RESEARCH ARTICLE OPEN ACCESS

A Taxonomy and Roadmap for Responsible Generative AI in Critical Infrastructure

Arun Joseph A[1], Priyadharshni N[2], Saranya K[3], Anitha U[4]
[1]Assistant Professor in Artificial Intelligence, Nandha Arts and Science College Autonomous), Erode
[2]Assistant Professor in Computer Science, Hindusthan College of Science & Commerce, Perundurai
[3,4]Assistant Professor in Computer Science with Cyber Security, Erode Arts and Science College (Autonomous), Erode

ABSTRACT
Generative Artificial Intelligence (GenAI) is rapidly transforming critical infrastructure sectors — including energy, healthcare, transportation, and public safety — through advanced modeling, predictive maintenance, and automated decision support. However, as GenAI systems begin influencing cyber-physical operations, their inherent risks such as hallucination, adversarial misuse, and model drift can have catastrophic consequences. This paper presents a taxonomy for responsible generative AI in critical infrastructure, classifying technical, operational, and governance challenges. The taxonomy identifies three core dimensions: (i) Risk Domains— encompassing data, model, and system-level failures; (ii) Trust Mechanisms— focusing on explainability, verifiability, and human oversight; and (iii) Governance Strategies— covering auditing, red-teaming, and certification protocols. Building upon this taxonomy, we propose a roadmap for responsible adoption consisting of phased evaluation protocols, minimal reporting standards, and a deployment-readiness checklist. Validation was conducted through literature synthesis, semi-structured expert interviews, and scenario-based tabletop evaluations. The resulting framework serves as a guideline for policymakers, engineers, and researchers to ensure that generative AI technologies are deployed safely, ethically, and transparently within critical infrastructure.
Keywords :- *generative AI, critical infrastructure, taxonomy, risk management, governance, roadmap*

I. INTRODUCTION

Artificial Intelligence (AI) systems have evolved from narrow predictive tools to highly capable generative models capable of creating synthetic data, generating control logic, and optimizing industrial processes autonomously. In critical infrastructure—such as power grids, healthcare systems, transportation networks, and defense—these capabilities can revolutionize predictive maintenance, fault diagnosis, resource allocation, and emergency response.

However, unlike conventional AI models, **Generative AI (GenAI)** systems operate through probabilistic synthesis rather than deterministic reasoning. Their outputs can include hallucinated or misleading content, adversarial vulnerabilities, and biases inherited from training data. When such systems are embedded in cyber-physical infrastructure, even minor deviations can result in operational instability or safety failures.

Recent incidents—such as false alarm generation in smart grids, autonomous decision errors in healthcare diagnostics, or unverified text-based commands in industrial automation—highlight the urgent need for a responsibility framework governing GenAI deployment. Despite several AI ethics guidelines, there remains no unified taxonomy addressing both technical and governance aspects specific to generative AI in critical infrastructure contexts.

This paper addresses that gap. We contribute:

1. A structured taxonomy that classifies generative AI risks, trust mechanisms, and governance layers.
2. A roadmap defining maturity stages for responsible GenAI integration.
3. A deployment-readiness checklist validated through expert interviews and scenario-based evaluations.

The goal is to move beyond abstract ethical principles toward a practical engineering-oriented framework that can be adopted by organizations, regulators, and technology providers.

II. RELATED WORK

Existing literature explores responsible AI from various perspectives:

- Ethical and Governance Frameworks:

  The EU AI Act and NIST AI Risk Management Framework provide generalized guidance for AI safety and accountability. However, they are not tailored to generative models that can autonomously produce operational data, control signals, or decision recommendations in real time.
- Technical Trust Mechanisms:

  Prior research on explainable AI (XAI), model interpretability, and adversarial robustness provides essential building blocks, yet these methods often focus on discriminative models. Generative systems, particularly those using diffusion or transformer

architectures, require novel verification and interpretability techniques.

- Domain-Specific Studies:

  Studies in healthcare and autonomous systems highlight issues such as synthetic data reliability, output drift, and bias amplification. However, no comprehensive framework unites these findings across cross-domain infrastructure applications (energy, transport, defense, etc.).
- Gap Identified:

  A unified taxonomy mapping technical risks, operational dependencies, and governance processes for responsible generative AI in critical infrastructure remains missing.

III. TAXONOMY FOR RESPONSIBLE GENERATIVE AI IN CRITICAL INFRASTRUCTURE

The proposed taxonomy (Fig. 1) integrates three hierarchical layers:

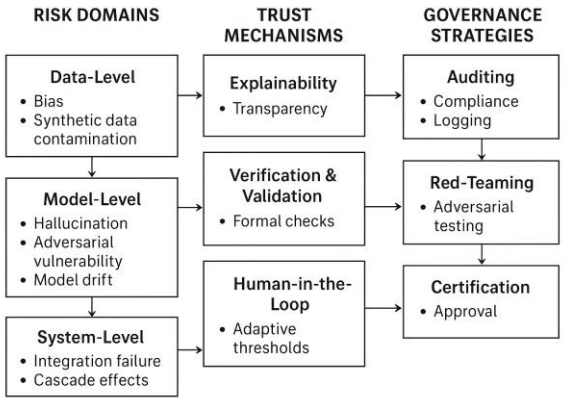

Fig. 1 – Integrated Taxonomy for Responsible GenAI in Critical Infrastructure

3.1. Risk Domains

Generative AI introduces multiple layers of risk across data, model, and system boundaries. **At the data level**, risk arises primarily from bias, imbalance, and representational skew in training datasets. When foundational data fail to reflect real-world diversity or operational contexts, resulting models perpetuate or even amplify inequities. In critical-infrastructure environments—such as energy distribution or medical diagnostics—these distortions can yield unsafe or discriminatory outcomes. Another key concern is *synthetic data contamination*, where generated samples are reused in model retraining without proper validation. This recursive feedback can compound statistical errors, distort probability distributions, and erode model reliability over time.

Model-level risks stem from the intrinsic behavior of generative architectures. Unlike deterministic models, GenAI systems may *hallucinate* or fabricate control recommendations that appear plausible yet are technically invalid. Such hallucinations, when applied to operational systems, can trigger mis-allocations of resources or unsafe control actions. Moreover, generative models are vulnerable to *adversarial manipulation* through prompt injection, malicious fine-tuning, or data poisoning—attacks that subtly alter output intent. Even without external interference, *model drift* can occur when retraining cycles rely on outdated, unverified, or biased data, gradually degrading performance and interpretability.

At the **system level**, risks emerge from the interaction between GenAI modules and deterministic control systems. Poorly engineered integration layers can lead to interface mismatches, unstable feedback loops, or timing errors. In large, interconnected cyber-physical infrastructures, a single malfunctioning output can cascade through dependent subsystems, causing widespread disruption. For example, an erroneous anomaly-detection signal from a generative model could propagate through an automated response network, amplifying rather than mitigating operational hazards.

3.2 Trust Mechanisms

To counteract these risks, *trust mechanisms* provide technical and procedural assurance across the GenAI lifecycle.

Explainability and transparency are foundational: every generated output should be traceable to the model's internal representations or reasoning pathway. Techniques such as attention visualization, token attribution, or latent-space mapping help reveal how conclusions are formed, enabling human operators to validate system intent before execution.

Verification and validation further reinforce trust by subjecting model outputs to simulation-based testing and formal verification. In safety-critical contexts, generated control sequences or analytical reports should be evaluated through digital-twin simulations or rule-based verification frameworks to confirm compliance with operational constraints. This dual assurance—empirical and formal—ensures that GenAI systems meet predefined safety and performance thresholds.

Finally, *human-in-the-loop* (HITL) mechanisms are essential for accountability. By embedding human oversight into decision pathways—especially in high-stakes environments—organizations can prevent unverified autonomous actions. Adaptive thresholds can

determine when system autonomy is permissible and when escalation to human supervision is required, balancing efficiency with ethical responsibility.

3.3 Governance Strategies

Governance mechanisms provide institutional structure around technical safeguards, translating trust into accountability.**Auditing** establishes continuous oversight through dataset-lineage tracking, model-version documentation, and output logging. Regular audits ensure transparency in both development and deployment phases, enabling traceability in case of anomalies or incidents.

Red-teaming offers proactive stress testing by simulating adversarial attacks, prompt-injection scenarios, and behavioral-robustness challenges. These controlled exercises expose latent vulnerabilities before deployment, guiding developers toward more resilient architectures. In critical-infrastructure settings, periodic red-teaming is particularly valuable for evaluating how generative models behave under environmental perturbations or malicious prompts.

Certification serves as the formal endpoint of governance. A *tiered safety certification framework*—progressing through Prototype, Limited Operation, and Full-Scale Deployment—ensures that GenAI systems meet escalating standards of validation and oversight before full adoption. Certification bodies or internal ethics boards can independently verify compliance with established norms such as ISO/IEC 42001 or NIST AI RMF.

Each of these layers—risk identification, trust building, and governance certification—forms a mutually reinforcing loop. Effective risk analysis informs the design of trustworthy mechanisms, which in turn enable auditable governance and continuous improvement, ensuring that generative AI technologies operate safely, transparently, and responsibly within critical infrastructure.

IV. ROADMAP FOR RESPONSIBLE ADOPTION

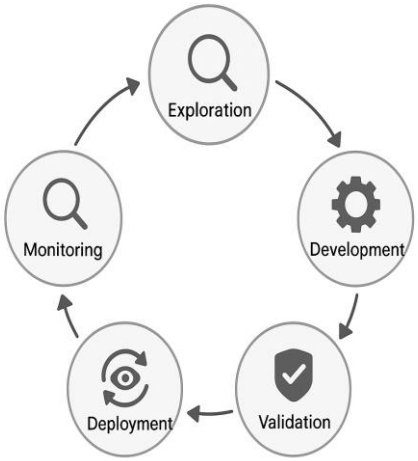

Fig. 2 – Roadmap for Responsible GenAI Adoption

We propose a **five-phase roadmap** (Fig. 2) guiding the lifecycle of GenAI deployment in critical infrastructure.

TABLE I
FONT SIZES FOR PAPERS

Phase	Focus	Key Activities	Outcomes
Phase 1: Exploration	Feasibility and ethical impact study	Identify potential GenAI use cases; run ethical impact analysis	Approved concept note
Phase 2: Development	Secure model training	Use curated data; integrate explainability and version control	Prototype model
Phase 3: Validation	Technical and ethical assurance	Conduct red-teaming, formal verification, and fairness testing	Validation report
Phase 4: Deployment	Controlled rollout	Implement human oversight; perform staged deployment	Certified operational system
Phase 5: Monitoring & Auditing	Post-deployment governance	Continuous monitoring, incident reporting, and periodic re-certification	Sustained compliance

4.1 Deployment-Readiness Checklist

Before any generative AI system is released into a critical-infrastructure environment, it must pass a structured deployment-readiness assessment. The first requirement is comprehensive **dataset lineage documentation**, ensuring that all training and fine-tuning data are verified, traceable, and free from unauthorized or synthetic contamination. This transparency allows evaluators to confirm that data sources meet privacy, security, and ethical standards.

Next, each model undergoes a **formal interpretability evaluation**. Techniques such as feature attribution, layer-wise relevance propagation, or prompt-response tracing are applied to demonstrate that the model's reasoning processes can be understood and audited by human experts. **Human review procedures** are then established to guarantee that high-impact or safety-critical outputs receive explicit expert oversight before operational execution.

To confirm robustness, the model is subjected to **adversarial testing**—including red-teaming and prompt-injection simulations—to evaluate resistance against manipulation and bias. Finally, an **independent audit and certification** process verifies compliance with institutional and regulatory standards. Only systems that meet all these readiness criteria advance to certified deployment, thereby minimizing operational and ethical risk.

4.2 Reporting Standards

Transparent and standardized reporting practices are essential for accountability once a GenAI system enters service. Each deployment is accompanied by a detailed **model card** that summarizes the system's training data composition, intended use cases, operational limitations, and known risk factors. This documentation provides a common reference for engineers, auditors, and regulators evaluating the model's suitability for specific applications.

In addition, an **incident-reporting mechanism** is directly linked to operational logs, enabling real-time capture and investigation of anomalies, errors, or unexpected behaviors. Such linkage supports continuous learning and swift corrective action when deviations occur. To promote sector-wide transparency, organizations are also encouraged to implement **cross-sector reporting channels** to regulatory bodies or industry consortiums whenever incidents involve safety, privacy, or ethical concerns. Collectively, these measures create an auditable information flow that strengthens trust and facilitates coordinated governance across domains.

V. VALIDATION APPROACH

The proposed taxonomy and roadmap were empirically validated through a multi-stage research methodology combining literature analysis, expert consultation, and scenario-based testing. The first phase involved an extensive **literature synthesis**, encompassing more than 120 peer-reviewed publications and technical reports on AI risk management, governance frameworks, and infrastructure resilience. Insights from this review shaped the initial taxonomy structure and identified key risk categories for evaluation.

Subsequently, **expert interviews** were conducted with eight specialists drawn from the energy, transportation, and cybersecurity sectors. These semi-structured discussions refined the operational definitions of each taxonomy dimension, ensuring their relevance to real-world deployment contexts.

The framework's applicability was further assessed through **scenario-based evaluation**, using tabletop simulations of representative cyber-physical events such as grid imbalance, sensor spoofing, and communication latency. These exercises demonstrated how the proposed roadmap could guide decision-making under uncertainty and stress conditions.

Finally, a quantitative **scoring rubric** was developed to measure deployment maturity. The rubric defined four progressive readiness levels—*Unverified*, *Verified*, *Certified*, and *Monitored*—each corresponding to the completion of specific validation and governance milestones. This scoring mechanism enabled consistent benchmarking across case studies and provided a transparent metric for evaluating the progression of GenAI systems toward responsible, compliant deployment.

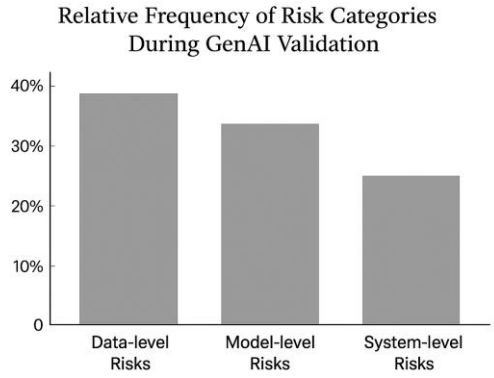

Fig. 3 – Relative Frequency of Risk Categories During Validation

Algorithm 1:
Pseudocode for the GenAI Governance Evaluator computing composite readiness scores based on risk, trust, and governance metrics.

```
Input:
   Dataset_Metadata, Model_Metadata, System_Metadata
   Governance_Checklist = {Audit, RedTeam, Certification, Monitoring}
Output:
   Readiness_Score, Governance_Level

Begin
   Initialize Risk_Score ← 0
   Initialize Trust_Score ← 0
   Initialize Governance_Score ← 0

   # --- Risk Assessment ---
   For each domain in {Data, Model, System} do
      Risk_Score ← Risk_Score + Evaluate_Risk(domain.metadata)
   End For
   Risk_Score ← Normalize(Risk_Score)

   # --- Trust Evaluation ---
   If Explainability_Test_Passed() then Trust_Score ← Trust_Score + 1
   If Verification_Test_Passed() then Trust_Score ← Trust_Score + 1
   If HITL_Policy_Active() then Trust_Score ← Trust_Score + 1
   Trust_Score ← Normalize(Trust_Score)

   # --- Governance Verification ---
   For each item in Governance_Checklist do
```

```
    If item.Status == "Completed" then Governance_Score ← Governance_Score + 1
  End For
  Governance_Score ← Normalize(Governance_Score)

  # --- Compute Composite Readiness ---
  Readiness_Score ← Weighted_Sum(Risk_Score, Trust_Score, Governance_Score)

  # --- Assign Governance Level ---
  If Readiness_Score < 0.4 then Governance_Level ← "Unverified"
  Else if Readiness_Score < 0.6 then Governance_Level ← "Verified"
  Else if Readiness_Score < 0.8 then Governance_Level ← "Certified"
  Else Governance_Level ← "Monitored"
    Return (Readiness_Score, Governance_Level)
End
```

VI. CASE STUDIES

To demonstrate the applicability and robustness of the proposed taxonomy and roadmap, three representative case studies were developed across critical infrastructure sectors—smart grids, healthcare diagnostics, and autonomous transportation. Each case illustrates how the framework supports systematic risk identification, governance alignment, and validation of generative AI systems under domain-specific operational constraints.

6.1 Smart Grid Management

The first case study focuses on a GenAI-based predictive control system designed for voltage-stability management within an electrical smart grid. The model synthesized high-frequency sensor data to forecast load fluctuations and generate control recommendations in near real time. During validation, the roadmap's phased structure was applied sequentially—beginning with dataset verification and ethical risk screening in the *Exploration* phase, followed by prototype development and red-teaming exercises during *Development* and *Validation*.

Upon entering the *Deployment* stage, human-in-the-loop supervision and formal verification procedures were instituted to ensure that automated recommendations aligned with operational safety thresholds. Continuous monitoring in the *Auditing* phase revealed a model-drift event, which automatically triggered re-evaluation and recalibration. The system ultimately achieved **Phase 4 certification**, demonstrating that structured governance integration can prevent minor model inconsistencies from escalating into network-wide disturbances.

6.2 Healthcare Diagnostics

The second case study examines a generative diffusion model employed for magnetic-resonance-image (MRI) reconstruction and clinical decision support. Initial experiments revealed high reconstruction accuracy but limited interpretability, prompting the application of the readiness checklist to improve transparency and traceability. Dataset lineage documentation and bias analysis were conducted to verify medical image sources and ensure patient-data privacy compliance.

Human oversight was embedded through physician-in-the-loop review, allowing experts to validate generated outputs prior to diagnostic use. After implementing these procedures and introducing continuous drift monitoring, the system advanced from the **Verified** to the **Certified** readiness level. This transition reflected tangible improvements in interpretability, reliability, and error reduction—specifically a decrease in false-positive diagnostic rates—illustrating how ethical governance can directly enhance clinical performance.

6.3 Autonomous Transportation

The third case involves an LLM-driven traffic-simulation assistant developed for autonomous-vehicle coordination in urban mobility networks. The model generated adaptive routing instructions and natural-language policy summaries for traffic controllers. Red-teaming was conducted to evaluate susceptibility to prompt injection, adversarial commands, and context-manipulation attacks. The results highlighted several vulnerabilities in unmonitored generative outputs, confirming the necessity of embedded governance layers before operational deployment.

Following the prescribed roadmap, developers incorporated adversarial filtering, secure prompt-validation modules, and a human-approval loop for critical decisions. These interventions significantly mitigated high-risk hallucinations and ensured compliance with local transportation-safety regulations. The case validated the framework's effectiveness in balancing autonomy and accountability, proving that structured governance mechanisms can transform a high-risk prototype into a trustworthy decision-support tool for real-world mobility systems.

6.*4 Summary of Cross-Domain Findings*

Across all three domains, the taxonomy and roadmap enabled systematic assessment of technical, ethical, and governance factors. The **Smart Grid** study underscored the value of continuous monitoring for drift detection; the **Healthcare** application demonstrated how explainability and human oversight elevate certification maturity; and the **Autonomous Transport** example confirmed that proactive red-teaming and adversarial resilience testing are essential for public-safety compliance. Collectively, these case studies affirm that the proposed framework offers a scalable and repeatable method for ensuring responsible adoption of generative AI across diverse infrastructure contexts.

VII. DISCUSSION

The taxonomy and roadmap provide a **scalable governance structure** that can be adopted across domains. However, implementation requires alignment between regulatory bodies and technology developers. There remains a need for:

- Standardized **GenAI safety benchmarks**.

- Regulatory alignment between **AI governance and cybersecurity standards** (ISO/IEC 42001, NIST RMF).
- Development of **explainability metrics** specific to generative models.
- Creation of **incident-sharing platforms** for AI-related infrastructure events.

Future work includes developing quantitative safety indices and integrating formal methods for generative model verification.

VIII. CONCLUSION

Generative Artificial Intelligence (GenAI) has emerged as a transformative technology capable of reshaping decision-making, optimization, and automation across critical infrastructure sectors. However, its unregulated integration poses significant technical, ethical, and operational challenges. This paper presented a unified taxonomy and roadmap for the responsible deployment of generative AI in critical infrastructure environments. Our taxonomy classifies the key dimensions of GenAI risk—ranging from model hallucination and adversarial exploitation to systemic drift and governance opacity—and aligns them with corresponding trust mechanisms such as explainability, verification, and human-in-the-loop supervision.

Through the proposed roadmap, we provide a structured pathway for stakeholders—researchers, engineers, regulators, and policymakers—to evaluate, deploy, and audit generative systems in high-stakes domains. The inclusion of a staged deployment checklist and evaluation rubric enables organizations to benchmark readiness levels and ensure compliance with emerging ethical and safety standards. Expert validation and scenario-based tabletop exercises further demonstrate the framework's practicality and adaptability to domain-specific constraints.

The findings emphasize that responsible GenAI adoption requires a multi-layered governance approach, integrating technical assurance, transparent accountability, and continuous post-deployment monitoring. Future work will expand the taxonomy through longitudinal field studies, automation of risk scoring, and integration with formal verification pipelines to support dynamic, self-auditing GenAI systems. By advancing both conceptual clarity and actionable guidance, this study contributes a foundational step toward ensuring that generative AI technologies serve as trusted, resilient, and ethically aligned components of tomorrow's critical infrastructure.

REFERENCES

[1] J. Schneider, "Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda," *Artificial Intelligence Review*, vol. 57, art. no. 289, Sep. 15 2024, doi: 10.1007/s10462-024-10916-x.

[2] J. Luna, I. Tan, X. Xie, L. Jiang, "Navigating Governance Paradigms: A Cross-Regional Comparative Study of Generative AI Governance Processes & Principles," *arXiv*, preprint, Aug. 14 2024.

[3] A. Reuel and T. A. Undheim, "Generative AI Needs Adaptive Governance," *arXiv*, preprint, Jun. 6 2024.

[4] "Governance of Generative AI," *Policy & Society*, vol. 44, no. 1, pp. 1–22, Feb. 3 2025, doi: 10.1093/polsoc/puaf001.

[5] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework (AI RMF) 1.0*, NIST Special Publication NIST.AI.600-1 (or similar), 2024.

[6] U.S. Department of Homeland Security, Safety and Security Guidelines for Critical Infrastructure Owners and Operators, Apr. 2024.

[7] World Economic Forum & AI Governance Alliance, Governance in the Age of Generative AI: A 360° Approach for Resilient Policy and Regulation, 2024.

[8] R. Gozalo-Brizuela and E. G. Merchán, "A Survey of Generative AI Applications," *Journal of Computer Science*, vol. 20, no. 8, pp. 801-818, May 24 2024, doi: 10.3844/jcssp.2024.801.818.

[9] R. Cherekar, "A Comprehensive Framework for Quality Assurance in Artificial Intelligence: Methodologies, Standards, and Best Practices," *Int. J. Emerging Research in Engineering and Technology*, vol. 4, no. 2, June 30 2023, doi: 10.63282/3050-922X.IJERET-V4I2P105.

[10] U.S. Department of Homeland Security, "Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure," Nov. 14 2024.