

T3-DEF: An AI Risk Framework for Recursive Socio-Technical Vulnerability and Verification Diversity

Dae-Ryung Lee

metav7388@gmail.com

TECH UNIVERSITY OF KOREA <https://orcid.org/0009-0009-6554-2613>

Research Article

Keywords: AI Risk, AI Risk Framework; T3-DEF; Socio-Technical Systems; Automation Density; Verification Diversity; Human–AI Interaction; AI Governance; Recursive Vulnerability

Posted Date: September 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10881279/v2>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

T3-DEF: An AI Risk Framework for Recursive Socio-Technical Vulnerability and Verification Diversity

Abstract

Artificial intelligence (AI) is increasingly embedded in socio-technical systems where perception, decision support, human oversight, and organizational routines interact dynamically. Existing AI risk-management and governance frameworks provide important guidance for managing and overseeing AI, but they offer more limited explanation of how localized vulnerabilities may propagate across operational and institutional layers. This study proposes T3-DEF (Three-Layer Distortion and Embedding of Failure), an exploratory mechanism-oriented framework comprising three interacting layers: T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding. The framework is complemented by two structural dimensions: Automation Density (AD), representing the concentration of automated perception, recommendation, execution, and reduced human override, and Verification Diversity (VD), representing the independence, diversity, challenge capability, and intervention authority of verification structures.

T3-DEF was operationally grounded using 100 maritime accident cases and exploratorily examined using 70 cross-domain AI-risk cases. The maritime corpus was used for case-level AD and VD coding, while the cross-domain corpus was used to examine whether T1–T2–T3 mechanisms were observable beyond the maritime domain. Across the maritime cases, mean AD was 0.504 and mean VD was 0.474. AD increased substantially across technological eras, whereas VD remained comparatively stable. The association between AD and VD was modest and negative (Pearson $r = -0.287$; Spearman $\rho = -0.247$).

These findings support a conditional rather than deterministic interpretation of AI-related vulnerability. Increasing automation does not necessarily imply declining verification diversity; rather, vulnerability may become more pronounced when automation concentration increases while verification structures remain insufficiently diverse, independent, or actionable. T3-DEF is therefore presented not as a predictive risk-scoring model, but as an exploratory, falsifiable, and extensible analytical framework that complements existing AI governance, regulatory, and management-system approaches by examining how socio-technical vulnerability may emerge, propagate, and become institutionally embedded in AI-assisted systems.

Keywords

AI Risk, AI Risk Framework; T3-DEF; Socio-Technical Systems; Automation Density; Verification Diversity; Human–AI Interaction; AI Governance; Recursive Vulnerability

1. Introduction

Artificial intelligence (AI) is increasingly embedded in socio-technical systems in which automated perception, algorithmic recommendation, human judgment, organizational procedures, and institutional accountability interact continuously. In such environments, vulnerability may arise not only from technical malfunction or model error, but also from interactions among information representation, human reliance, organizational routines, and weakened opportunities for independent verification [1–8]. As AI systems assume more consequential roles in monitoring, prioritization, recommendation, and operational decision support, understanding these interactions becomes increasingly important.

Research on automation and human–machine interaction has long shown that technological assistance can improve performance while also introducing dependency, automation bias, reduced situation awareness, and difficulties in maintaining effective human control [9–17]. System-safety and resilience research has similarly demonstrated that major failures often emerge through interactions among technical, human, and organizational components rather than from isolated component failure [3–8,18]. Dependability research further emphasizes redundancy, fault tolerance, and verification as important properties of reliable systems [19]. Together, these literatures establish that the safety of AI-assisted systems cannot be understood solely through technical performance.

Contemporary AI risk-management and governance frameworks address important aspects of this challenge. The NIST AI Risk Management Framework provides a structured approach to governing and managing AI risks across the lifecycle [20], while regulatory and national governance initiatives such as the EU AI Act and Korea's AI Basic Act emphasize risk management, accountability, transparency, and human oversight [21,22]. These approaches are essential for governance and compliance. However, a complementary analytical question remains: how can a localized distortion in an AI-assisted system gain decision influence, become reinforced through repeated operation, and eventually become embedded within organizational practice?

This question is especially relevant because the formal presence of human oversight does not necessarily guarantee effective independent verification. Human operators may retain nominal authority while depending on the same automated information pathway, lacking alternative interpretations, or possessing insufficient practical authority to challenge or correct system outputs. Thus, the analytical problem extends beyond whether oversight exists to whether the surrounding socio-technical system preserves sufficiently independent and actionable verification mechanisms.

To address this gap, this study proposes T3-DEF (Three-Layer Distortion and Embedding of Failure) as an exploratory, mechanism-oriented AI risk framework. T3-DEF conceptualizes recursive socio-technical vulnerability through three interacting layers: T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding. The framework is complemented by two structural dimensions: Automation Density (AD) and Verification Diversity (VD). AD characterizes the concentration of automated perception, recommendation, execution, and reduced human override, whereas VD characterizes the independence, diversity, challenge capability, and intervention authority of verification structures.

A central premise of T3-DEF is that automation concentration and verification diversity should not be treated as simple opposites. Increasing automation does not necessarily imply declining verification diversity, nor is high automation inherently hazardous. Rather, vulnerability may become more pronounced under particular socio-technical configurations in which automation becomes more concentrated while verification mechanisms remain insufficiently diverse, independent, or actionable. The framework therefore adopts a conditional and configuration-based interpretation of AI-related vulnerability.

The present study provides initial operational grounding for this framework using two complementary corpora. One hundred maritime accident cases are used for case-level examination of AD and VD because maritime investigation reports provide comparatively detailed information on operational processes, automation roles, human intervention, and organizational conditions. Seventy cross-domain AI-risk cases are used separately for exploratory examination of whether T1–T2–T3 mechanisms are observable beyond the maritime domain. The two corpora serve different analytical purposes and are not treated as a single homogeneous statistical sample. This distinction is consistent with the exploratory purpose of the study and avoids overstating the empirical status of the framework.

The study is guided by three research questions:

RQ1. How can socio-technical vulnerability propagate recursively across perceptual, decisional, and institutional layers in AI-assisted systems?

RQ2. How can Automation Density and Verification Diversity characterize structural conditions associated with recursive socio-technical vulnerability?

RQ3. To what extent are T3-DEF mechanisms observable across maritime operational cases and heterogeneous cross-domain AI-risk cases?

This study makes three contributions. First, it proposes a three-layer mechanism for examining how vulnerability may propagate from information representation to decision influence and institutional embedding. Second, it operationalizes AD and VD as complementary structural dimensions for examining automation concentration and verification architecture. Third, it positions T3-DEF as a complementary analytical layer that may connect AI governance and management frameworks with the operational mechanisms through which vulnerability emerges, gains decision influence, and becomes institutionally embedded.

These contributions should be interpreted within the exploratory scope of the study. The maritime corpus is purposively assembled rather than statistically representative, AD and VD remain heuristic constructs requiring independent validation, and the cross-domain corpus is used for mechanism-level exploration rather than formal quantitative validation.

T3-DEF is not proposed as a validated predictive model, universal risk score, or fixed numerical threshold. Rather, it is intended as an exploratory and falsifiable framework whose propositions can be independently tested, refined, narrowed, or rejected. The following sections position T3-DEF within relevant socio-technical and AI-risk literature, develop its conceptual structure, describe its operationalization, and examine its initial empirical patterns.

2. Related Work and Conceptual Positioning

2.1. Socio-Technical Systems and Systemic Safety

Research on complex socio-technical systems has long shown that major failures cannot be adequately explained by isolated technical malfunction or individual human error. Rasmussen conceptualized risk as emerging from interactions among technological systems, human operators, organizational structures, management priorities, and regulatory environments [3]. Reason likewise demonstrated that adverse outcomes often arise through the alignment of latent organizational conditions and weakened defenses rather than from a single frontline error [4]. Perrow further emphasized that tightly coupled and interactively complex systems can produce unexpected failure pathways that are difficult to anticipate or control [18].

Leveson's systems-theoretic approach extended this perspective by treating safety as a problem of inadequate constraints, interactions, and feedback across multiple system levels [8]. These approaches are especially relevant to AI-assisted environments, where perception, recommendation, human interpretation, execution, and organizational routines may interact dynamically. They provide the broader socio-technical foundation on which T3-DEF is developed.

T3-DEF does not replace these theories. Its narrower purpose is to examine how vulnerability may propagate from distorted or incomplete information, through decision processes, and into organizational practice.

2.2. Resilience Engineering and Adaptive Capacity

Resilience engineering emphasizes the capacity of socio-technical systems to monitor, adapt, respond, and recover under changing or uncertain conditions [5–7,23]. Safety therefore depends not only on preventing component failure, but also on preserving the ability to detect emerging problems, question dominant interpretations, and intervene when conditions deviate from expectations.

This perspective is important for AI-assisted systems because formal oversight does not necessarily imply effective adaptive capacity. A human operator may remain present while relying on the same automated information channel as the system being monitored. Under such conditions, the practical ability to independently verify or challenge automated outputs may be limited.

This concern provides an important conceptual foundation for Verification Diversity, which focuses on the functional independence and actionability of verification pathways rather than on the mere presence of multiple actors or monitoring channels.

2.3. Automation Dependency and Human Oversight

The risks associated with automation dependence have been recognized for decades. Bainbridge described the "ironies of automation," in which automation reduces routine human involvement while increasing the importance of effective human intervention during abnormal situations [9]. Parasuraman and Riley distinguished appropriate use of automation from misuse, disuse, and abuse [1], while Parasuraman, Sheridan, and Wickens showed that automation can

redistribute functions across information acquisition, analysis, decision selection, and action implementation [10].

Related research has documented automation bias, reduced supervisory engagement, degraded situation awareness, and difficulty in maintaining effective human control [11,12,16,17]. These findings are directly relevant to AI-assisted decision environments, where automated systems increasingly participate not only in execution but also in filtering, ranking, recommendation, and interpretation.

T3-DEF builds on this literature by distinguishing the concentration of automation from the quality of verification architecture. This distinction avoids treating automation itself as an inherently unsafe condition.

2.4. Human–AI Coordination and Effective Oversight

Research on human–AI collaboration emphasizes that system performance depends on more than algorithmic capability. Effective operation requires coordination, appropriate reliance, transparency, feedback, controllability, and meaningful allocation of responsibility [13–15].

A particularly important issue is the distinction between nominal human involvement and effective human oversight. A human operator may formally approve an AI recommendation while lacking independent information, alternative interpretations, sufficient time, or effective intervention authority. In such circumstances, human presence alone may provide only limited protection against automation-related vulnerability.

This distinction is relevant to T3-DEF because the framework treats meaningful verification as a socio-technical capability rather than a binary human-in-the-loop condition.

2.5. Dependability, Redundancy, and Verification

Dependability research has long emphasized fault detection, redundancy, recovery, and tolerance of component failure [19]. However, multiple channels do not necessarily create meaningful verification if they depend on the same underlying data, assumptions, algorithms, or organizational structures.

This distinction is increasingly important in AI-assisted systems. Multiple sensors, dashboards, models, or reviewers may appear to provide redundancy while still reproducing the same informational or decision pathway. Therefore, numerical redundancy should not be equated automatically with independent verification.

T3-DEF draws on this insight by conceptualizing Verification Diversity in terms of functional independence, interpretive diversity, challenge capability, and intervention authority.

2.6. AI Risk Management and Governance

Contemporary AI safety, ethics, risk-management, and governance research addresses important dimensions of trustworthy AI, including technical safety concerns and broader ethical and socio-technical risks [24,25]. The NIST AI Risk Management Framework organizes AI risk management around the functions of GOVERN, MAP, MEASURE, and MANAGE [20]. The European Union AI Act establishes regulatory obligations according to risk and application context [21]. Korea's AI Basic Act provides a national governance framework emphasizing trust, safety, risk management, and human supervision [22].

These frameworks are important for organizational governance, regulatory compliance, and accountability. However, their principal purpose differs from that of a mechanism-oriented analytical framework. They specify how AI risk should be governed, managed, or regulated, but they do not primarily seek to explain how a localized informational distortion may acquire decision authority, become reinforced through repeated operation, and eventually become embedded in organizational practice.

This distinction defines the conceptual space in which T3-DEF is positioned.

2.7. Positioning T3-DEF

T3-DEF is positioned as a complementary, mechanism-oriented analytical framework for examining how vulnerability may emerge, propagate, and become institutionally embedded in AI-assisted socio-technical systems. Its central analytical question is:

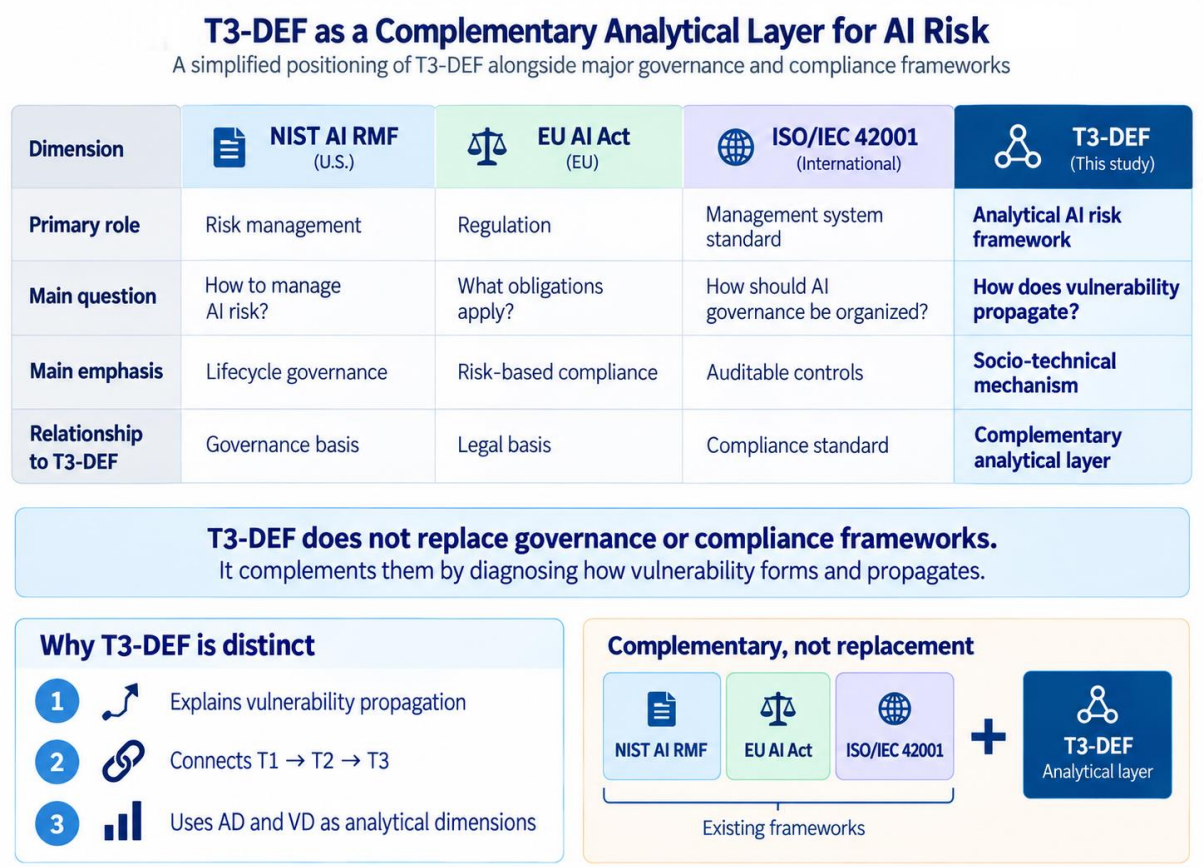
How does socio-technical vulnerability emerge, propagate, and become institutionally embedded in AI-assisted systems?

This complementary role is particularly relevant to established governance, regulatory, and management frameworks. The NIST AI Risk Management Framework structures organizational AI risk management through the functions of GOVERN, MAP, MEASURE, and MANAGE; the EU AI Act establishes risk-based regulatory obligations and compliance requirements; and ISO/IEC 42001 provides an organizational AI management-system structure for establishing, implementing, maintaining, monitoring, and continually improving AI governance and management processes.

T3-DEF does not duplicate or replace these functions. Instead, it offers a mechanism-oriented analytical lens for examining how vulnerability may develop and propagate through perceptual, decisional, and institutional processes within the socio-technical environments governed by such frameworks. In particular, T3-DEF examines how localized distortion may influence decision-making, acquire greater operational influence through repeated reliance, and ultimately become embedded in organizational routines and practices.

In this sense, governance, regulatory, and management-system frameworks help organizations determine what should be governed, controlled, assessed, or improved, whereas T3-DEF focuses on how vulnerability may form, propagate, and become institutionally embedded.

Figure 1. Comparative Positioning of T3-DEF alongside NIST AI RMF, the EU AI Act, and ISO/IEC 42001



Amplification, and T3 Institutional Embedding. Automation Density (AD) and Verification Diversity (VD) complement these mechanisms by characterizing the surrounding structural conditions of automation concentration and verification architecture.

Taken together, the frameworks should therefore be viewed as complementary rather than competing. Governance, regulatory, and management-system approaches provide structures for managing, regulating, and controlling AI-related risk, while T3-DEF provides an analytical vocabulary for examining how vulnerability may emerge, gain decision influence, recur through feedback, and become embedded within socio-technical systems.

Accordingly, T3-DEF contributes not by proposing a universal risk score or replacing established governance structures, but by offering a transparent and falsifiable analytical structure that future studies can independently test, refine, narrow, or apply across different AI-assisted domains.

3. The T3-DEF Framework

3.1. Conceptual Foundation

T3-DEF (Three-Layer Distortion and Embedding of Failure) is proposed as a mechanism-oriented framework for analyzing how vulnerability may emerge and propagate within AI-assisted socio-technical systems. Its purpose is not to replace established approaches to system safety, resilience engineering, human–automation interaction, dependability, or AI governance, but to provide a focused analytical structure for examining how localized vulnerability may acquire decision influence and become embedded in organizational practice.

The framework comprises three interacting layers:

T1: Perceptual Distortion

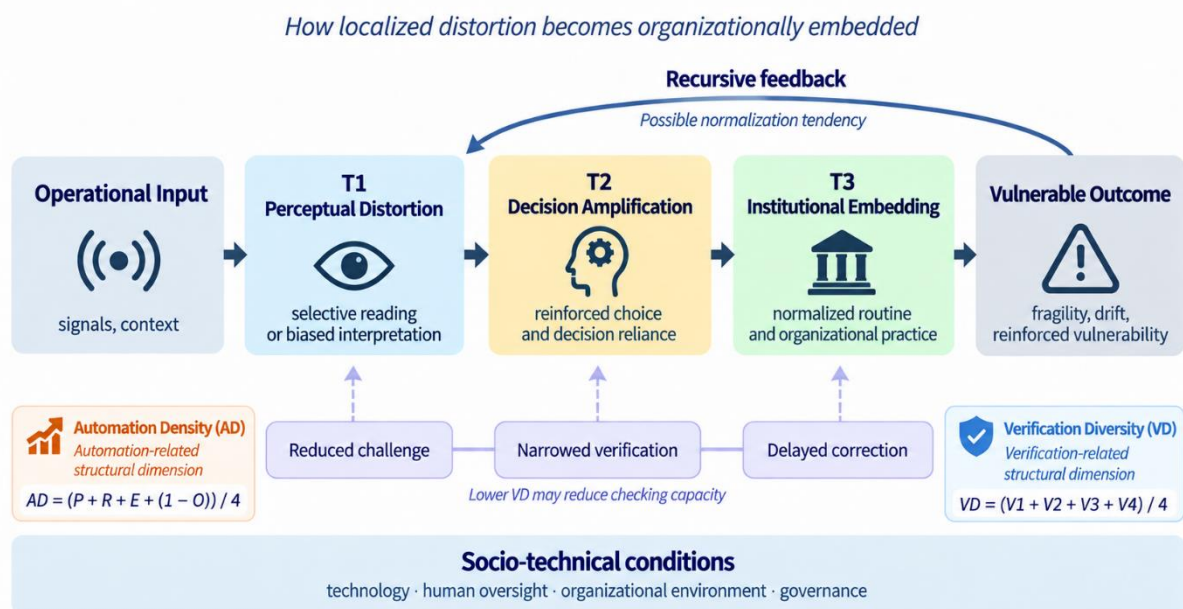
T2: Decision Amplification

T3: Institutional Embedding

These layers are not assumed to operate as a rigid linear sequence. They may overlap, recur, or reinforce one another through feedback. T3-DEF therefore treats vulnerability as a potentially recursive socio-technical process rather than as a single chain of technical or human failure.

Two structural dimensions complement the three-layer mechanism: Automation Density (AD) and Verification Diversity (VD). AD characterizes the concentration of automated functions within an operational workflow, whereas VD characterizes the independence and effectiveness of the system's verification architecture.

Figure 2. Conceptual Architecture of the T3-DEF AI Risk Framework.



The figure illustrates the conceptual relationship among T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding, together with Automation Density (AD) and Verification Diversity (VD) as complementary structural dimensions. AD and VD are presented as contextual conditions that may influence how vulnerability is reinforced, interrupted, or redirected within AI-assisted socio-technical systems. The arrows indicate

potential pathways of influence and recursive feedback rather than deterministic causal effects or universally validated sequences. Broader technological, human, organizational, and regulatory factors are treated as contextual influences surrounding the T3-DEF mechanism.

3.2. T1: Perceptual Distortion

T1 Perceptual Distortion refers to conditions in which the representation of an operational situation is selectively filtered, reduced, emphasized, or otherwise altered before or during interpretation.

Such distortion may arise from sensor limitations, data preprocessing, interface design, ranking, filtering, model uncertainty, missing contextual information, or selective salience. T1 does not require an objectively false AI output. A technically plausible output may still create vulnerability if it narrows the information available to the decision-maker or suppresses relevant alternatives.

The central T1 question is:

What information reaches the decision-maker, through which pathway, and in what representational form?

T1 therefore focuses on the informational conditions that shape subsequent decision processes.

3.3. T2: Decision Amplification

T2 Decision Amplification describes the process through which a particular representation, recommendation, or interpretation acquires increasing influence over subsequent decisions.

Amplification may occur through automated recommendations, prioritization, confidence indicators, defaults, optimization outputs, or repeated reliance on apparently consistent system outputs. The central issue is not simply whether a human remains formally involved, but whether the human retains sufficient independent basis to evaluate or challenge the automated recommendation.

The central T2 question is:

How does a local representation acquire additional weight or authority within the decision process?

T2 therefore captures the transition from informational influence to decisional influence.

3.4. T3: Institutional Embedding

T3 Institutional Embedding refers to the process through which repeated decisions, assumptions, or system-mediated practices become incorporated into organizational routines, procedures, responsibility structures, or accepted norms.

Once embedded, a vulnerable practice may persist beyond the original event or individual operator. Repeated reliance on a particular information source, recommendation system, or workflow may gradually shape standard operating practice and expectations.

The central T3 question is:

How does a local decision pattern become normalized or embedded within organizational practice?

T3 therefore connects operational decision processes with longer-term organizational structure and governance.

3.5. Recursive Propagation of Vulnerability

The defining feature of T3-DEF is the possibility of recursive feedback among perception, decision, and institutional practice.

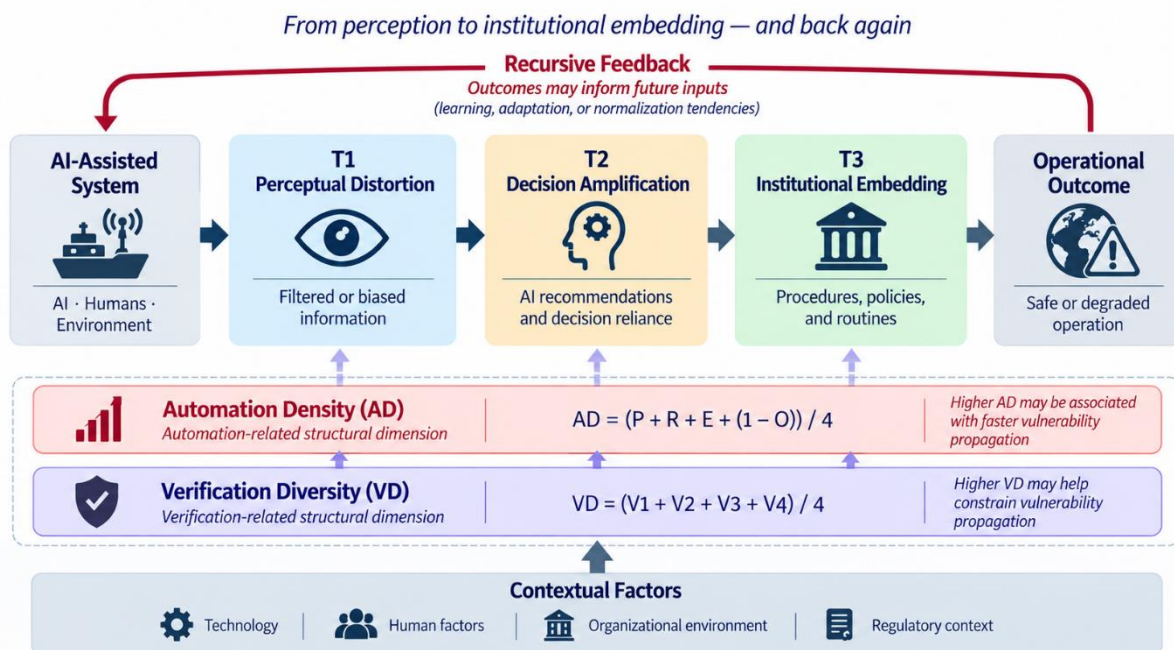
A simplified pathway is:

T1 → T2 → T3 → Operational Outcome → Updated System Conditions

Operational outcomes may reinforce confidence in existing system configurations, alter procedures, influence training, reshape future data collection, or modify expectations regarding human intervention. In this way, prior outcomes may affect subsequent perception and decision processes.

However, recursion does not necessarily produce escalating vulnerability. Feedback may also strengthen verification, improve procedures, or trigger redesign. T3-DEF therefore accommodates both vulnerability-reinforcing feedback and corrective feedback.

Figure 3. Recursive Propagation of Vulnerability in T3-DEF.



The figure illustrates a potential recursive pathway in which T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding may interact over time within AI-assisted socio-technical systems. Operational outcomes may feed back into subsequent perception, decision processes, organizational routines, and verification practices. Automation Density (AD) and Verification Diversity (VD) are shown as structural conditions that may be associated with the reinforcement or interruption of vulnerability propagation. The recursive relationships depicted here are conceptual and conditional rather than deterministic, and their manifestation may vary according to technological, human, organizational, and regulatory context.

3.6. Automation Density

Automation Density (AD) represents the relative concentration of automated or AI-assisted functions within the operational workflow.

AD is operationalized through four dimensions:

- ***P* — Perception:** the extent to which information acquisition and interpretation are automated;
- ***R* — Recommendation:** the extent to which automated systems generate or shape decision options;
- ***E* — Execution:** the extent to which operational actions are carried out automatically;
- ***O* — Override:** the extent to which humans retain effective authority to intervene or retake control.

The normalized heuristic formulation is:

$$AD = \frac{P + R + E + (1 - O)}{4}$$

The inverse treatment of *O* reflects the principle that stronger effective human override reduces the concentration of operational authority within automation.

AD is used as a **comparative structural descriptor**. It should not be interpreted as a probability of failure, validated engineering risk score, or universal threshold.

3.7. Verification Diversity

Verification Diversity (VD) represents the structural diversity and independence of pathways available for checking, contesting, or correcting automated or AI-assisted outputs.

VD is operationalized through four dimensions:

- ***V1* — Information-Source Independence**
- ***V2* — Interpretive Diversity**
- ***V3* — Challenge Capability**
- ***V4* — Independent Intervention Authority**

The heuristic formulation is:

$$VD = \frac{V1 + V2 + V3 + V4}{4}$$

VD differs from simple redundancy. Multiple actors, sensors, or systems do not necessarily provide meaningful verification if they rely on the same data, assumptions, algorithms, or decision pathways.

VD is therefore treated as a structural property of the verification architecture rather than as a simple count of monitoring channels.

3.8. AD–VD Configuration and Central Proposition

A central feature of T3-DEF is that AD and VD are treated as **partially independent dimensions**.

The framework does not assume:

$$AD\uparrow \not\Rightarrow VD\downarrow$$

Instead, different socio-technical configurations are possible, including high AD/high VD, high AD/low VD, low AD/high VD, and low AD/low VD.

This distinction shifts attention away from automation itself and toward the configuration of automation and verification structures. A highly automated system may remain resilient if independent verification and effective intervention are preserved. Conversely, a less automated system may remain vulnerable when information sources, challenge mechanisms, or intervention authority are weak.

T3-DEF therefore advances the following conditional proposition:

Recursive socio-technical vulnerability may become more likely when increasing automation concentration is combined with insufficiently diverse, independent, or actionable verification structures.

This proposition does not imply that automation causes vulnerability or that low VD inevitably produces failure. The relationship may also be shaped by contextual factors such as operator workload, organizational culture, training, interface design, management incentives, and regulatory conditions.

3.9. Scope and Falsifiability

T3-DEF is proposed as an **exploratory and falsifiable analytical framework**, not as a completed theory or predictive risk model.

Its central propositions are therefore open to independent testing, refinement, narrowing, or rejection. The present study provides initial operational grounding through maritime and cross-domain cases, while the reliability, validity, and generalizability of AD, VD, and the T1–T2–T3 mechanism require further independent examination.

The methodology used for this initial examination is described in Section 4.

4. Research Methodology

4.1 Research Design

This study adopted an exploratory multi-case design to provide initial operational grounding for the T3-DEF framework [26,27]. The analysis was designed to examine whether the proposed T1–T2–T3 mechanism could be identified in documented socio-technical events and whether Automation Density (AD) and Verification Diversity (VD) could be operationalized consistently enough to support structured comparative analysis.

Two complementary corpora were used for different analytical purposes. A primary corpus of **100 maritime accident and incident cases** was used for case-level operationalization of AD and VD. A second corpus of **70 cross-domain AI-risk cases** was used to examine whether T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding were observable beyond the maritime domain.

The two corpora were not treated as a single homogeneous statistical sample because their source structure, level of operational detail, and availability of evidence concerning automation and verification differed substantially.

Table 1. Analytical Roles of the Maritime and Cross-Domain Case Corpora

Corpus	N	Primary Analytical Role	AD/VD Quantitative Coding	Use in Combined Analysis
Maritime accident corpus	100	Primary operational grounding of T3-DEF	Yes	Descriptive comparison
Cross-domain AI-risk corpus	70	Exploratory replication of T1–T2–T3 mechanisms	No	Descriptive cross-domain examination
Combined corpus	170	Examination of framework applicability across domains	Not treated as a homogeneous quantitative sample	Mechanism-level comparison only

4.2. Maritime Case Corpus

The maritime corpus comprises 100 accident and incident cases spanning different vessel types, accident categories, operational environments, and technological periods. Maritime cases were selected because formal investigation reports generally provide comparatively detailed information concerning operational processes, navigation and control systems, automation use, human intervention, monitoring, and organizational conditions.

The maritime corpus was purposively assembled to support mechanism-oriented and comparative analysis rather than statistical representation of the global population of maritime accidents. Cases were retained when publicly available investigation reports or other authoritative documentary sources provided sufficient information to reconstruct the operational context and to assess, at least at a case level, automation roles, human intervention or override, verification pathways, and relevant socio-technical conditions. The corpus was

also assembled to preserve variation across accident types, vessel types, technological periods, and operational contexts. Cases for which the available documentation did not provide sufficient evidence for meaningful case-level assessment of the relevant T3-DEF dimensions were not used for the primary AD–VD analysis. Accordingly, the resulting 100-case corpus should be interpreted as a purposive analytical sample rather than as a statistically representative sample of maritime accidents.

The unit of analysis was the **case-level operational event** rather than an individual operator, device, or isolated technical component. Each case was examined as a socio-technical configuration involving interactions among technical systems, human operators, organizational practices, and contextual conditions.

Available documentary evidence was reviewed for information concerning automated perception, recommendation and decision support, execution, human override capability, verification pathways, intervention structures, and relevant T3-DEF mechanisms.

4.3. Cross-Domain AI-Risk Corpus

The second corpus comprises 70 cross-domain AI-risk cases derived from the publicly available AI risk case collection used in this study [28]. These cases were used to examine whether patterns consistent with the T1–T2–T3 mechanism could be observed across heterogeneous AI application contexts.

The cross-domain cases were **not subjected to the same quantitative AD/VD analysis** as the maritime corpus. Many did not provide sufficient operational evidence concerning automated execution, effective override authority, information-source independence, or verification architecture to support measurement-equivalent AD and VD coding.

Accordingly, the cross-domain corpus was used for exploratory mechanism-level examination of whether cases showed evidence consistent with:

- T1 Perceptual Distortion,
- T2 Decision Amplification,
- T3 Institutional Embedding,
- or combinations of these mechanisms.

This distinction prevents the 170 cases from being interpreted as a single measurement-equivalent population.

4.4. Coding Procedure and Unit of Analysis

The analytical procedure involved four stages:

1. extraction of relevant evidence from the available case documentation;
2. identification of T1–T2–T3 mechanisms;
3. structural coding of AD and VD independently of the T1–T2–T3 mechanism classification for the maritime corpus;
4. documentation of the coding rationale and case-level verification status.

Mechanism classification and structural coding were treated separately. T1, T2, and T3 describe how vulnerability may propagate, whereas AD and VD describe the structural conditions surrounding that propagation.

AD was not inferred from accident severity, technological era, failure type, or T-layer classification. Likewise, VD was not inferred from the AD value or from the simple number of actors, sensors, or monitoring systems present.

The coding procedure followed a structured qualitative content-analysis logic, with predefined analytical categories and documented case-level rationales used to improve transparency and reproducibility [29,30].

This separation was used to reduce circularity between the theoretical mechanism and the structural indicators.

4.5. Operationalization of Automation Density

Automation Density represents the relative concentration of automated or AI-assisted functions in the operational workflow.

Four dimensions were used:

- ***P* — Automated Perception**
- ***R* — Operational Recommendation**
- ***E* — Autonomous Execution**
- ***O* — Human Override Capability**

AD was calculated as:

$$AD = \frac{P + R + E + (1 - O)}{4}$$

The inverse treatment of $(1 - O)$ reflects the assumption that stronger effective human override reduces the concentration of operational authority within automation.

The four dimensions were assigned using predefined ordinally anchored heuristic criteria, recognizing that ordinal categories support ordered comparison without necessarily possessing interval-scale properties [31].

Detailed interpretive anchors and guidance for intermediate *P*, *R*, *E*, and *O* scores are provided in Supplementary Dataset S2 (AD_Rubric sheet). Intermediate values represent documentary judgments positioned between the descriptive anchor conditions and should not be interpreted as validated interval-scale increments.

AD is therefore interpreted as a **comparative structural descriptor**, not as a probability of failure, causal coefficient, validated interval-scale measure, or engineering risk threshold.

4.6. Operationalization of Verification Diversity

Verification Diversity represents the degree to which a system preserves functionally independent and actionable pathways for checking, questioning, and correcting automated or AI-assisted outputs.

VD comprises four dimensions:

- ***V1* — Information-Source Independence**
- ***V2* — Interpretive Diversity**
- ***V3* — Cross-Check / Challenge Capability**
- ***V4* — Independent Intervention Authority**

VD was calculated as:

$$VD = \frac{V1 + V2 + V3 + V4}{4}$$

VD differs from numerical redundancy. Multiple sensors, operators, systems, or organizational actors do not necessarily constitute diverse verification if they rely on common data, assumptions, algorithms, or decision pathways.

Higher VD therefore reflects greater functional independence in information, interpretation, challenge, and intervention.

Like AD, VD is an **ordinally anchored heuristic construct** and is not interpreted as a validated causal or interval-scale measure [31].

Detailed VD anchor definitions, coding principles, and independence rules are provided in Supplementary Dataset S2 (VD_Rubric sheet).

Table 2. Operational Coding Anchors for Automation Density and Verification Diversity

Dimension	Low Anchor	Neutral / Intermediate Anchor	High Anchor
P – Perception	Human/manual perception dominates	Shared human–system perception	Automated perception dominates
R – Recommendation	Human develops decision alternatives	System supports or recommends options	System strongly determines preferred options
E – Execution	Human executes actions	Shared or assisted execution	Automated execution dominates
O – Override	Immediate and effective human control	Conditional or constrained override	Little or ineffective human override
V1 – Independence	Single/dependent information pathway	Limited or uncertain independence	Multiple functionally independent sources
V2 – Diversity	Single interpretive pathway	Standard cross-checking	Diverse and competing interpretations
V3 – Challenge	Weak/no meaningful challenge	Formal but limited challenge	Strong ability to question and correct
V4 – Authority	Nominal or ineffective intervention	Constrained intervention	Independent and effective intervention authority

4.7. Evidence-Limited and Neutral-Anchor Coding

When available documentary evidence did not support classification toward either a lower or higher anchor, a neutral exploratory value was retained rather than assigning a directional score without adequate evidence.

This issue was particularly relevant to VD because accident investigation reports often describe immediate technical and operational causes more extensively than information-source independence, organizational challenge structures, or intervention authority.

A neutral value of **0.50** should therefore not be interpreted automatically as evidence of objectively moderate verification diversity. Instead, such observations indicate that the documentary evidence did not justify stronger directional classification.

These cases were retained as evidence-limited observations and interpreted cautiously in the subsequent analysis.

4.8. Technological Era Classification

The maritime cases were grouped into three broad technological eras based on the year of the accident: **Era 1 (1987–2004, N = 29)**, **Era 2 (2005–2015, N = 43)**, and **Era 3 (2016–2024, N = 28)**. These era categories were used as descriptive temporal–technological contexts for examining changes in Automation Density (AD) and Verification Diversity (VD) across the corpus. They were not intended to represent universal stages of maritime automation or fixed technological thresholds.

Era classification was treated solely as a contextual descriptor. A later-era case did not automatically receive a higher AD value, and era membership was not used to infer VD, T1–T2–T3 classification, or failure type. This separation was maintained to reduce circularity between temporal context and the structural variables being analyzed.

4.9. Case-Level Review and Computational Verification

During revision, the maritime coding was re-examined at the case level to identify inconsistencies in classification, component scoring, and supporting rationale. Boundary or causally complex cases were reviewed individually, and final coding decisions were documented in Supplementary Dataset S2.

The revised analysis does not claim a newly established formal inter-rater reliability statistic for the complete 100-case corpus. The coding should therefore be interpreted as a structured and auditable case-level analysis rather than as an independently replicated multi-coder validation study.

Computational verification was used to recalculate Automation Density (AD) from the four component scores—P, R, E, and O—using the specified heuristic formulation, and to recalculate Verification Diversity (VD) from V1–V4. These checks were used to identify discrepancies between recorded values and formula-derived values and to confirm arithmetic consistency across the maritime dataset.

Detailed interpretive guidance for the AD and VD coding structures is provided in Supplementary Dataset S2. The **AD_Rubric** sheet documents the descriptive anchor conditions, coding principles, independence rules, and treatment of intermediate P, R, E, and O scores. The **VD_Rubric** sheet documents the anchor definitions, coding principles, and independence rules for V1–V4. These rubric materials are intended to improve transparency and reproducibility of the coding process and should not be interpreted as establishing validated interval-scale measurement properties.

Computational verification establishes arithmetic consistency but does not establish inter-rater reliability, construct validity, or causal validity. Accordingly, the final AD and VD values should be interpreted as exploratory, ordinally anchored heuristic descriptors for comparative analysis.

Supplementary Dataset S1 contains the cross-domain T3-DEF operational coding, while Supplementary Dataset S2 contains the detailed maritime case-level coding, AD and VD rubric documentation, summary statistics, and case-level verification information.

4.10. Statistical Analysis

Quantitative analysis was limited to the 100-case maritime corpus and was exploratory in purpose.

The following analyses were conducted:

- descriptive statistics for AD and VD;
- comparison of mean AD and VD across technological eras;
- Pearson correlation between AD and VD;
- Spearman rank correlation between AD and VD; and
- graphical examination of the case-level relationship using a scatterplot and descriptive LOESS smoother.

Both Pearson and Spearman coefficients were reported because AD and VD are heuristic, ordinally anchored constructs. Pearson correlation provides a descriptive measure of linear association, while Spearman correlation provides a rank-based comparison that is less dependent on interval-scale assumptions.

The analyses were not designed to establish causality, a universal transition point, or a validated numerical threshold. Accordingly, no specific AD value is interpreted as separating safe from unsafe systems, and the LOESS smoother is used only as a descriptive visualization of the observed case distribution.

5. Results

5.1. Overview

The empirical analysis generated two complementary sets of findings. The 100-case maritime corpus provided case-level results for Automation Density (AD), Verification Diversity (VD), technological-era patterns, and the descriptive association between AD and VD. The 70 cross-domain AI-risk cases were examined separately to determine whether patterns consistent with T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding were observable beyond the maritime domain.

Because the two corpora served different analytical purposes, the results are reported separately and are not interpreted as statistical findings from a homogeneous 170-case population.

5.2. Descriptive Characteristics of AD and VD

Across the 100 maritime cases, mean AD was **0.504** and mean VD was **0.474**. AD showed substantially greater dispersion than VD, with a standard deviation of **0.134**, compared with **0.051** for VD. AD values ranged from **0.260 to 0.780**, whereas VD values ranged from approximately **0.313 to 0.563**.

The narrower VD distribution is consistent with the conservative coding approach described in Section 4, particularly the use of neutral anchors where available documentation did not support stronger directional classification. Accordingly, the distribution should be interpreted as a property of the coded corpus rather than as evidence that most cases possessed objectively intermediate verification diversity.

Table 3. Descriptive Statistics for Automation Density and Verification Diversity in the Maritime Corpus

Measure	Automation Density (AD)	Verification Diversity (VD)
N	100	100
Mean	0.504	0.474
Standard deviation	0.134	0.051
Minimum	0.26	0.313
Maximum	0.78	0.563

The difference in dispersion between AD and VD is analytically important. AD varies substantially across the 100 cases, reflecting differences in automation concentration, recommendation support, execution authority, and override conditions. VD is more tightly clustered, particularly around the neutral anchor. This narrower distribution places an important constraint on the interpretation of subsequent association analyses.

5.3. Era-Wise Patterns in AD and VD

Mean AD increased substantially across the three eras:

- **Era 1: 0.336**
- **Era 2: 0.515**
- **Era 3: 0.664**

In contrast, mean VD changed only modestly:

- **Era 1: 0.485**
- **Era 2: 0.471**
- **Era 3: 0.469**

Thus, automation concentration increased markedly across successive technological eras, whereas verification diversity remained comparatively stable within the coded corpus.

Table 4. Era-Wise Mean Automation Density and Verification Diversity

Technological Era	Mean AD	Mean VD
Era 1	0.336	0.485
Era 2	0.515	0.471
Era 3	0.664	0.469

5.4. Association Between Automation Density and Verification Diversity

The association between AD and VD was negative but modest in magnitude. The Pearson correlation was:

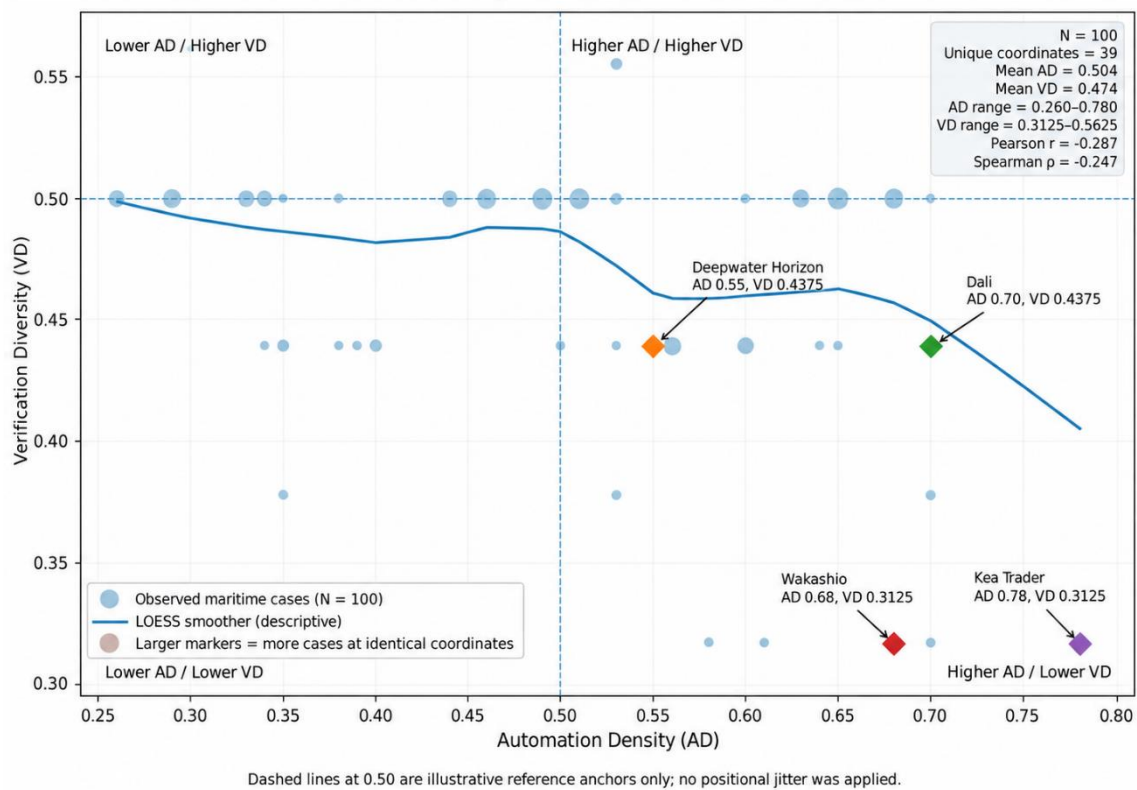
$$r = -0.287$$

and the Spearman rank correlation was:

$$\rho = -0.247$$

The similar direction of the two coefficients indicates a modest negative descriptive association between automation concentration and verification diversity in the maritime corpus. However, the magnitude of the relationship does not indicate that higher AD corresponds uniformly to lower VD.

Figure 4. Descriptive Relationship between Automation Density and Verification Diversity in the Maritime Corpus (N = 100).



The case-level distribution in Figure 4 shows substantial heterogeneity across AD–VD configurations. The LOESS curve provides a descriptive representation of the overall pattern, while the dispersion of observations demonstrates that similar levels of automation density can coexist with different levels of verification diversity.

The horizontal and vertical reference lines at 0.50 are **illustrative anchors only** and should not be interpreted as validated thresholds, safety boundaries, or causal breakpoints.

5.5. T1–T2–T3 Mechanisms in the Maritime Corpus

The maritime case analysis identified patterns consistent with each of the three T3-DEF mechanisms.

T1 Perceptual Distortion was observed in cases where sensing, information selection, presentation, or interpretation constrained the operational picture available to decision-makers.

T2 Decision Amplification was observed where recommendations, system outputs, or narrowed interpretations acquired increasing influence over subsequent operational decisions.

T3 Institutional Embedding was observed where repeated assumptions, procedures, or system-mediated practices became normalized or incorporated into organizational routines.

The mechanisms were not mutually exclusive. In some cases, perceptual constraints were followed by amplified decision reliance and subsequently reinforced through organizational practice. Conversely, existing organizational routines could also shape later perception and

decision processes. These observations are consistent with the recursive structure proposed in Section 3.

5.6. Cross-Domain Observation of T1–T2–T3 Mechanisms

The 70 cross-domain AI-risk cases were examined separately from the maritime AD–VD analysis. Their purpose was not to provide quantitative validation of AD or VD, but to examine whether patterns corresponding to the three-layer mechanism were observable across heterogeneous AI-related contexts.

Across these cases, recurring patterns included:

- incomplete or selectively represented information;
- increased reliance on automated recommendations or outputs;
- weakened independent challenge;
- diffusion of responsibility; and
- normalization of AI-mediated assumptions or practices.

These observations indicate that patterns corresponding to T1, T2, and T3 can be identified outside the maritime corpus. However, because the cross-domain cases differ substantially in source quality, operational detail, and system architecture, these findings are treated as **exploratory mechanism-level observations rather than formal cross-domain validation.**

5.7. Summary of Results

Three principal empirical patterns emerged from the analysis. First, Automation Density increased substantially across technological eras, while Verification Diversity remained comparatively stable. Second, the AD–VD association was negative but modest, indicating substantial variation in verification architecture at similar levels of automation concentration. Third, patterns corresponding to T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding were observable in both maritime and cross-domain cases.

These findings provide the empirical basis for the interpretation developed in Section 6.

6. Discussion

6.1. From Automation Risk to Configurational Vulnerability

The findings refine the central proposition of T3-DEF by shifting attention away from automation as an isolated source of risk and toward the **configuration of automation and verification within socio-technical systems**.

The modest negative association between Automation Density (AD) and Verification Diversity (VD), together with substantial case-level heterogeneity, suggests that the two dimensions should not be treated as simple opposites. Increasing automation concentration does not necessarily imply a proportional decline in verification diversity. A highly automated system may preserve diverse and independent verification mechanisms, whereas a less automated system may remain vulnerable if information pathways, challenge mechanisms, or intervention authority are weak.

This supports a conditional formulation of the framework:

Recursive socio-technical vulnerability may become more likely when increasing automation concentration is combined with insufficiently diverse, independent, or actionable verification structures.

The analytical focus therefore shifts from the amount of automation alone to the interaction between automation concentration, verification architecture, and organizational context.

6.2. Interpreting the Modest AD–VD Association

The modest magnitude of the AD–VD association is analytically important because it suggests that automation and verification structures may evolve through partially independent processes. Several factors may contribute to this pattern. First, VD values were more tightly clustered than AD values, partly because neutral anchors were retained where documentary evidence did not support directional classification. This conservative approach reduces speculative coding but may also restrict variance and attenuate observable associations.

Second, automation concentration and verification architecture may develop through different organizational processes. Automation may increase through equipment modernization, software integration, or efficiency-driven operational changes, whereas verification structures depend more heavily on staffing, training, procedural design, organizational culture, and allocation of authority.

Third, the maritime corpus contains heterogeneous failure mechanisms. Technical, interpretive, navigational, structural, power-related, and mixed socio-technical events may not exhibit identical AD–VD relationships.

Finally, AD and VD remain ordinally anchored heuristic constructs rather than direct physical measurements. The observed association should therefore be interpreted as a descriptive property of the coded corpus rather than as evidence of a universal causal relationship.

6.3. A Possible Verification Lag

The era-wise results showed that AD increased substantially across technological eras, whereas VD remained comparatively stable. This asymmetry suggests a provisional research hypothesis that may be described as **verification lag**.

Verification lag refers to:

a condition in which automation concentration or capability develops more rapidly than the diversity, independence, and operational effectiveness of the structures used to verify, challenge, and correct it.

This concept should be interpreted cautiously. The present study does not establish verification lag as a general law, and the era categories are broad contextual descriptors. Nevertheless, the pattern raises an important socio-technical question: whether the evolution of verification architecture keeps pace with the increasing concentration of automated perception, recommendation, and execution.

This issue may be particularly relevant in contemporary AI-assisted systems, where formal human oversight can remain in place even as independent information sources, alternative interpretations, challenge authority, or effective intervention mechanisms fail to evolve proportionally.

6.4. Recursive Vulnerability Across T1, T2, and T3

The findings also support the analytical value of distinguishing among T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding.

T1 focuses on how operational information is represented, filtered, or interpreted. T2 addresses how a particular representation or recommendation gains influence over subsequent decision-making. T3 concerns how repeated reliance becomes normalized or embedded in organizational routines and responsibility structures.

The principal contribution of this layered structure is not simply the identification of three categories, but the recognition that they can interact recursively. Organizational practices may influence what information is collected, how automated outputs are interpreted, and how future decisions are made. Likewise, operational outcomes may reinforce or challenge existing assumptions.

For example, when repeated reliance on an automated recommendation becomes institutionalized as standard operating practice at T3, training, interface use, and operator expectations may progressively orient attention toward the same information pathway. This institutionalized expectation can then feed back into T1 by narrowing the perceptual frame through which subsequent system outputs and alternative signals are interpreted, potentially initiating another T1–T2–T3 cycle.

This recursive mechanism distinguishes T3-DEF from explanations that treat AI-related failure as a one-directional sequence from technical error to adverse outcome. At the same time, recursion is not necessarily harmful: feedback may also produce learning, redesign, stronger verification, or improved intervention authority.

6.5. Verification Diversity and Meaningful Human Oversight

A further implication concerns the meaning of human oversight.

A binary distinction between “human-in-the-loop” and “human-out-of-the-loop” may be insufficient for evaluating socio-technical resilience. Human operators may retain formal responsibility while lacking independent information, alternative interpretations, sufficient challenge capability, or effective authority to intervene.

T3-DEF therefore conceptualizes meaningful oversight through Verification Diversity, which includes:

- information-source independence;
- interpretive diversity;
- challenge capability; and
- intervention authority.

This formulation emphasizes that human presence alone is not equivalent to independent verification. A system may retain nominal human supervision while effectively depending on the same information pathway, assumptions, or automated recommendations. Conversely, a highly automated system may preserve strong oversight if independent verification and intervention mechanisms remain operationally effective.

6.6. Implications for AI Risk Management and Governance

T3-DEF is intended to complement rather than replace established AI risk-management and governance frameworks.

Frameworks such as the NIST AI RMF, the EU AI Act, and Korea's AI Basic Act address essential requirements concerning risk management, transparency, accountability, monitoring, and human oversight [20-22]. International standards such as ISO/IEC 23894 and ISO/IEC 42001 similarly provide structured guidance for AI risk management and organizational management systems [32,33].

T3-DEF addresses a related but distinct analytical question:

How can vulnerability propagate through an operational socio-technical system even when formal governance structures are present?

This distinction can be summarized as follows:

Governance asks what controls should exist; T3-DEF asks whether those controls remain operationally independent and effective in practice.

For example, a governance framework may require human oversight, while T3-DEF asks whether that oversight possesses sufficient verification diversity. Similarly, risk-management procedures may document monitoring controls while operational practice remains dependent on shared information sources or common automated assumptions.

T3-DEF may therefore function as a **complementary diagnostic layer** linking formal governance requirements with operational socio-technical conditions, without becoming a compliance or certification framework.

6.7. Cross-Domain and Theoretical Implications

The 70 cross-domain AI-risk cases suggest that T1–T2–T3 mechanisms may have relevance beyond maritime operations, although they do not constitute formal cross-domain validation. Patterns consistent with perceptual distortion, decision amplification, weakened independent challenge, responsibility diffusion, and institutional embedding were observable across heterogeneous AI-related contexts. The significance of this finding lies less in claiming universality than in identifying a common analytical vocabulary that may support comparison across otherwise different socio-technical domains.

The broader theoretical contribution of T3-DEF lies in linking three levels of analysis that are often studied separately:

- information representation and perception;
- decision influence and reliance; and
- institutionalization and organizational practice.

AD and VD complement these mechanisms by characterizing the surrounding automation and verification architecture.

The framework is deliberately structured so that future studies can apply T1–T2–T3 mechanism coding and AD/VD structural analysis independently, enabling comparison across domains without requiring adoption of a fixed risk score.

T3-DEF therefore contributes not by proposing a universal numerical risk score, but by specifying a set of structured and falsifiable relationships through which recursive socio-technical vulnerability can be examined across AI-assisted environments

6.8. Limitations and Future Research

Several limitations define the appropriate scope of the present findings.

First, the 100 maritime cases constitute a purposively assembled analytical corpus rather than a statistically representative population sample. Second, AD and VD remain heuristic constructs whose construct validity requires independent external testing. Third, VD is particularly sensitive to documentary limitations because accident investigation reports often provide more detailed information about immediate technical causes than about information-source independence, organizational challenge, or intervention authority.

Fourth, neutral-anchor coding may have restricted VD variance.

Future validation should also examine the robustness of VD estimates through independent expert-panel or Delphi-based coding and sensitivity analyses comparing the current neutral-anchor treatment with alternative approaches for handling evidence-limited cases.

Fifth, the full maritime corpus has not yet undergone prospective independent multi-coder validation. Computational consistency therefore should not be interpreted as equivalent to inter-rater or construct validity. Sixth, technological-era categories are contextual descriptors and not causal predictors. Finally, the present analysis cannot establish that T1–T2–T3 propagation or specific AD–VD configurations cause adverse outcomes.

Future research should therefore prioritize independent multi-coder validation of AD and VD, construct-validation studies, subgroup analysis by failure mechanism, longitudinal examination of the verification-lag hypothesis, and cross-domain replication using evidence-rich cases. Further work should also test interaction effects between AD and VD and examine how T3-DEF maps onto major AI governance and risk-management frameworks such as NIST AI RMF, the EU AI Act, Korea's AI Basic Act, ISO/IEC 23894, and ISO/IEC 42001 [20-22,32,33].

Most importantly, future empirical work should remain open to confirming, modifying, narrowing, or rejecting individual propositions of T3-DEF. Such falsifiability is central to the intended development of the framework.

7. Conclusion

This study proposed **T3-DEF (Three-Layer Distortion and Embedding of Failure)** as an exploratory, mechanism-oriented framework for examining recursive socio-technical vulnerability in AI-assisted systems. The framework links **T1 Perceptual Distortion, T2 Decision Amplification, and T3 Institutional Embedding** with two complementary structural dimensions: **Automation Density (AD)** and **Verification Diversity (VD)**.

The maritime analysis provided initial operational grounding for these constructs. Automation Density increased substantially across technological eras, whereas Verification Diversity remained comparatively stable, and the association between the two was modest rather than deterministic. These findings support a configuration-based interpretation in which automation concentration and verification architecture should be treated as partially independent dimensions.

Accordingly, the central proposition of T3-DEF is conditional: **recursive socio-technical vulnerability may become more likely when increasing automation concentration is combined with insufficiently diverse, independent, or actionable verification structures**. High automation is therefore not treated as inherently unsafe, and low verification diversity is not assumed to be an inevitable consequence of automation.

The cross-domain cases further suggest that the T1–T2–T3 mechanism may be applicable beyond maritime operations, although the present study should be understood as an initial framework-formulation and operational-grounding study rather than as a completed validation. AD and VD remain heuristic constructs, the maritime corpus is purposively assembled, and independent multi-coder and cross-domain validation remain necessary.

T3-DEF asks not only whether AI risk is governed, but how vulnerability forms, propagates, and becomes embedded.

T3-DEF is therefore best viewed as a **transparent, falsifiable, and extensible analytical framework** that complements existing AI risk-management and governance approaches by focusing on how vulnerability can emerge, amplify, and become institutionally embedded in socio-technical systems. Future research should test, refine, narrow, or reject its propositions through independent coding, construct validation, longitudinal analysis, and cross-domain replication.

Disclosures

Conflicts of Interest:

The author declares no conflicts of interest related to this study.

Funding:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical Approval:

This study is based on publicly available accident investigation reports, published AI-risk cases, and secondary documentary sources. No human participants, identifiable private data, or experimental interventions were involved; therefore, institutional review board approval was not required.

Data Availability:

The case-level analytical datasets used in this study are provided as supplementary materials. Supplementary Dataset S1 contains the cross-domain T3-DEF operational coding cases, and Supplementary Dataset S2 contains the maritime accident T3-DEF coding, Automation Density (AD), Verification Diversity (VD), detailed AD and VD coding rubrics, summary statistics, and case-level verification information. Source traceability and coding rationales are documented where available.

Use of Generative AI and AI-Assisted Tools:

Generative AI and AI-assisted tools were used only to support language refinement, structural editing, formatting, and computational assistance during manuscript preparation. The conceptual framework, research design, case selection, coding criteria, interpretation of results, and final scholarly judgments were determined and verified by the author. No AI-generated content was accepted without author review and revision.

Author Responsibility:

The author takes full responsibility for the integrity of the study, the accuracy of the reported analyses, the interpretation of the findings, and the final content of the manuscript.

References

1. Parasuraman R, Riley V: Humans and automation: use, misuse, disuse, abuse. *Hum Factors*. 1997, 39:230-253. 10.1518/001872097778543886
2. Floridi L, Cowls J, Beltrametti M, et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach*. 2018, 28:689-707. 10.1007/s11023-018-9482-5
3. Rasmussen J: Risk management in a dynamic society: a modelling problem. *Saf Sci*. 1997, 27:183-213. 10.1016/S0925-7535(97)00052-0
4. Reason J: *Managing the Risks of Organizational Accidents*. Ashgate Publishing Ltd, Aldershot; 1997.
5. Hollnagel E, Pariès J, Woods DD, Wreathall J: *Resilience Engineering in Practice: A Guidebook*. Ashgate Publishing Ltd, Farnham; 2011.
6. Vincent C, Burnett S, Carthey J: *The Measurement and Monitoring of Safety*. The Health Foundation, London; 2013.
7. Woods DD: Four concepts for resilience and the implications for the future of resilience engineering. *Reliab Eng Syst Saf*. 2015, 141:5-9. 10.1016/j.res.2015.03.018
8. Leveson NG: *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge; 2011.
9. Bainbridge L: Ironies of automation. *Automatica*. 1983, 19:775-779. 10.1016/0005-1098(83)90046-8
10. Parasuraman R, Sheridan TB, Wickens CD: A model for types and levels of human interaction with automation. *IEEE Trans Syst Man Cybern A Syst Hum*. 2000, 30:286-297. 10.1109/3468.844354
11. Cummings ML: Automation bias in intelligent time critical decision support systems. *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics, Reston; 2004. 10.2514/6.2004-6313
12. Sarter NB, Woods DD, Billings CE: Automation surprises. *Handbook of Human Factors and Ergonomics*. Salvendy G (ed): Wiley, New York; 1997. 1926-1943.
13. Klein G, Woods DD, Bradshaw JM, Hoffman RR, Feltovich PJ: Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intell Syst*. 2004, 19:91-95. 10.1109/MIS.2004.74
14. Seeber I, Bittner E, Briggs RO, et al.: Machines as teammates: A research agenda on AI in team collaboration. *Inf Manag*. 2020, 57:103174. 10.1016/j.im.2019.103174
15. Amershi S, Weld D, Vorvoreanu M, et al.: Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019:1-13. 10.1145/3290605.3300233
16. Sheridan TB: *Humans and Automation: System Design and Research Issues*. John Wiley & Sons, New York; 2002.
17. Endsley MR: From here to autonomy: lessons learned from human-automation research. *Hum Factors*. 2017, 59:5-27. 10.1177/0018720816681350
18. Perrow C: *Normal Accidents: Living with High-Risk Technologies*. Princeton University Press, Princeton; 1984.

19. Avizienis A, Laprie JC, Randell B, Landwehr C: Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans Dependable Secure Comput.* 2004, 1:11-33. 10.1109/TDSC.2004.2
20. National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce, Gaithersburg; 2023. 10.6028/NIST.AI.100-1
21. European Parliament, Council of the European Union: Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*; 2024.
22. Republic of Korea: Act on the Development of Artificial Intelligence and Establishment of Trust-Based Foundations (AI Basic Act). Act No. 21311; 2026. Effective July 21, 2026.
23. Hillmann J, Guenther E: Organizational resilience: A valuable construct for management research? *Int J Manag Rev.* 2021, 23:7-44. 10.1111/ijmr.12239
24. Amodei D, Olah C, Steinhardt J, et al.: Concrete problems in AI safety. *arXiv.* 2016. 10.48550/arXiv.1606.06565
25. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L: The ethics of algorithms: mapping the debate. *Big Data Soc.* 2016, 3:1-21. 10.1177/2053951716679679
26. Yin RK: *Case Study Research and Applications: Design and Methods.* 6th ed. SAGE Publications, Thousand Oaks; 2018.
27. Eisenhardt KM: Building theories from case study research. *Acad Manage Rev.* 1989, 14:532-550. 10.5465/amr.1989.4308385
28. National Intelligence Service, National AI Security Center: *AI Risk Casebook.* National Intelligence Service, Seoul; 2025.
29. Hsieh HF, Shannon SE: Three approaches to qualitative content analysis. *Qual Health Res.* 2005, 15:1277-1288. 10.1177/1049732305276687
30. Krippendorff K: *Content Analysis: An Introduction to Its Methodology.* 4th ed. SAGE Publications, Thousand Oaks; 2018.
31. Stevens SS: On the theory of scales of measurement. *Science.* 1946, 103:677-680. 10.1126/science.103.2684.677
32. International Organization for Standardization, International Electrotechnical Commission: *ISO/IEC 23894:2023. Information technology—Artificial intelligence—Guidance on risk management.* ISO, Geneva; 2023.
33. International Organization for Standardization, International Electrotechnical Commission: *ISO/IEC 42001:2023. Information technology—Artificial intelligence—Management system.* ISO, Geneva; 2023.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SuppS1.xlsx](#)
- [SupplementaryDatasetS2ADVDCompletedwithADRubric.xlsx](#)