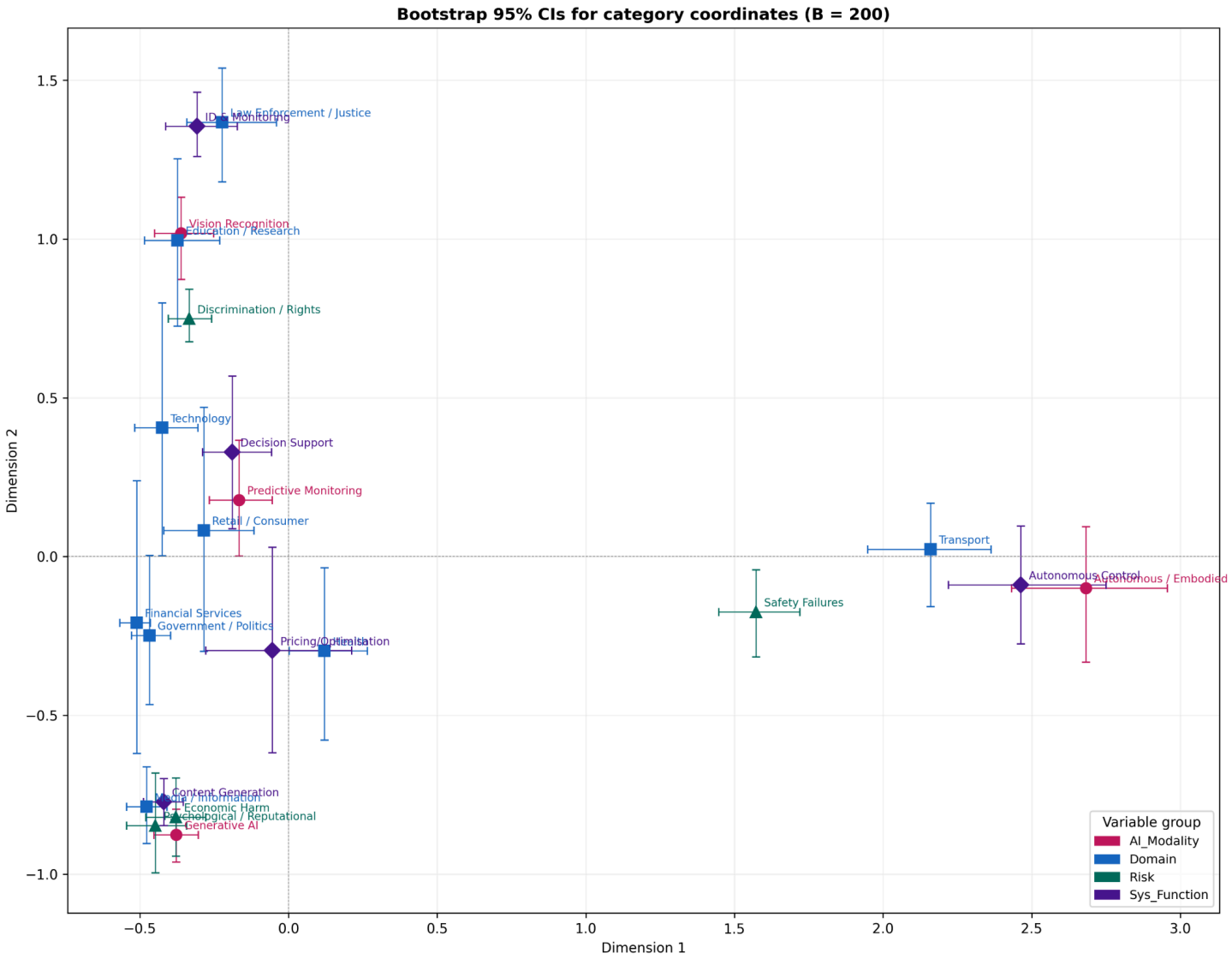

Figure 4

Bootstrap 95% confidence intervals for each active category on Dimensions 1 and 2 (B = 200 resamples)

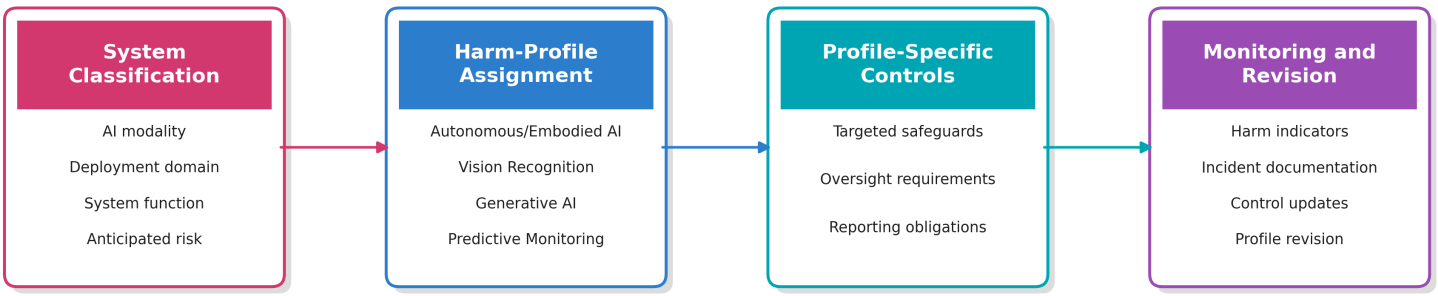

Figure 5

Four-stage governance pipeline: system context, harm-profile mapping, governance response, and adaptive revision