International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Artificial Intelligence Governance and Cybersecurity Trust: A Comparative Study of Public Infrastructure Systems in India, Singapore, and the United Kingdom

A K M Fazlur Rahman[1], Dr Dulari A Rajput[2]

[1]CEO
[2]Dissertation Director, Doctorate, IIBM

Abstract

The increasing integration of Artificial Intelligence (AI) into critical public infrastructure systems (e.g., energy, transport, water, digital identity) introduces both operational efficiencies and unprecedented cybersecurity vulnerabilities. However, the relationship between national AI governance frameworks and the cultivation of cybersecurity trust remains underexplored, particularly across divergent socio-technical and regulatory contexts. This paper addresses the central research question: How do different AI governance models in India, Singapore, and the United Kingdom influence cybersecurity trust outcomes in public infrastructure systems?

A comparative multiple-case study design was employed, selecting one major AI-enabled public infrastructure system per country: India's DigiYatra (biometric air travel), Singapore's Smart Water Assessment Network (SWAN), and the UK's National Grid AI Demand Forecasting System. Data were collected from policy document analysis (2019–2024), semi-structured interviews with 45 cybersecurity and infrastructure governance experts, and publicly available incident reports. A thematic analysis was guided by a conceptual framework integrating institutional trust theory, the NIST AI Risk Management Framework, and GDPR/UK GDPR data protection principles. Cross-case comparison used a most-different systems design to isolate governance effects.

Results reveal three distinct governance-trust configurations. India's hybrid governance (non-binding guidelines plus sectoral mandates) fosters rapid AI deployment but produces fragmented trust, with high citizen usage alongside low institutional confidence in breach response. Singapore's centralized, risk-based model (Model AI Governance Framework, amended Cybersecurity Act) generates managed trust—predictable but brittle, with private infrastructure partners exhibiting compliance fatigue. The UK's principles-based, cross-sectoral approach (e.g., CDEI, NCSC guidance) yields negotiated trust, characterized by active public contestation and slower adoption but higher resilience to adversarial attacks. Cross-cutting findings show that technical robustness alone does not predict trust; instead, transparency mechanisms (e.g., algorithmic impact assessments) and redress pathways are stronger determinants. Notably, all three systems struggle with trust asymmetries: AI operators over-trust automated defenses while citizens under-trust anomaly detection systems.

We conclude that no single governance model universally optimizes cybersecurity trust. India's agility suits resource-constrained scaling but requires independent oversight for trust repair. Singapore's

precision reduces known risks but may fail against novel AI attacks. The UK's deliberative model builds legitimacy but at the cost of speed. For policymakers, we recommend (1) embedding 'trust audits' as mandatory components of AI system certifications, (2) establishing cross-jurisdictional learning mechanisms for incident response, and (3) moving from static compliance to dynamic trust calibration. Future work should extend the comparison to Global South contexts and empirically test the proposed trust-governance typology.

Introduction

1. Background, Historical Data, Definitions, and Key Terms

For most of human history, trust in public infrastructure—whether a power grid, a water treatment plant, or a railway signalling system—rested on predictable, mechanical, and human-supervised operations. An engineer turning a valve, a switchboard operator rerouting power, or a signalman pulling levers: these were visible, understandable, and accountable actions. The digital revolution of the 1990s and 2000s began to erode that visibility, but the core logic remained human-authored code. The past decade, however, has witnessed a quieter but more profound shift: the insertion of autonomous Artificial Intelligence (AI) systems into the nervous system of critical public infrastructure.

Consider what this means in practical terms. An AI now predicts electricity demand across a national grid and automatically adjusts supply. Another AI flags potential leaks in a smart water network, bypassing human review for emergency shut-offs. A biometric AI system clears travellers through airport security without a human officer verifying each match. These are not futuristic scenarios; they are operational realities in countries like Singapore, the United Kingdom, and increasingly India.

However, this transition has introduced a novel problem: cybersecurity trust—a term that requires careful unpacking. Unlike traditional trust in infrastructure (which meant "it works reliably and safely"), cybersecurity trust in the age of AI carries three distinct dimensions. First, technical trust: does the AI system correctly resist, detect, and recover from cyberattacks? Second, institutional trust: do the governing bodies and infrastructure operators have the competence, transparency, and accountability to manage AI-specific risks? Third, societal trust: do citizens, businesses, and frontline workers believe the system is secure enough to depend upon, even when they cannot understand its internal workings?

The historical backdrop here is critical. From the late 1990s to the early 2010s, cybersecurity of public infrastructure focused primarily on perimeter defence (firewalls, air gaps) and human-in-the-loop verification. The 2015 Ukraine power grid cyberattack was a watershed moment, demonstrating that determined adversaries could bypass conventional defences. The subsequent wave of AI adoption in infrastructure—accelerated globally between 2017 and 2022—was partly a defensive response: AI promised faster anomaly detection, predictive maintenance, and automated containment of threats. But as we have learned painfully since, AI itself became a new attack surface. Adversarial machine learning, data poisoning, and model inversion attacks are no longer theoretical; they have been demonstrated in laboratory and, in a few unconfirmed cases, in operational settings.

Key terms used throughout this paper are defined as follows: AI Governance refers to the ensemble of laws, policies, standards, and institutional practices that shape the development, deployment, and oversight of AI systems. Cybersecurity Trust is operationalised as the justified confidence of stakeholders that an AI-enabled infrastructure system will maintain confidentiality, integrity, and availability of data and operations despite malicious interference. Public Infrastructure Systems encompass assets and services deemed critical by national governments, including energy, water, transport, and digital identity platforms.

2. Existing Evidence – Literature Survey

The scholarly landscape on AI governance and cybersecurity has grown rapidly but remains curiously siloed. One robust stream of literature, primarily from computer science and engineering, focuses on technical robustness. Researchers such as Papernot et al. (2018) and Carlini and Wagner (2020) have meticulously documented adversarial attack vectors against machine learning models. Complementary work in cybersecurity journals (e.g., Shneiderman, 2020; Sarker et al., 2021) has proposed defensive architectures, including adversarial training, input sanitisation, and differential privacy. This literature is rigorous but overwhelmingly technical, treating trust as a binary outcome (secure vs. compromised) rather than a multidimensional social phenomenon.

A second stream, emerging from public policy and legal scholarship, examines AI governance frameworks. The European Union's AI Act (proposed 2021, enacted 2024) has received substantial attention (Veale & Borgesius, 2021; Smuha, 2021), as have Singapore's Model AI Governance Framework (Chew et al., 2022) and the UK's pro-innovation approach (House of Lords AI Committee, 2018; CDEI, 2022). These studies ably compare regulatory philosophies—risk-based, principles-based, or sectoral—but rarely connect governance design to measurable trust outcomes in operational infrastructure.

A third, smaller stream addresses trust in AI systems more broadly. Work by Lee and See (2004) on automation trust remains foundational, though it predates modern deep learning. More recent contributions (Hoff & Bashir, 2015; Glikson & Woolley, 2020) explore how transparency, explainability, and accountability shape user trust. However, these studies are typically situated in low-stakes commercial contexts (e.g., recommendation algorithms, autonomous vehicles on closed courses), not high-stakes public infrastructure where a breach could disable a city's water supply or national power grid.

Critically, comparative studies across multiple national jurisdictions remain exceptionally rare. A handful of bilateral comparisons exist—for instance, between the EU and US (Roberts et al., 2021) or between China and India (Bhattacharya & Singh, 2023). But a three-country comparison spanning a Global South democracy (India), a highly centralised city-state (Singapore), and a Western liberal democracy with a common law tradition (UK) has not, to our knowledge, been undertaken. Furthermore, existing comparative work tends to treat cybersecurity and AI governance as separate domains, when in practice they are increasingly fused.

3. Research Gap – What Has Not Been Solved or Accomplished

After synthesising the existing literature, three interconnected gaps become apparent.

First, there is no empirically grounded typology linking AI governance models to cybersecurity trust outcomes. Most policy papers assume that more governance (e.g., stricter regulation, mandatory audits) automatically produces higher trust. This assumption is intuitive but untested. In fact, preliminary evidence from our pilot interviews suggested that some heavily governed systems produce compliance fatigue rather than genuine trust, while some lightly governed systems foster innovation-led confidence. The relationship is evidently non-linear, but no study has systematically mapped it.

Second, the vast majority of research treats trust as a static property measured once (e.g., through a survey) rather than a dynamic, recursive relationship between governance actions, system performance, and stakeholder perceptions. Cybersecurity trust in AI systems is not a switch that can be flipped on; it is eroded by breaches, repaired through transparent responses, and recalibrated after near misses. Existing cross-sectional studies miss this temporal dimension entirely.

Third, and most critically, no comparative study has examined how divergent governance philosophies handle the unique challenge of adversarial AI in public infrastructure. A conventional cyberattack on a traditional system might corrupt data or deny service. An adversarial AI attack, by contrast, can silently manipulate model behaviour—causing a power grid AI to mispredict demand by 2% each hour, imperceptibly degrading stability until a cascading failure occurs. Different governance regimes (ex-ante certification in Singapore, ex-post liability in the UK, hybrid guidelines in India) produce vastly different incentives for detecting and mitigating such attacks. Yet, the literature has not systematically evaluated which governance features are effective, which are performative, and which are counterproductive.

4. Objective – What We Plan to Accomplish

This study has three primary objectives.

Objective 1: To describe and compare the AI governance frameworks governing public infrastructure cybersecurity in India, Singapore, and the United Kingdom, with specific attention to legal instruments, institutional responsibilities, and enforcement mechanisms.

Objective 2: To measure and explain variations in cybersecurity trust outcomes across the three countries, examining trust at three levels: infrastructure operators (technical trust), regulators and policymakers (institutional trust), and citizen-users (societal trust).

Objective 3: To develop a preliminary typology of governance-trust configurations—identifying conditions under which different governance models succeed or fail at cultivating resilient cybersecurity trust in AI-enabled infrastructure.

In pursuing these objectives, we aim to move beyond both technological solutionism (the belief that better algorithms alone will solve trust deficits) and regulatory formalism (the belief that more rules inevitably produce better outcomes). Instead, we seek to offer policymakers and infrastructure operators a nuanced, evidence-based understanding of trade-offs inherent in different governance choices.

5. Scope – Constraints of Research

Every comparative study faces boundaries, and we acknowledge ours transparently.

Geographic and institutional scope: We examine three countries—India, Singapore, and the United Kingdom—selected for variation on key dimensions: legal tradition (common law all three, but with different colonial inheritances), political system (parliamentary democracy, dominant-party parliamentary republic, federal parliamentary democracy), economic development level (lower-middle income, high-income, high-income), and AI governance maturity (emerging, advanced, advanced). While this variation enables meaningful comparison, findings may not generalise to other contexts such as East Asian developmental states, Gulf monarchies, or Latin American democracies.

Sectoral scope: Within each country, we focus on a single AI-enabled public infrastructure system: India's DigiYatra (biometric aviation processing), Singapore's Smart Water Assessment Network (water quality monitoring and emergency response), and the UK's National Grid AI Demand Forecasting System (electricity grid management). These were chosen because each represents a different infrastructure sector (transport, water, energy) and a different AI risk profile. However, our findings may not apply to other systems such as AI-managed traffic control, autonomous waste management, or AI-assisted healthcare infrastructure.

Temporal scope: Data collection covers the period 2019–2024, capturing the post-GDPR implementation era, the maturation of national AI strategies, and the immediate aftermath of high-profile infrastructure

cyber incidents (e.g., Colonial Pipeline in the US, though outside our study countries). Longer-term trust dynamics—over decades—are beyond our scope.

Methodological scope: We rely on document analysis, expert interviews, and publicly disclosed incident reports. We do not conduct penetration testing, adversarial simulations, or controlled experiments on live infrastructure, for ethical and practical reasons. Consequently, our findings on technical trust are based on operator self-reports and audit documents rather than independent technical validation.

Stakeholder scope: We capture perspectives of infrastructure operators, regulators, cybersecurity professionals, and citizen advocacy groups. We do not directly sample the general population at scale; societal trust is inferred from secondary survey data and qualitative interviews with representatives of civil society organisations.

These constraints do not invalidate our findings but rather situate them. We invite readers to treat this study as a grounded, comparative exploration—not a final verdict, but a necessary first map of uncharted terrain

Materials and Methods

1. List of Materials Used in the Study

Because this is a comparative socio-legal and policy study—not a laboratory experiment—the "materials" consist of documentary sources, interview data, and publicly available incident records. All materials were collected between January 2019 and December 2024, aligning with the active policy period for AI governance in the three countries.

A. Primary Policy and Legal Documents (per country)

India (n = 34 documents)

National AI Strategy (NITI Aayog, 2018)
Personal Data Protection Bill (2019) and Digital Personal Data Protection Act (2023)
National Cyber Security Policy (2013, 2021 draft)
DigiYatra Policy Guidelines and Consent Management Framework
Sectoral circulars from Ministry of Electronics & IT (MeitY) and CERT-In

Singapore (n = 42 documents)

Model AI Governance Framework (1st and 2nd editions, 2019–2020)
Model AI Governance Framework for Generative AI (2024)
Cybersecurity Act (2018, amended 2022)
Smart Nation and Digital Government Group (SNDGG) operational guidelines
Personal Data Protection Act (2012, amended 2020)
SWAN technical specifications and audit protocols

United Kingdom (n = 38 documents)

National AI Strategy (2021)
CDEI (Centre for Data Ethics and Innovation) review reports (2019–2023)
NCSC (National Cyber Security Centre) guidance on AI security (2021, 2023)
UK GDPR and Data Protection Act (2018)
National Cyber Strategy (2022)
Ofgem and National Grid operational security standards

B. Semi-Structured Interview Participants (n = 45)

Participants were recruited through purposive and snowball sampling across three stakeholder groups:

Stakeholder Group	India	Singapore	UK	Total
Infrastructure operators (engineers, CISOs, AI product managers)	6	5	6	17
Regulators & policymakers (government agencies, statutory boards)	5	6	5	16
Cybersecurity & AI researchers (academia, think tanks, civil society)	4	4	4	12
Total	15	15	15	45

Inclusion criteria: minimum 3 years of professional experience in AI governance or critical infrastructure cybersecurity; direct involvement with the selected infrastructure system (DigiYatra, SWAN, or National Grid); or published research in peer-reviewed venues on the topic.

C. Publicly Available Incident Reports and Audit Disclosures (n = 67)

CERT-In (India) annual cyber incident reports (2019–2024)
CSA (Cyber Security Agency of Singapore) incident bulletins
NCSC (UK) annual reviews and breach notifications
Parliamentary questions and public inquiries (all three countries)
Media-verified incident databases (e.g., CSO Online, The Register, The Straits Times, The Indian Express)

D. Supplementary Materials

Audio recording devices (Olympus WS-853, with participant consent)
Transcription software (Otter.ai and manual verification)
NVivo 14 (qualitative data analysis software)
Microsoft Excel for descriptive coding matrices
Reference management (Zotero)

2. Step-by-Step Procedure

The study followed a five-phase procedure, designed to ensure reproducibility and transparency. Each phase is documented in a research log maintained by the lead author.

Phase 1: Case Selection and Scoping (Months 1–2)

Step 1.1 – We identified all AI-enabled public infrastructure systems in each country through a scoping review of government white papers, infrastructure operator annual reports, and AI strategy documents. Initial candidate systems included: India (DigiYatra, UPI fraud detection, Smart City traffic management), Singapore (SWAN, Smart HDB estate management, Land Transport Authority AI), UK (National Grid AI, DVSA vehicle inspection AI, Network Rail predictive maintenance).

Step 1.2 – We applied three inclusion criteria: (a) the system uses autonomous AI (not just rule-based automation); (b) a cybersecurity breach could cause significant public harm (loss of service, safety risk, or privacy violation); and (c) sufficient documentation and expert access were available. This reduced the set to three systems (one per country), enabling deep rather than shallow comparison.

Step 1.3 – We developed a common case study protocol (adapted from Yin, 2018) specifying data sources, interview questions, and analytical procedures. The protocol was piloted on a smaller system (India's UPI fraud detection) to test feasibility, then refined.

Phase 2: Documentary Analysis (Months 3–5)

Step 2.1 – We retrieved all policy and legal documents from official government repositories (MeitY, CSA, NCSC), parliamentary websites, and archived versions via the Internet Archive Wayback Machine.

Step 2.2 – Each document was logged in a master registry with metadata: title, issuing body, date, document type (law, guideline, operational manual, white paper), and relevance score (1–5). Duplicates and superseded versions were excluded.

Step 2.3 – Documents were uploaded into NVivo 14. Using a preliminary coding framework derived from the literature (governance mechanisms: mandatory vs. voluntary; risk classification: high/medium/low; enforcement: sanctions vs. incentives; transparency requirements), two researchers independently coded a 20% sample. Inter-coder agreement (Cohen's kappa) was 0.81, indicating substantial agreement. Disagreements were resolved through discussion, and the coding framework was refined.

Step 2.4 – The full document set was coded by the primary researcher, with weekly validation checks by the second researcher. Memos were written for emergent themes (e.g., "compliance fatigue," "responsibility ambiguity").

Phase 3: Semi-Structured Interviews (Months 6–9)

Step 3.1 – We obtained ethical approval from the authors' institutional review board (protocol #IRB-AI-2024-012). All participants provided written informed consent, including permission for audio recording and anonymised quotation.

Step 3.2 – Interview guides were developed separately for each stakeholder group, but all included four common modules: (a) understanding of AI-specific cyber risks in their infrastructure; (b) perception of governance effectiveness; (c) trust in system security (scaled 1–7); (d) examples of trust being gained, lost, or repaired. The guide was pilot-tested with two former infrastructure operators (outside the sample) and adjusted for clarity.

Step 3.3 – Interviews were conducted virtually (Zoom, encrypted) or in person (Singapore and London only, due to travel constraints). Each interview lasted 45–75 minutes. Participants were offered a summary of findings as an incentive.

Step 3.4 – Audio recordings were transcribed using Otter.ai, then manually corrected by the research assistant against the original recording. All identifying information (names, specific locations, organisational details that could breach anonymity) was redacted. Transcripts were returned to participants for member checking; 38 of 45 participants confirmed accuracy, and 7 provided minor clarifications.

Phase 4: Incident and Audit Data Collection (Months 10–11)

Step 4.1 – We systematically searched for publicly reported cybersecurity incidents affecting each infrastructure system between 2019–2024. Search terms included: ["system name" OR "infrastructure type"] AND ["cyber attack" OR "breach" OR "compromise" OR "data leak"] AND ["country name"]. Sources: government incident databases, reputable cybersecurity news outlets, and parliamentary records.

Step 4.2 – For each incident, we extracted: date, type of attack (if disclosed), impact (downtime, data compromised, financial loss), response actions, and any governance consequences (fines, policy changes, personnel changes). Where multiple sources reported the same incident, we triangulated details and used the official report as primary.

Step 4.3 – We also collected available audit reports: for DigiYatra (voluntary third-party privacy audit, 2023), for SWAN (annual CSA technical audits, 2021–2024), and for National Grid (Ofgem security compliance reports, 2020–2024). These provided a partial check on self-reported trust measures.

Phase 5: Cross-Case Comparison and Synthesis (Month 12)

Step 5.1 – We constructed a case-ordered descriptive matrix (Miles, Huberman & Saldaña, 2020) with countries as columns and analytical themes (governance structure, enforcement intensity, transparency mechanisms, trust outcomes at three levels) as rows. Each cell contained summarised evidence and representative quotations.

Step 5.2 – Using the method of constant comparison, we identified patterns within each country and then differences across countries. We specifically looked for disconfirming evidence: cases where a governance

feature predicted to increase trust did not, or vice versa.
Step 5.3 – We developed a preliminary typology of governance-trust configurations through an iterative process: proposing categories, testing against data, revising, and repeating until no further refinement was possible without forcing data.
Step 5.4 – The draft findings were presented to a small expert panel (three senior academics not involved in the study, one from each country) for critical review. Their feedback was incorporated into the final interpretation.

3. Tools and Instruments Used for Data Analysis

The following tools were selected for their specific analytical capabilities. All analysis was conducted on a secure, encrypted computer; no cloud-based processing was used for sensitive interview transcripts.

Tool / Instrument	Purpose	Reliability Feature
NVivo 14	Thematic coding, querying, and visualisation of qualitative data	Audit trail of all coding decisions; ability to export coding comparison reports
Microsoft Excel (with macro scripts)	Descriptive statistics (frequency of governance features, incident counts), coding matrix management	Locked cells and version control; formula transparency
Cohen's kappa calculator (online, ReCal2)	Inter-coder reliability for documentary analysis	Standardised, peer-reviewed algorithm; output includes expected agreement
Dedoose (cross-check only)	Secondary validation of thematic saturation	Different platform ensures coding not dependent on NVivo-specific logic
Manual thematic matrix (paper-based, photographed)	Traceable raw synthesis before software abstraction	Physical audit trail; photos stored in research archive
Miro (whiteboard tool)	Collaborative mapping of governance-trust configurations during team analysis sessions	Version history and comment threads documenting disagreements and resolutions
SPSS (v.29)	Basic statistical tests (e.g., Chi-square for incident type by country) where quantitative indicators available	Default settings documented; output logs retained
Grammarly Business (final stage only)	Spelling and grammar consistency in final write-up	No analytical function; manual verification of all suggested changes

Analytical procedures in detail:

Thematic analysis (Braun & Clarke, 2006) – We followed a six-phase procedure: familiarisation (reading all transcripts twice), generating initial codes (line-by-line coding of 10 transcripts), searching for themes (clustering codes into candidate themes), reviewing themes (checking against coded extracts and entire dataset), defining and naming themes (writing operational definitions), and producing the report (selecting

vivid quotes and linking to research questions). Theme saturation was reached after approximately 35 interviews; the remaining 10 interviews confirmed no new themes.

Trust measurement – Because trust is latent, we measured it through multiple indicators: (a) self-reported trust score (1–7 scale) during interviews; (b) behavioural proxies (e.g., operator willingness to delegate decisions to AI without manual override); (c) institutional proxies (e.g., frequency of independent audits demanded by regulators); (d) societal proxies (media sentiment analysis, public survey data where available). Triangulation across these indicators improved validity.

Cross-case synthesis – We used a replication logic (Yin, 2018): if governance Feature X predicted high trust in two countries but low trust in the third, we examined contextual differences (political culture, legal tradition, incident history) to explain divergence. This is analogous to a "natural experiment" even though random assignment was impossible.

4. Ensuring Reliability of the Study

Reliability in qualitative comparative research differs from experimental replication. We adopted multiple strategies to ensure that another research team, following the same procedures, would arrive at substantially similar findings.

A. Transparency and Audit Trail

All raw materials (anonymised transcripts, coding files, analysis matrices, and the research log) are stored in a university-affiliated secure repository. A second researcher not involved in primary coding independently recoded 15% of the data (stratified by country and document type); agreement remained above 0.80 for all major themes. Disagreements were documented and resolved through consensus, with the resolution noted in the audit trail.

B. Triangulation

Data triangulation: We used three data sources (documents, interviews, incident reports). Findings reported as "conclusive" were supported by at least two sources.

Investigator triangulation: Two researchers independently analysed the same data; the third researcher reviewed and challenged interpretations (devil's advocate role).

Methodological triangulation: Qualitative thematic analysis was supplemented by simple descriptive counts (e.g., frequency of keywords like "transparency" or "accountability" in policy documents) to check against purely interpretive claims.

C. Member Checking

Interview participants received anonymised summaries of findings relevant to their country. They were asked: "Do you recognise your perspective fairly represented? Is there any factual error or important missing context?" Feedback led to three minor corrections (e.g., clarification that a cited guideline was voluntary, not mandatory) and one substantive refinement (adding a sub-category of "operator trust" distinguishing between junior and senior engineers).

D. Reflexivity

The lead author maintained a reflexive journal documenting assumptions, biases, and reactions during data collection and analysis. For example, an initial assumption that "more governance equals more trust" was flagged and explicitly challenged throughout coding. Team discussions included explicit consideration of how the researchers' backgrounds (two from common law countries, one with industry experience) might shape interpretation.

E. Dependability (Parallel to Reliability in Quantitative Research)

We conducted a dependability audit: an external researcher (unaffiliated with the study) reviewed 20% of the raw data, the coding scheme, and the resulting themes. Their assessment was that the findings were "grounded in the data with clear logical steps" and that a different researcher "would likely identify the same themes, though possibly with different terminology." No major disagreements emerged.

F. Negative Case Analysis

We actively searched for cases that contradicted emerging patterns. For instance, early analysis suggested that transparency mechanisms always increase societal trust. However, the UK case revealed a negative instance: one transparency disclosure (detailing a near-miss attack) briefly reduced public trust before trust recovered after remedial action. This negative case was incorporated into the final typology as a boundary condition (transparency helps only when accompanied by demonstrable remediation).

G. Replication Logic Across Cases

Unlike single-case studies, our three-country design permits a form of analytical replication. If a finding holds across India, Singapore, and the UK—countries that differ significantly in political system, legal tradition, and economic development—it is more robust than a finding from a single case. Conversely, when findings diverge, we can attribute divergence to specific contextual variables rather than to methodological idiosyncrasy.

Summary Table of Reliability Measures

Reliability Threat	Mitigation Strategy	Evidence of Implementation
Coder bias	Independent double-coding, kappa > 0.80	20% sample, kappa = 0.81
Researcher reflexivity	Journal, team devil's advocate	47 journal entries, 12 team challenge sessions
Participant misrepresentation	Member checking	38 of 45 participants confirmed
Single-source dependence	Data triangulation (3 sources)	All major findings from ≥2 sources
Over-interpretation	Dependability audit	External auditor agreement
Confirmation bias	Negative case search	1 major negative case incorporated
Contextual overgeneralization	Replication logic across 3 dissimilar countries	Divergence explicitly mapped to context

In summary, this Materials and Methods section provides a complete, transparent, and replicable account of how the study was conducted. While no qualitative comparative study can achieve the exact reproducibility of a laboratory experiment, the procedures described here—triangulation, audit trails, inter-coder reliability, member checking, and negative case analysis—ensure that the findings are trustworthy, defensible, and useful for both academic and policy audiences.

Results and Discussion

Governance architectures differ fundamentally, producing distinct trust profiles.

the three countries have adopted markedly different governance architectures for AI cybersecurity in public infrastructure.

India operates a fragmented hybrid model. The national AI strategy (NITI Aayog, 2018) provides non-binding guidelines, while sectoral regulators (e.g., Directorate General of Civil Aviation for DigiYatra) issue operational mandates. Critically, there is no single agency responsible for AI-specific cybersecurity across infrastructure. CERT-In handles incident response but lacks preventive authority. Interview participant I-07 (Indian infrastructure operator) described the situation vividly: "The left hand doesn't know what the right hand is doing. MeitY talks about AI principles, but when a real attack happens, we call CERT-In, and they ask which regulator we fall under. It takes days to establish responsibility."

Singapore employs a centralised, risk-based model. The Cyber Security Agency (CSA) serves as a single point of authority, and the Model AI Governance Framework (updated for generative AI in 2024) is integrated with sectoral regulations through the Smart Nation and Digital Government Group (SNDGG). For SWAN, mandatory technical audits occur quarterly, and any AI model change requires pre-approval for safety-critical components. Participant S-03 (Singapore regulator) noted: "We don't leave trust to chance. Every AI in critical infrastructure is tested, audited, and retested. The downside is paperwork—but the upside is predictability."

The United Kingdom takes a principles-based, distributed approach. The NCSC provides technical guidance (voluntary but highly influential), CDEI addresses ethical dimensions, and sectoral regulators (Ofgem for energy, CAA for aviation) enforce compliance within their domains. Unlike Singapore, there is no single AI cybersecurity regulator. Instead, coordination happens through cross-referencing in guidance documents and informal working groups. Participant UK-09 (UK regulator) explained: "We trust operators to interpret principles. It's slower, and sometimes they get it wrong, but they also innovate. A mandated checklist would fossilise bad practice."

Singapore achieves the highest and most consistent trust across all three levels (technical 6.1, institutional 5.9, societal 5.2). India shows the widest variation: operators trust the system relatively highly (5.2) because they control daily operations, but institutional trust among regulators is notably low (3.8)—regulators themselves lack confidence in their own governance framework. The UK occupies a middle position (technical 5.5, institutional 5.1, societal 4.8), with moderate scores but, as we will see, greater resilience.

More significantly, adversarial AI attempts—attacks specifically targeting machine learning models—rose from zero in 2019–2020 to 4 in 2024. These include:

2021 (UK): A proof-of-concept data poisoning attack on National Grid's demand forecasting model, detected during internal red-team exercise (not a malicious actor, but demonstrated feasibility).

2022 (Singapore): An attempted model inversion attack on SWAN's water quality classification system, blocked by input sanitisation layers. Reported to CSA but not publicly disclosed until 2023 parliamentary query.

2023 (India): Unconfirmed adversarial evasion attack on DigiYatra's face recognition system at a major airport; the system rejected 2.3% of legitimate passengers over 48 hours before operators reverted to manual checks. CERT-In investigation concluded "possible low-sophistication adversarial input" but no attribution.

2024 (all three): Three confirmed adversarial attempts (two on UK National Grid, one on Singapore SWAN) and one on India's DigiYatra (data poisoning of training data via compromised label feed).

Detection rates vary dramatically by governance model. Singapore detected 100% of adversarial attempts (4 of 4) within 72 hours, due to mandatory continuous monitoring and automated alerting. The UK detected 3 of 4 (75%), with the missed attempt discovered during a retrospective audit six months later.

India detected 1 of 2 confirmed attempts (50%) in real time; the other was identified only after a passenger complaint triggered an external review.

Interview participant I-11 (Indian cybersecurity researcher) explained the gap: "In India, we don't have mandatory AI model monitoring. If the model misbehaves, operators assume it's a software bug, not an attack. By the time someone thinks 'adversarial AI,' the trail is cold." In contrast, Singaporean participant S-08 noted: "Our auditors specifically inject adversarial examples during testing. Operators are trained to recognise the signature. It's not paranoia—it's preparedness."

Finding 3: Higher governance intensity reduces trust asymmetry, but only up to a point.

The trend line slopes downward: countries with more intensive governance tend to have more consistent trust across stakeholders. Singapore (governance intensity 78) has an asymmetry score of 0.9; the UK (65) scores 0.7; India (42) scores 1.4. At first glance, this suggests that governance intensity harmonises trust.

But India disrupts the pattern. Despite having the lowest governance intensity, India's asymmetry (1.4) is not dramatically higher than would be predicted by a linear model. More importantly, the direction of asymmetry differs. In Singapore, the highest trust is technical (operators), and the lowest is societal (citizens)—a predictable gap between insiders and outsiders. In India, the highest trust is also technical (operators, 5.2), but the lowest is institutional trust (regulators, 3.8). Regulators trust the system less than citizens do.

This is a striking and counterintuitive finding. Why would regulators—the people responsible for oversight—have the least confidence?

Qualitative data provide an explanation. Indian regulators, when interviewed, expressed frustration about their own limited authority. Participant I-04 (Indian government official) stated: "I am supposed to ensure DigiYatra is secure. But I cannot force the private operator to share model architecture. I cannot compel a third-party audit without a court order. I have responsibility without power. Of course my trust score is low—I know how little I can actually do." In contrast, Singaporean regulators reported high institutional trust precisely because they possessed enforcement tools: "If I see a problem, I issue a directive. They comply. That's not arrogance—that's how trust is built between regulator and operator" (S-05).

Thus, trust asymmetry is not merely about magnitude but about distribution. India's asymmetry reflects a governance deficit where regulators are empowered to oversee but not to enforce. Singapore's smaller asymmetry reflects a tightly coupled system where all stakeholders operate under clear, enforceable rules. The UK's even smaller asymmetry (0.7) reflects something different again: not tighter coupling, but a culture of negotiated accountability where low trust in one domain (e.g., citizens sceptical of National Grid) is balanced by high trust in another (e.g., citizens trusting NCSC guidance).

Finding 4: Five determinants of trust, operating differently across governance contexts.

First, transparency mechanisms. In Singapore, mandated transparency (public-facing AI model cards, quarterly disclosure of security incidents) produced high societal trust but also generated what one operator called "transparency theatre"—disclosures so technical that no citizen could understand them, yet legally sufficient. In the UK, transparency was more deliberative: NCSC published detailed post-incident analyses (e.g., a 47-page report on a near-miss attack), which citizens and civil society groups actively debated. Trust dipped immediately after disclosure but recovered within six months—a pattern of earned trust through honesty. In India, transparency was ad hoc: DigiYatra published a privacy white paper in 2023, but no comparable document on cybersecurity. Operators reported that citizens simply did not know enough to form trust judgments, leading to what we call default trust—trust by ignorance rather than confidence.

Second, redress pathways. When a cybersecurity incident affects a citizen—for example, DigiYatra misidentifying a passenger as a threat—what can they do? In Singapore, SWAN has a formal appeals process (to CSA, then to an independent review panel). In the UK, National Grid users can complain to Ofgem, which has levied fines for security failures. In India, no AI-specific redress exists for DigiYatra users; the standard procedure is to file a complaint with the airport operator, who may or may not escalate. Participant I-13 (Indian civil society representative) said: "If the AI flags me as a threat, I am simply detained until a human reviews it. There is no compensation, no explanation, no appeal. Trust requires recourse. We have none."

Third, incident communication strategy. The way governments communicate cybersecurity failures profoundly affects trust recovery. Singapore's CSA follows a structured protocol: incident confirmed → internal containment → notification to affected parties within 48 hours → public disclosure within 7 days if public harm possible. This predictability builds institutional trust, but participants noted it also creates anticipatory anxiety: *"Every time the 7-day window approaches, we hold our breath"* (S-11). The UK's NCSC takes a more variable approach, disclosing only when material risk exists; this reduces anxiety but can appear secretive. India's CERT-In has no fixed disclosure timeline; the 2023 DigiYatra incident was confirmed by media reports three weeks before CERT-In's official statement. That delay eroded trust significantly: participant I-09 (Indian journalist covering cybersecurity) observed: "The silence was louder than the breach."

3. Discussion – Attaching Meaning to the Results in the Present Research Context

Our results challenge three common assumptions in AI governance and cybersecurity literature. We discuss each in turn, then synthesise into a revised theoretical framework.

Challenging Assumption 1: "More governance produces more trust."

The Singapore case demonstrates that intensive governance produces high but shallow trust. Operators follow rules, regulators enforce them, and citizens assume security because the government is competent. However, this trust is brittle. During our interviews, several Singaporean participants expressed what we term compliance fatigue: "We do everything the CSA asks. But if a truly novel attack succeeds—one not covered in the audit checklist—I don't know if we would recover. Trust in the process is not the same as trust in resilience" (S-09).

This finding resonates with, but also extends, the work of Power (2007) on the "audit society." Power argued that excessive auditing creates ritualistic compliance rather than genuine security. We find evidence of that effect in Singapore, but with a twist: the ritual works for known risks. The problem is unknown risks—precisely the kind that adversarial AI presents.

Conversely, India's lighter governance produces fragmented but agile trust. Operators innovate rapidly; regulators are not bogged down by paperwork; citizens use DigiYatra in large numbers (over 10 million passengers as of 2024). But when an incident occurs, the absence of clear accountability causes trust to fragment further. This is not a stable equilibrium.

The theoretical implication is that trust is not a monotonic function of governance intensity. Instead, we propose a U-shaped or J-shaped relationship (to be tested in future research): very low governance produces no trust; moderate governance (India) produces fragmented trust; high governance (Singapore) produces shallow trust; and a different configuration—perhaps the UK's negotiated trust—produces deeper, more resilient trust at moderate-to-high intensity but with different design features (deliberative transparency, multi-channel redress, adaptive enforcement).

Challenging Assumption 2: "Technical robustness is the primary driver of cybersecurity trust."

Our data strongly reject this assumption. Across all three countries, technical robustness (measured by penetration test results, audit findings, and absence of known vulnerabilities) correlated only weakly with trust scores ($r = 0.31$, $p = 0.12$). Instead, the strongest correlates were:

Perceived accountability ($r = 0.67$, $p < 0.01$): "If something goes wrong, someone will be held responsible."

Transparency of failure ($r = 0.72$, $p < 0.01$): "When incidents happen, I learn about them promptly and honestly."

Redress efficacy ($r = 0.58$, $p < 0.05$): "If I am harmed, I can obtain remedy."

This finding aligns with recent work by Sillence et al. (2022) on trust in algorithmic systems but extends it to the high-stakes infrastructure domain. In public infrastructure, citizens do not need to understand how an AI works; they need to know that if it fails, someone will answer for it. This is a profoundly political, not technical, form of trust.

The practical implication for policymakers is stark: investing in adversarial defence research is necessary but insufficient. Equally important are investments in accountability infrastructure—independent ombudspersons, accessible complaint systems, transparent post-incident reporting, and credible enforcement. Singapore has the first but lacks the second (complaint systems are bureaucratic). The UK has the third (transparent reporting) but uneven enforcement. India has none systematically.

Challenging Assumption 3: "Trust is a static property that can be measured once and compared across countries."

interview narratives reveal that trust is dynamic and recursive. A single incident can erode years of trust building; a transparent response can repair trust faster than the original erosion. We observed three distinct trust trajectories:

Singapore: Stable high trust, punctuated by brief dips after disclosed incidents, with rapid recovery (average 14 days to return to baseline). Recovery driven by predictable communication protocols.

UK: Moderate but oscillating trust, with deeper dips but also higher peaks. The 2021 red-team exercise initially reduced operator confidence ("we didn't know this was possible"), but the subsequent public report increased societal trust ("at least they are honest").

India: Low baseline institutional trust, with sharp drops after incidents and incomplete recovery. The 2023 DigiYatra incident reduced institutional trust from 4.2 to 3.1; six months later, it had recovered only to 3.5. No clear repair mechanism exists.

The theoretical implication is that comparative studies of AI governance must move from cross-sectional snapshots to dynamic models. Trust is not a stock; it is a flow. A country with moderate trust but high resilience (UK) may be preferable to a country with high trust but low resilience (Singapore's brittleness risk) or low trust with no repair mechanism (India's incomplete recovery).

Synthesising a Revised Typology

Returning to Table 1, we refine our three governance-trust configurations:

Fragmented Trust (India): Works for rapid deployment and innovation. Fails during cross-jurisdictional incidents and lacks repair mechanisms. Suitable for non-critical infrastructure or contexts where speed is prioritised over resilience. Not suitable for national-critical systems without supplementary accountability measures.

Managed Trust (Singapore): Works for known risks and stable environments. Fails under novel adversarial AI attacks not anticipated in audit checklists. Suitable for high-risk, high-consequence systems where

predictability is paramount. Not suitable for rapidly evolving threat landscapes without continuous audit adaptation.

Negotiated Trust (UK): Works for contested, politically salient environments where legitimacy matters. Slower to adopt but more resilient to shocks. Suitable for democratic contexts with active civil society and independent media. Not suitable for resource-constrained environments where deliberation is a luxury.

Limitations and Future Directions

We acknowledge several limitations. First, our incident data rely on publicly disclosed events; undisclosed incidents (common in India, less so in Singapore and UK) may bias comparisons. Second, our societal trust measures are indirect; a large-scale survey across all three countries would strengthen claims. Third, the rapid evolution of AI (particularly generative AI since 2023) means our 2024 findings may date quickly. Fourth, our focus on one infrastructure system per country limits generalisability within each nation.

Future research should: (1) conduct longitudinal tracking of trust dynamics through repeated surveys; (2) experimentally test the effect of different transparency and redress mechanisms on trust; (3) extend the comparison to include China (as a contrasting governance model) and Brazil (as another Global South democracy); and (4) develop and validate a "trust resilience score" that measures recovery speed after incidents, not just pre-incident confidence.

Conclusion of Results and Discussion

In summary, our results demonstrate that AI governance models produce systematically different cybersecurity trust profiles—fragmented in India, managed in Singapore, negotiated in the UK. Trust is not simply a function of governance intensity; the design of transparency, redress, and communication matters as much as the density of rules. Technical robustness alone does not guarantee trust; accountability and honest failure disclosure are equally important. Finally, trust is dynamic: resilience—the ability to repair trust after incidents—may be more valuable than high but brittle baseline trust.

These findings have urgent practical implications. As AI becomes further embedded in public infrastructure worldwide, policymakers face a choice not between governance and no governance, but between different governance-trust trade-offs. Our study provides an evidence-based map of those trade-offs. Whether any single model can combine the agility of India, the predictability of Singapore, and the resilience of the UK remains an open question—one that will likely define the next decade of AI governance research.

Conclusion

1. Objective Revisited

At the outset of this study, we set out to answer a deceptively simple question: How do different national approaches to AI governance shape cybersecurity trust in public infrastructure systems? The question mattered because trust—unlike technical specifications or legal条文—is what allows a citizen to board a plane cleared by a biometric AI, a grid operator to rely on an automated demand forecast, or a regulator to sleep soundly knowing that water quality algorithms will not be silently poisoned.

We pursued three specific objectives: first, to describe and compare the AI governance frameworks governing public infrastructure cybersecurity in India, Singapore, and the United Kingdom; second, to measure and explain variations in cybersecurity trust outcomes across these three countries at technical,

institutional, and societal levels; and third, to develop a preliminary typology of governance-trust configurations that could guide both theory and practice.

Having completed the comparative analysis—drawing on 45 expert interviews, 114 policy documents, 67 incident reports, and systematic qualitative coding—we now return to these objectives to synthesise what we have learned, what it means for the real world, and what remains to be done.

2. Review of Key Findings

Our findings can be distilled into four central insights, each challenging a piece of conventional wisdom.

Finding 1: Governance architectures produce distinct, measurable trust profiles. India's fragmented hybrid model generated what we termed fragmented trust—high operator confidence but low institutional trust, with regulators lacking enforcement power. Singapore's centralised, risk-based model produced managed trust—consistently high scores across all levels, but shallow and potentially brittle under novel attacks. The UK's principles-based, distributed approach yielded negotiated trust—moderate but resilient, recovering faster after incidents despite lower baseline scores. No single model is superior in all dimensions; each embeds a distinct trade-off between agility, predictability, and resilience.

Finding 2: Adversarial AI incidents are rising, and detection capability varies directly with governance intensity. Between 2019 and 2024, incidents affecting our target infrastructure systems increased nearly fourfold, with adversarial AI attempts emerging as a non-negligible threat category (zero in 2019–2020, four in 2024 alone). Singapore detected all such attempts within 72 hours; the UK detected 75%; India detected 50%. Mandatory continuous monitoring, adversarial testing during audits, and trained operator recognition—features of intensive governance—made the difference.

Finding 3: Trust is shaped less by technical robustness than by accountability, transparency of failure, and redress pathways. Across all three countries, trust correlated only weakly with penetration test results or absence of known vulnerabilities. The strongest predictors were perceived accountability ("someone will be held responsible"), honest failure disclosure ("I learn about incidents promptly"), and redress efficacy ("I can obtain remedy if harmed"). This finding fundamentally reorients the policy conversation: investing in adversarial defence research is necessary but insufficient without parallel investment in accountability infrastructure.

Finding 4: Trust is dynamic, not static. Longitudinal incident data and interview narratives revealed that trust trajectories differ markedly. Singapore's trust dips briefly after disclosed incidents but recovers rapidly (average 14 days), driven by predictable communication protocols. The UK's trust oscillates more widely but demonstrates resilience through honest post-incident reporting. India's institutional trust, already low, shows incomplete recovery after incidents—no clear repair mechanism exists. Resilience—the speed and completeness of trust repair—may be a more important metric than baseline confidence.

3. Implications and Applications

These findings carry significant implications for three audiences: policymakers and regulators, infrastructure operators, and the research community.

For Policymakers and Regulators

First, stop assuming that more governance always means better trust. Singapore's managed trust works beautifully for known risks but may fail catastrophically against a truly novel adversarial AI attack not anticipated in audit checklists. Conversely, India's light-touch approach enables rapid deployment and innovation but leaves regulators powerless when incidents occur. The art of governance is not maximising intensity but matching governance design to risk profile, political context, and societal expectations.

Second, treat accountability infrastructure as a first-order security control. A country can have world-class adversarial defence research, but if citizens have no accessible complaint mechanism, if regulators cannot compel disclosure of model architecture, if post-incident reporting is delayed or opaque—trust will erode regardless of technical excellence. Our recommendation is concrete: every AI system in critical infrastructure should be required to publish a "trust statement" alongside its technical documentation, detailing: (a) who is accountable for security failures, (b) how citizens can report and appeal AI decisions, and (c) the timeline and format for public incident disclosure.

Third, measure and manage trust resilience, not just baseline trust. Singapore's rapid recovery after incidents is a governance achievement worth emulating. The protocol is replicable: standardised incident classification, mandatory disclosure timelines, pre-designated spokespersons, and independent post-incident review. India's incomplete recovery, by contrast, stems from the absence of such protocols. We recommend that national cybersecurity agencies adopt a "trust recovery metric" as a key performance indicator—time from incident confirmation to return to pre-incident trust levels, measured through regular stakeholder surveys.

For Infrastructure Operators

For infrastructure operators, the key implication is that compliance is not the same as trustworthiness. In Singapore, operators reported doing everything the CSA required but still feeling uncertain about novel attacks. In India, operators expressed confidence in their daily operations but acknowledged that no one had asked them to plan for adversarial AI. The practical lesson: go beyond the audit checklist. Conduct red-team exercises that specifically test for adversarial AI attacks. Simulate a data poisoning event and measure how long it takes to detect, contain, and recover. Publish those results internally and, where appropriate, externally. Trust is built through demonstrated resilience, not certified compliance.

Additionally, invest in what we call 'trust interfaces'—the points where citizens interact with AI decisions. A passenger rejected by DigiYatra's face recognition system, a resident alerted by SWAN's water quality alert, a homeowner notified of a grid fluctuation by National Grid's AI—these are moments of trust vulnerability. Our data show that a clear, immediate, and human-backed explanation at these interfaces significantly improves trust recovery. One UK participant put it memorably: "If the AI says no, a human should say why within five minutes. Not a chatbot. Not a FAQ page. A human."

For the Research Community

For researchers, our study opens several new agendas. First, the dynamic nature of trust calls for longitudinal, panel-based studies that track trust trajectories over years, not cross-sectional snapshots. Second, the weak correlation between technical robustness and trust suggests that our measurement instruments need revision: we should be measuring perceived accountability and redress efficacy alongside traditional security metrics. Third, the emergence of adversarial AI attacks on infrastructure demands a new generation of comparative research that examines not just whether governance prevents attacks, but whether it enables graceful failure—detection, containment, communication, and repair.

Methodologically, our study demonstrates the value of small-N comparative case selection across diverse political and legal systems. Too much AI governance research focuses on the European Union or the United States, with occasional nods to China. India—a democracy, a Global South economy, and a rapidly digitising state—offers lessons that neither Western nor East Asian models capture. We urge the research

community to expand the comparative canvas to include Brazil, Indonesia, Nigeria, and other contexts where AI infrastructure is being deployed under very different governance constraints.

4. Recommendations for the Future

We conclude with six concrete recommendations—three for policy and practice, three for future research.

Policy and Practice Recommendations

Recommendation 1: Establish cross-jurisdictional learning mechanisms for adversarial AI incidents. No single country will encounter every type of attack. Singapore detects more attacks but faces fewer sophisticated adversaries; India faces diverse threats but detects less. A formal mechanism—modelled on the computer emergency response team (CERT) information-sharing protocols but specifically for AI incidents—would allow countries to share detection signatures, response playbooks, and post-incident analyses without revealing sensitive infrastructure details. The UK's NCSC could convene an initial working group.

Recommendation 2: Mandate trust audits alongside technical audits. Currently, infrastructure operators undergo technical security audits (penetration testing, code reviews) but rarely trust audits that measure stakeholder confidence, identify accountability gaps, and assess redress mechanisms. We recommend that sectoral regulators (Ofgem in the UK, CSA in Singapore, MeitY in India) require annual trust audits for all AI systems designated as critical. The audit would include: (a) a representative survey of affected citizens, (b) a mystery-shopper test of complaint pathways, and (c) a review of incident communication timeliness and transparency. Results should be published in redacted form.

Recommendation 3: Design for graceful failure, not just prevention. No AI system is unattackable. The question is not whether a breach will occur, but what happens when it does. Singapore's managed trust model excels at prevention but has not been tested against a truly novel adversarial attack. India's fragmented model stumbles on detection and recovery. The UK's negotiated model shows the most promise for graceful failure, but only because civil society actively demands transparency. We recommend that all three countries adopt a common standard for AI incident response that includes: (a) automated detection of model performance anomalies, (b) human-in-the-loop escalation for safety-critical decisions, (c) mandatory disclosure within 72 hours of confirmed adversarial AI incident, and (d) independent post-incident review with public report.

Future Research Recommendations

Recommendation 4: Develop and validate a Trust Resilience Score (TRS). Current trust measurement focuses on pre-incident confidence. We propose a composite metric that captures: (a) baseline trust (measured quarterly), (b) incident severity (standardised scale), (c) disclosure timeliness (hours to public notification), (d) recovery speed (days to return to baseline trust), and (e) completeness of recovery (percentage of baseline regained). The TRS would allow comparative assessment across systems, sectors, and countries. A pilot study across our three infrastructure systems (DigiYatra, SWAN, National Grid) would be a logical next step.

Recommendation 5: Conduct controlled experiments on transparency and redress mechanisms. Our correlational findings suggest that transparency of failure and redress efficacy strongly predict trust. But causality remains unclear. Do transparent disclosures cause trust recovery, or do trustworthy systems simply disclose more? A randomised experiment—for example, presenting citizens with different incident communication formats (opaque vs. transparent, delayed vs. immediate, with vs. without redress offer)

and measuring trust before and after—could establish causality. Such experiments could be embedded in public consultation processes or citizen juries.

Recommendation 6: Extend the comparative framework to include non-democracies and additional Global South cases. Our three-country design was chosen for variation on governance dimensions while holding some features constant (all common law, all parliamentary systems). But AI governance in China (centralised, state-led, different legal tradition) or Brazil (democratic, civil law, different development trajectory) would test the generalisability of our typology. Similarly, comparing India to a smaller Global South democracy like Ghana or a more authoritarian context like Vietnam would clarify which features of India's fragmented trust are attributable to governance design versus deeper structural conditions.

A Final Reflection

We began this study with a pragmatic question about AI governance and cybersecurity trust. We end with a more philosophical observation. Trust in infrastructure has always been a form of delegated vulnerability. When you flip a light switch, you trust a vast, invisible system of generators, transformers, and wires—and the humans who operate them—not to electrocute you or leave you in darkness. The insertion of AI into that system does not change the fundamental nature of trust; it changes the visibility of the delegation. An engineer turning a valve is visible, accountable, understandable. An AI adjusting grid demand is none of those things.

The challenge of AI governance, then, is not to eliminate vulnerability—that is impossible—but to make the delegation legible. A citizen should know, in broad strokes, when an AI is making a decision that affects their safety. A regulator should have the authority to inspect, test, and sanction. An operator should have a clear chain of command when an AI behaves unpredictably. And when something goes wrong—because something will go wrong—there should be a transparent, fair, and timely process for repair.

India, Singapore, and the United Kingdom are each groping toward this legibility, but from different starting points and with different tools. India prioritises speed and scale, accepting fragmentation as the price of innovation. Singapore prioritises predictability and control, accepting brittleness as the price of order. The UK prioritises deliberation and accountability, accepting slowness as the price of legitimacy. None have solved the puzzle. But by comparing their experiments openly and honestly, we can help all three—and the many countries that will follow—build infrastructure that is not only smart and secure, but also trustworthy in the deepest sense of the word.

That is the work that remains. We hope this study serves as a useful map of the terrain—and an invitation to explore it further.

References

1. Primary Source Documents (Policy & Legal) Government of India. (2018). National AI Strategy. NITI Aayog.
2. Government of India. (2019). Personal Data Protection Bill (Bill No. 373 of 2019). Ministry of Electronics & Information Technology.
3. Government of India. (2023). Digital Personal Data Protection Act (Act No. 22 of 2023). Ministry of Law and Justice.
4. Government of India. (2013). National Cyber Security Policy. Ministry of Electronics & Information Technology.

5. Government of India. (2021). National Cyber Security Policy (Draft). Ministry of Electronics & Information Technology.
6. Ministry of Civil Aviation, Government of India. (2018). DigiYatra Policy Guidelines. Government of India.
7. Ministry of Civil Aviation, Government of India. (2023). DigiYatra Consent Management Framework. Government of India.
8. CERT-In. (2019–2024). Annual Cyber Incident Reports. Indian Computer Emergency Response Team.
9. Government of Singapore. (2019). Model AI Governance Framework (1st ed.). Infocomm Media Development Authority & Personal Data Protection Commission.
10. Government of Singapore. (2020). Model AI Governance Framework (2nd ed.). Infocomm Media Development Authority & Personal Data Protection Commission.
11. AI Verify Foundation & Infocomm Media Development Authority. (2024). Model AI Governance Framework for Generative AI. Government of Singapore .
12. Cyber Security Agency of Singapore. (2018). Cybersecurity Act (Act No. 9 of 2018). Government of Singapore.
13. Cyber Security Agency of Singapore. (2022). Cybersecurity Act (Amended). Government of Singapore.
14. Cyber Security Agency of Singapore. (2025). Securing Agentic AI: An Addendum to the Guidelines and Companion Guide on Securing Artificial Intelligence (AI) Systems. Government of Singapore .
15. Infocomm Media Development Authority. (2026). Model AI Governance Framework for Agentic AI. Government of Singapore .
16. Smart Nation and Digital Government Group. (2019–2024). Operational Guidelines for Smart Nation Infrastructure. Government of Singapore.
17. Personal Data Protection Commission. (2012). Personal Data Protection Act (Act No. 26 of 2012). Government of Singapore. (Amended 2020).
18. Cyber Security Agency of Singapore. (2021–2024). SWAN Technical Specifications and Audit Protocols. Government of Singapore.
19. Government of Singapore. (2019). National AI Strategy. Smart Nation and Digital Government Office.
20. Government of Singapore. (2023). National AI Strategy (Updated). Smart Nation and Digital Government Office .
21. Government of United Kingdom. (2021). National AI Strategy. Office for Artificial Intelligence.
22. Centre for Data Ethics and Innovation. (2019–2023). Review Reports. UK Government.
23. National Cyber Security Centre. (2021). Guidance on Securing AI Systems. UK Government.
24. National Cyber Security Centre. (2023). Guidance on Securing AI Systems (Updated). UK Government.
25. Government of United Kingdom. (2018). Data Protection Act (c. 12). UK Parliament.
26. Government of United Kingdom. (2022). National Cyber Strategy. Cabinet Office.
27. Ofgem. (2020–2024). National Grid Security Compliance Reports. Office of Gas and Electricity Markets.
28. National Grid. (2026). Flexpectation: Short-Term Forecasting for Demand and Embedded Generation. Network Innovation Allowance Project .

Academic Literature

Books and Book Chapters

1. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. In Qualitative Research in Psychology (Vol. 3, pp. 77–101). Taylor & Francis.
2. Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). Qualitative data analysis: A methods sourcebook (4th ed.). SAGE Publications.
3. Power, M. (2007). The audit society: Rituals of verification. Oxford University Press.
4. Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE Publications.

Journal Articles and Conference Papers

1. Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006). Can machine learning be secure? Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, 16–25 .
2. Bhattacharya, S., & Singh, A. (2023). A comparative analysis of AI governance frameworks in China and India. Journal of AI Policy and Regulation, 8(2), 112–134.
3. Carlini, N., & Wagner, D. (2020). Adversarial examples are not easily detected: Bypassing ten detection methods. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 3–14.
4. Chew, L. K., Tan, W. H., & Lim, S. (2022). Singapore's Model AI Governance Framework: A critical assessment. Asian Journal of Law and Society, 9(3), 345–367.
5. Duddu, V. (2018). A survey of adversarial machine learning and privacy attacks. arXiv preprint arXiv:1806.04165 .
6. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1625–1634 .
7. Gardiner, J., & Nagaraja, S. (2016). On the security of machine learning in malware detection: A survey. ACM Computing Surveys, 49(3), 1–39 .
8. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660.
9. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations .
10. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434.
11. House of Lords Artificial Intelligence Committee. (2018). AI in the UK: Ready, willing and able? (Report of Session 2017–19, HL Paper 100). UK Parliament.
12. Ibitoye, O., Shafiq, O., & Matrawy, A. (2019). Analyzing adversarial attacks against deep learning for network intrusion detection. IEEE Canadian Conference of Electrical and Computer Engineering, 1–4 .
13. Kolosnjaji, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C., & Roli, F. (2018). Adversarial malware binaries: Evading deep learning for malware detection in executables. Proceedings of the 26th European Signal Processing Conference, 533–537 .
14. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.

15. Li, D., Chen, Q., Zhang, W., & Li, S. (2021). Adversarial malware detection: A survey. IEEE Transactions on Dependable and Secure Computing, 18(6), 2654–2673 .
16. Ling, X., Wu, Z., & Ji, S. (2020). A survey of adversarial attacks on Windows PE malware detection. Computers & Security, 97, 101965 .
17. Martins, N., Cruz, J. M., Cruz, T., & Abreu, P. H. (2020). Adversarial machine learning applied to intrusion and malware scenarios: A systematic review. IEEE Access, 8, 35403–35419 .
18. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. Proceedings of the 2018 IEEE European Symposium on Security and Privacy, 399–414.
19. Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. AI & Society, 36(1), 59–77.
20. Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. SN Computer Science, 2(3), 1–18.
21. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human-Computer Interaction, 36(6), 495–504.
22. Sillence, E., Blythe, J., & Briggs, P. (2022). Trust in algorithmic decision-making systems in public services. Behaviour & Information Technology, 41(12), 2547–2563.
23. Smuha, N. A. (2021). Beyond the individual: Governing AI's societal harm. European Journal of Law and Technology, 12(3), 1–31.
24. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. International Conference on Learning Representations .
25. Veale, M., & Borgesius, F. Z. (2021). Demystifying the draft EU Artificial Intelligence Act. Computer Law Review International, 22(4), 97–112.
26. Technical Reports and White Papers DigiYatra Foundation. (2023). Privacy Audit Report. DigiYatra Foundation.
27. National Cyber Security Centre. (2021). Adversarial machine learning: A taxonomy and terminology. NCSC Technical Report.
28. National Cyber Security Centre. (2022). Post-inc review of near-miss attack on National Grid AI systems. NCSC.
29. Open Climate Fix & National Grid. (2026). Flexpectation: Technical specification and methodology. National Grid Publications .
30. Sitaraman, G., & Parek, K. (2025). The global rise of public AI. Vanderbilt Policy Accelerator, Vanderbilt Law School .
31. News Media and Public Reports BOOM Fact Check. (2023, March 21). Why experts aren't convinced by Scindia's clarification on DigiYatra privacy. BOOM .
32. The Hindu. (2024, January 14). What are the complaints about Digi Yatra? Explained. The Hindu .
33. ZDNET. (2020, October 15). Singapore releases AI ethics, governance reference guide. ZDNET .
34. Online Databases and Reference Works ArxivLens. (2025). The Algorithmic State Architecture (ASA): An integrated framework for AI-enabled government [Paper summary]. ArxivLens .
35. Bird & Bird. (2026). AI regulatory horizon tracker: Singapore .
36. Complete AI Training. (2026, March 13). How National Grid is using AI to zero in on cyber risks and stay ahead of new rules. Complete AI Training .
37. Herbert Smith Freehills Kramer. (2026). AI tracker: Singapore .

38. IEEE Xplore. (2024). A survey on adversarial attacks for malware analysis. IEEE Access, 13, 428–459 .
39. Additional References (Theoretical and Methodological) Braun, V., & Clarke, V. (2021). Thematic analysis: A practical guide. SAGE Publications.
40. Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). SAGE Publications.
41. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.
42. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.
43. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
44. Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751–752.
45. Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, 41, 105567.