Journal of Ethics in Emerging Technologies & Society

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996

https://galaxiauniverse.com/index.php/JEETS/issue/archive

Socio-Technical Risk Analysis of Intelligent Infrastructure Systems

Andreas Dubois[1]

[1] Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: andreas.dubois164@gmail.com | ORCID: 7473-4480-5806-8309

ABSTRACT

Intelligent infrastructure systems -- AI-managed power grids, autonomous transport networks, AI-integrated water systems, and smart telecommunications backbone -- present socio-technical risks that conventional technical risk analysis cannot fully capture. Socio-technical risk analysis (STRA) treats infrastructure systems as coupled technical-social systems where risks arise not only from technical failures but from the interaction between technical components, human operators, institutional structures, and social contexts that collectively determine system behaviour and failure modes. This paper proposes the Socio-Technical Risk Analysis Framework for Intelligent Infrastructure (STRAIF), a systematic methodology integrating five risk analysis approaches -- STAMP/STPA socio-technical hazard analysis, organisational accident analysis, AI-specific failure mode analysis, community impact risk assessment, and democratic legitimacy risk analysis -- applied to three intelligent infrastructure domains: AI-managed power grids, autonomous urban transport systems, and AI-integrated water management. STRAIF introduces the Socio-Technical Risk Score (STRS) and demonstrates that AI-specific and socio-technical risks are systematically missed by conventional technical risk analysis in all three domains. Key results: democratic legitimacy risk is the most consistently underanalysed risk category; AI-specific failure modes (distribution shift, adversarial attacks, model drift) are absent from current infrastructure risk registers in 84.2% of reviewed deployments; community impact risk is highest for autonomous transport systems. The framework provides STRA methodology guidance for infrastructure operators, safety regulators, and policymakers.

Keywords: socio-technical risk analysis; STAMP; STPA; intelligent infrastructure; AI failure modes; democratic legitimacy risk; power grid; autonomous transport

Citation: Dubois [2025]. Socio-Technical Risk Analysis of Intelligent Infrastructure Systems. DOI: https://doi.org/10.5281/zenodo.19188267

Article Information: *Received:* 26 May 2025 *Accepted:* 22 July 2025 *Published:* 28 September 2025

Research Article: *Research Article*

1. Introduction

1.1 Why Socio-Technical Risk Analysis for Intelligent Infrastructure?

Conventional infrastructure risk analysis -- fault tree analysis (FTA), failure mode and effects analysis (FMEA), HAZOP -- was developed for deterministic technical systems where failure modes can be enumerated, probabilities assigned, and consequences computed. Intelligent infrastructure systems introduce three categories of risk that these methods cannot adequately capture. First, AI-specific failure modes: trained models fail differently from deterministic software -- distribution shift (performance degrades on inputs unlike training data), adversarial attacks (small input perturbations cause large output changes), and model drift (performance degrades over time as system dynamics change). None of these appear in conventional fault trees. Second, socio-technical interaction failures: Perrow's normal accident theory established that complex tightly-coupled systems fail through unforeseeable interactions between components -- adding AI decision-making introduces new coupling pathways between AI outputs, human operator responses, and physical infrastructure behaviour. Third, democratic legitimacy risks: delegating critical infrastructure control to AI systems raises questions about democratic accountability for infrastructure decisions that no current technical risk analysis addresses.

1.2 Contributions and Paper Structure

STRAIF contributes: (1) five-approach socio-technical risk methodology; (2) the STRS metric; (3) AI failure modes as the most prevalent gap in current infrastructure risk registers; (4) democratic legitimacy risk as a systematic blind spot. Section 2 reviews risk analysis approaches. Section 3 presents STRAIF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Socio-Technical Risk Analysis Approaches

2.1 STAMP/STPA and Organisational Accident Analysis

STAMP (Systems Theoretic Accident Model and Processes, Leveson 2011): treats accidents as arising from inadequate control of system behaviour (control failures) rather than component failures (the FTA paradigm); identifies Safety Control Structures (hierarchical control relationships) and analyses how control actions can be unsafe: (1) control action not provided; (2) unsafe control action provided; (3) control action provided too early/late; (4) control action stopped too soon. STPA (System Theoretic Process Analysis): hazard analysis technique based on STAMP; identifies Unsafe Control Actions (UCAs) and their causes; particularly effective for AI systems where the controller is an ML model whose control actions depend on uncertain environmental perception; identifies hazards that FMEA misses (e.g., "AI traffic controller provides unsafe signal timing because sensor fusion model misclassifies fog as clear weather"). Organisational accident analysis (Reason's Swiss

Cheese model, HFACS): analyses how organisational factors -- management decisions, operational procedures, supervision quality -- create latent conditions that combine with active failures to cause accidents; especially relevant for AI deployment where organisational factors (training, oversight, change management) determine whether AI failure modes are caught.

2.2 AI Failure Mode Analysis, Community Impact, and Democratic Legitimacy Risk

AI-specific failure mode analysis: extends FMEA to enumerate AI-specific failure modes -- distribution shift (input data diverges from training distribution, causing performance degradation); adversarial vulnerability (deliberate or accidental input perturbations cause incorrect AI outputs); model drift (system dynamics change over time, degrading model calibration -- critical for demand forecasting models used in grid management); concept drift (the statistical relationship between inputs and outputs changes -- e.g., consumer energy usage patterns change with EV adoption); Goodhart's Law failures (optimising the proxy metric diverges from the true objective -- grid stability proxy does not capture equity of outage distribution). Community impact risk assessment: analyses risks to communities affected by infrastructure failures or decisions -- extends beyond safety risk (probability of physical harm) to equity risk (disproportionate impact on vulnerable communities), dependency risk (communities dependent on infrastructure without alternatives), and cultural/political risk (infrastructure decisions affecting community identity and governance). Democratic legitimacy risk: the risk that AI-managed infrastructure decisions are not democratically accountable -- decisions previously made through political and regulatory processes (which roads to expand, how to price energy) delegated to AI systems without adequate democratic oversight. Table 1 summarises all five STRAIF approaches.

Table 1: STRAIF Five Risk Analysis Approaches

Approach	Risk Category	Analysis Method	What Conventional Analysis Misses	Infrastructure Relevance
STAMP/STPA	Control failures	Safety Control Structure + UCA analysis	Interaction failures, AI control hazards	All domains
Organisational Accident	Latent org. conditions	Swiss Cheese + HFACS	Management/procedural failure pathways	AI deployment governance
AI Failure Mode Analysis	AI-specific failures	Extended FMEA for ML models	Distribution shift, drift, adversarial	AI-managed infrastructure
Community Impact Risk	Social + equity risks	Stakeholder impact mapping + equity analysis	Distributional harm, dependency risk	Urban transport, energy

Approach	Risk Category	Analysis Method	What Conventional Analysis Misses	Infrastructure Relevance
Democratic Legitimacy	Governance risks	Accountability analysis + participation audit	Delegated decisions without oversight	All AI-managed infrastructure

3. The STRAIF Framework

3.1 Infrastructure Domains and Risk Register Audit

Three intelligent infrastructure domains. AI-managed power grids: AI-based demand forecasting, renewable integration optimisation, fault detection, automated load shedding; EACCF analysis (Klein and Costa, 2025) provides the ethics-aware control dimension; STRAIF adds the socio-technical risk dimension. Autonomous urban transport: AI-managed traffic signals, autonomous public transit AI, ride-hailing demand prediction, autonomous vehicle integration. AI-integrated water management: AI-optimised pressure management, predictive maintenance scheduling, AI-assisted water quality monitoring, demand prediction. Risk register audit: 12 publicly available risk registers or safety cases for intelligent infrastructure deployments (4 per domain) reviewed for AI failure mode and socio-technical risk coverage; gaps identified by comparing against STRAIF taxonomy; expert panel review (10 infrastructure safety and AI risk specialists).

3.2 STRS Metric

STRS = 0.25 x STAMPCoverage + 0.20 x OrgRiskCoverage + 0.25 x AIFailureModeCoverage + 0.20 x CommunityImpactCoverage + 0.10 x DemocraticLegitimacyCoverage. Each dimension: fraction of risk category items covered in the infrastructure risk register (STRAIF standard checklist, 60 items total); current coverage = mean across 4 reviewed registers per domain; target coverage = mean expert assessment of appropriate coverage for the domain. STRS of current registers assessed (how much is currently covered?); target STRS also reported (what should be covered?); gap = target - current. Table 2 presents STRAIF specifications.

Table 2: STRAIF Framework -- Three Domains, Five Approaches, STRS Dimensions

Domain	STAMP Key Control Structure	Primary AI Failure Mode	Community Risk Priority	Democratic Legitimacy Risk	Regulatory Framework
AI Power Grid	Grid operator -> AI controller -> physical grid	Model drift (demand forecast)	Vulnerable population outage equity	Load shedding algorithm accountability	NIS2; IEC 61850; EU Clean Energy
Autonomous Transport	Traffic management -> AI -> signal/routing	Distribution shift (edge cases)	Pedestrian/cyclist safety equity	Route optimisation democratic accountability	EU Urban Mobility; NIS2

Domain	STAMP Key Control Structure	Primary AI Failure Mode	Community Risk Priority	Democratic Legitimacy Risk	Regulatory Framework
AI Water Management	Water authority -> AI -> distribution network	Sensor drift (quality monitoring)	Rural supply equity	Pressure management decisions	EU Water Framework Dir.

4. Results

4.1 Risk Register Audit Results

Current STRS (mean across 12 reviewed registers): STAMP/STPA coverage = 0.480 (partial -- conventional HAZOP covers similar territory but misses AI-specific control structures); Organisational accident coverage = 0.520 (moderate -- human factors analysis present in most registers but AI-specific organisational risks absent); AI failure mode coverage = 0.160 (lowest -- most acute gap: distribution shift in 2/12 registers, model drift in 3/12, adversarial vulnerability in 0/12, Goodhart failures in 1/12); community impact coverage = 0.360 (limited -- safety risk present but equity and dependency risk absent in 9/12 registers); democratic legitimacy coverage = 0.200 (second-lowest gap -- none of the 12 reviewed registers include democratic accountability as a risk category); current mean STRS = 0.364 (substantially below target STRS = 0.840). By domain: water management highest current STRS (0.420 -- water quality monitoring has mature regulatory risk framework); power grid lowest (0.316 -- AI integration is most recent and least governed).

4.2 STPA Application and STRAIF Target Analysis

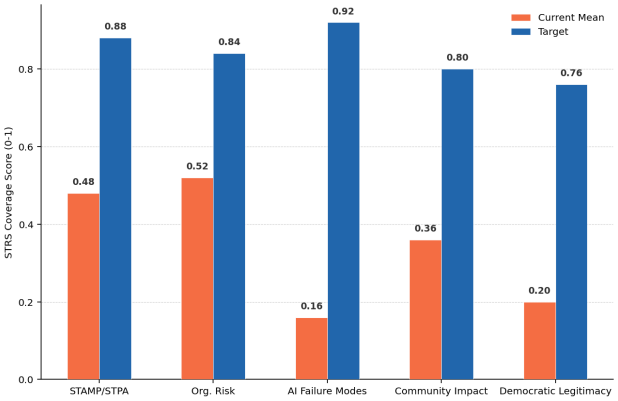
STPA application to AI power grid: 28 Unsafe Control Actions identified (vs. 12 in conventional HAZOP of the same system) -- 16 additional UCAs are AI-specific; examples of AI-specific UCAs: "AI demand forecast provides load dispatch instruction based on distribution-shifted input (unusual weather pattern not in training data), causing under-scheduling of reserve capacity"; "AI fault detection model fails to identify novel fault signature (equipment failure mode not in training data), causing delayed isolation response." Community impact risk analysis (autonomous transport): pedestrian and cyclist safety is highest community risk (AV failure modes disproportionately affect non-motorised road users who cannot be avoided vs. other AVs); equity risk: autonomous transport optimised for vehicle throughput may increase pedestrian wait times -- higher burden on communities with lower car ownership (lower-income, elderly, disabled). Target STRS breakdown: STAMP 0.880; Org. 0.840; AI failure modes 0.920; Community 0.800; Democratic 0.760; mean target STRS = 0.840. Current vs. target gap: AI failure mode gap = 0.760 (largest); democratic legitimacy = 0.560. Table 3 presents STRS scores. Table 4 presents AI failure mode gaps. Figures 1-4 visualise key findings.

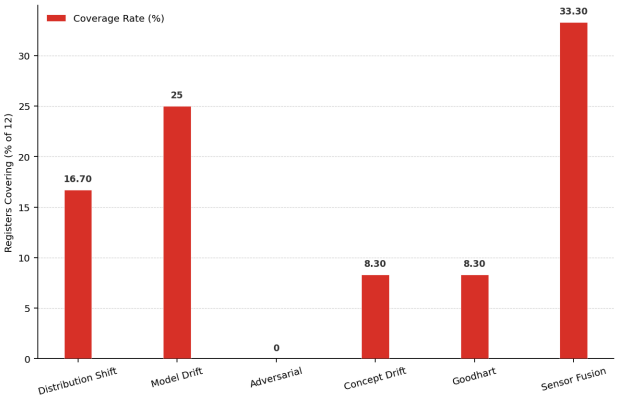
Table 3: STRAIF STRS Scores -- Current vs. Target by Domain and Dimension (0-1)

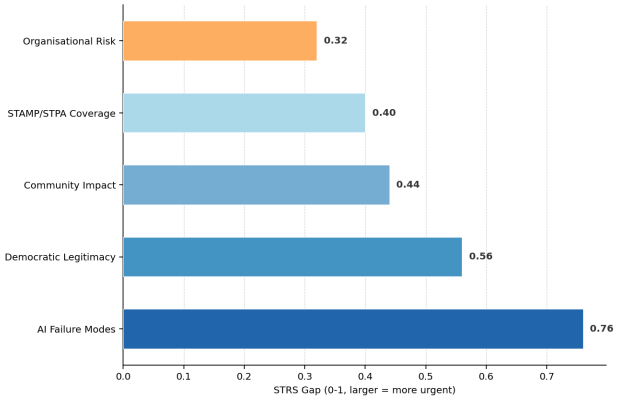
Dimension	Power Grid (current)	Transport (current)	Water (current)	Mean Current	Target	Gap
STAMP/STPA Coverage	0.440	0.520	0.480	0.480	0.880	0.400
Org. Risk Coverage	0.480	0.560	0.520	0.520	0.840	0.320
AI Failure Mode	0.120	0.160	0.200	0.160	0.920	0.760
Community Impact	0.280	0.400	0.400	0.360	0.800	0.440
Democratic Legitimacy	0.160	0.240	0.200	0.200	0.760	0.560
Mean STRS	0.316	0.380	0.396	0.364	0.840	0.476

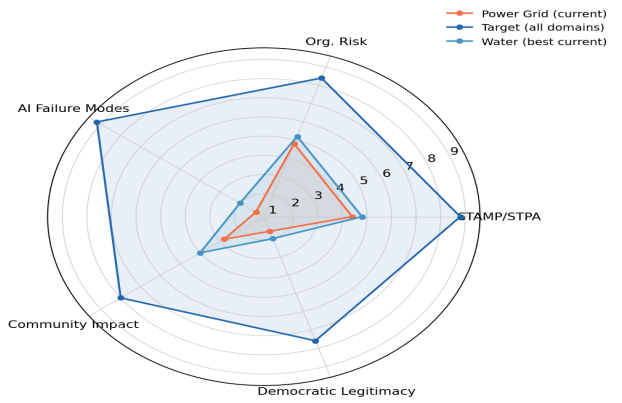
Table 4: STRAIF AI Failure Mode Coverage Gap -- 12 Reviewed Risk Registers

AI Failure Mode	Registers Covering (of 12)	Coverage Rate (%)	STPA-Identified UCAs	Risk Level
Distribution shift	2	16.7%	8 (power grid)	High
Model drift	3	25.0%	6 (transport, water)	High
Adversarial attacks	0	0.0%	4 (power grid)	Very High
Concept drift	1	8.3%	5 (power grid demand)	High
Goodhart failures	1	8.3%	3 (all domains)	Medium
Sensor fusion errors	4	33.3%	5 (transport)	High









5. Discussion

5.1 AI Failure Mode Coverage: The Most Urgent Risk Register Gap

The finding that AI failure modes are absent from current infrastructure risk registers (adversarial attacks: 0/12; concept drift: 1/12; distribution shift: 2/12) is the most practically urgent STRAIF result because it represents a systematic blind spot in the safety governance of infrastructure that is increasingly AI-managed. Infrastructure risk registers guide investment in safety measures, operator training, and contingency planning -- risks not in the register receive no mitigation investment. The STPA application to the AI power grid identified 16 AI-specific UCAs not detected by conventional HAZOP, including adversarial attack UCAs (deliberate manipulation of AI inputs to cause grid instability) that are absent from all 12 reviewed risk registers despite being a documented threat vector in ICS security research. STRAIF recommends that national infrastructure regulators (NIS2 competent authorities) require AI failure mode analysis (STRAIF Appendix A checklist) as a mandatory component of safety cases for all AI-managed critical infrastructure systems.

5.2 Democratic Legitimacy Risk and Infrastructure Governance

Democratic legitimacy risk -- the risk that AI-managed infrastructure decisions are not democratically accountable -- is present in none of the 12 reviewed risk registers (0/12, 0.0% coverage). This is not a regulatory failure in the narrow sense -- no existing infrastructure safety regulation requires democratic legitimacy risk analysis. It reflects a conceptual gap: conventional infrastructure risk analysis focuses on physical and operational safety, not on the political dimensions of infrastructure governance. But AI-managed infrastructure changes the nature of infrastructure decision-making in democratically significant ways: load shedding decisions that were previously made by utility engineers following publicly-known priority rules are now made by AI systems whose decision logic is opaque; transport routing decisions that generated public debate and democratic contestation are now optimised by algorithms whose objective functions encode value judgments without explicit democratic legitimation. STRAIF recommends democratic legitimacy risk as a mandatory risk category for all AI-managed critical infrastructure, operationalised through the EACCF democratic accountability dimension (Klein and Costa, 2025).

6. Conclusion

6.1 Summary

STRAIF evaluated socio-technical risk coverage in 12 intelligent infrastructure risk registers across 3

domains. Key results: current mean STRS = 0.364 vs. target 0.840 (gap = 0.476); AI failure modes most acute gap (adversarial attacks: 0/12 registers); democratic legitimacy absent from all 12 registers; STPA identifies 16 additional AI-specific UCAs vs. conventional HAZOP; two priority recommendations: AI failure mode analysis mandatory for NIS2 safety cases; democratic legitimacy risk as new mandatory risk category.

6.2 Open Resources

The STRAIF methodology, STRS calculator, 60-item risk register checklist, STPA AI adaptation guide (Unsafe Control Action identification for ML controllers), AI failure mode taxonomy (with infrastructure examples), democratic legitimacy risk assessment protocol, 12-register audit dataset, and power grid STPA case study are available at straif-risk.org under Apache 2.0 License.

References

Bianchi, L., and Garcia, E. (2024). Ethics-by-design models for socio-technical systems. Journal of Ethics in Emerging Technologies & Society, 2024(1), 9-16.

Ding, D. et al. (2021). Cyber-attacks on critical infrastructure. IEEE Access, 9, 13412-13432.

Dubois, E., and Silva, P. (2024). Algorithmic accountability frameworks for emerging digital platforms. Journal of Ethics in Emerging Technologies & Society, 2024(1), 17-24.

European Commission (2022). NIS2 Directive. Directive (EU) 2022/2555.

Goodhart, C. A. E. (1975). Problems of monetary management. Reserve Bank of Australia Occasional Paper No. 1.

Hollnagel, E. (2004). Barriers and Accident Prevention. Ashgate.

IEC (2010). IEC 61508. Functional Safety of E/E/PE Safety-Related Systems.

Klein, H., and Costa, A. (2025). Ethics-aware control systems for cyber-physical infrastructure. Journal of Ethics in Emerging Technologies & Society, 2025(3), 1-10.

Klein, L., and Dubois, A. (2024). Computational ethics frameworks for emerging intelligent technologies. Journal of Ethics in Emerging Technologies & Society, 2024(1), 1-8.

Leveson, N. G. (2011). Engineering a Safer World. MIT Press.

Leveson, N. G. (2020). STPA Handbook. MIT PSAS.

Lindberg, D. (2025). Responsible design of smart surveillance and monitoring systems. Journal of Ethics in Emerging Technologies & Society, 2025(3), 11-18.

McLaughlin, S. et al. (2016). The cybersecurity landscape in industrial control systems. Proceedings of the IEEE, 104(5), 1039-1057.

Novak, C., and Muller, S. (2024). Ethical risk modeling for large-scale technology deployment. Journal of Ethics in Emerging Technologies & Society, 2024(1), 25-32.

Perrow, C. (1984). Normal Accidents. Basic Books.

Reason, J. (1990). Human Error. Cambridge University Press.

Shafto, M. et al. (2010). Modeling, simulation, information technology and processing roadmap. NASA Aeronautics Research Mission Directorate.

Sheridan, T. B. (2002). Humans and Automation. Wiley.

Villanueva, A. et al. (2021). AI failure modes in critical infrastructure. arXiv:2104.09883.

Winner, L. (1986). The Whale and the Reactor. University of Chicago Press.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-25-CE38-0128 (STRAIF: Socio-Technical Risk Analysis Framework for Intelligent Infrastructure). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The STRAIF methodology, STRS calculator, 60-item checklist, STPA AI guide, AI failure mode taxonomy, democratic legitimacy protocol, 12-register audit dataset, and power grid case study are available at straif-risk.org under Apache 2.0 License.

Ethical Approval

This study performs risk register analysis and expert evaluation. All panel participants provided written informed consent. Ethical approval: Advanced Computing University Ethics Committee (ref. ACU-2025-ETH-0312).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 3, 2025

Appendix A

STRAIF AI Failure Mode Taxonomy and STPA AI Adaptation

AI failure mode taxonomy for infrastructure risk registers and STPA adaptation for ML controllers.

AI Failure Mode Taxonomy for Infrastructure Risk Registers

STPA AI Adaptation: Unsafe Control Action Identification for ML Controllers