ECLECTIC TECHNOLOGY

Critical Infrastructure Cybersecurity and AI Assessment

Tiered Assessment Framework for Cognitive Errors in Generative AI Systems for Critical Infrastructure Applications

A Quality Assurance Methodology for Industrial Control Systems

Version 2.8 — September 2026

Status: Released

For Use by Asset Owners/Operators, ICS Suppliers, and Critical Infrastructure Sector Regulators

SUGGESTED CITATION

T. Roxey, "Tiered Assessment Framework for Cognitive Errors in Generative AI Systems," Version 2.8, Eclectic Technologies, September 2026.
Concept DOI: https://doi.org/10.5281/zenodo.21363865
Version DOI (this version): 10.5281/zenodo.22726480 (supersedes v2.7, 10.5281/zenodo.22726006)

Erratum notice — the published predecessors. Two published records of this framework carry a sentence that this version corrects. Version 2.5 (version DOI 10.5281/zenodo.21363866) and Version 2.6 (version DOI 10.5281/zenodo.21813691) state at §2.4: "Consequently, deterministic ML can be deployed autonomously, even in Highest-Criticality applications, provided appropriate verification is completed." The paragraph containing it is byte-identical in both.

That sentence allows a technology class to settle a placement question. Determinism sets the scope of the Command Broker obligation and not the placement of actions: a component deterministic within a stated input bound is outside the obligation, but a deterministic component can still produce an action above the Bright Line, and placement is decided action by action on the consequence of the action (FD-BL §4.2–§4.4; FD-LD). The corrected statement is at §2.4 of this version.

Neither published record is edited. A version DOI names a fixed text, and editing the text a version DOI resolves to destroys the only thing a version DOI does; a notice carried by the successor is the remedy for a fixed text that is wrong. The concept DOI 10.5281/zenodo.21363865 resolves to the most recent version and will resolve to this one on publication.

Erratum notice — the revision history of Version 2.7. Version 2.7 (version DOI 10.5281/zenodo.22726006) carries a revision history that is incomplete and that contains a sentence belonging to its own drafting rather than to its record. The history holds no entry for Version 2.6, running from 2.7 directly to 2.5, so that published record cannot tell its reader what the August 2026 version changed. The Version 2.7 entry closes a passage with the sentence “All four corrections are now in and this version is ready to publish,” which was true of the draft and is not true of the published document, and which carries two literal asterisk pairs that are an artifact of the drafting medium rather than emphasis.

This version writes the missing Version 2.6 entry and removes the drafting sentence from the Version 2.7 entry. The removal is disclosed here and in the Version 2.8 entry, and the sentence is quoted above, so that the correction is recorded rather than absorbed: a revision history edited silently to match present state stops being evidence of anything. Version 2.7 is not edited and retains both defects as the record of what issued.

Document Control Information

Document Title	Tiered Assessment Framework for Cognitive Errors in Generative AI Systems for Critical Infrastructure Applications
Version	2.8
Date	September 2026
Status	Released
Developed by	Eclectic Technology
Applicable Standards	ASME NQA-1, NRC Regulatory Guides (1.152, 1.168, 1.169, 1.173), NERC CIP Standards, IEEE 1012, IEC 61508, ISA/IEC 62443, ISAGCA Bright Line Criteria, NIST AI RMF
Companion Document	A Study of AI Risks (Vendor-Agnostic) — provides regulatory context and sector-specific use cases to inform the standards development requirements in this framework; available upon request.

Purpose

This framework establishes systematic methods for assessing cognitive errors in generative AI systems deployed in critical-infrastructure environments. The tiered criticality model enables proportional allocation of assessment resources while maintaining comprehensive coverage of safety-critical, quality-of-service, and compliance-driven error categories. The framework supports regulatory due diligence requirements for the sixteen critical infrastructure sectors defined in the National Infrastructure Protection Plan (NIPP).

Intended Audience

- Asset Owners and Operators (AOO) in critical infrastructure sectors
- Industrial Control System (ICS) and Operational Technology (OT) suppliers
- Generative AI system developers and vendors
- Critical Infrastructure and Key Resources (CIKR) sector regulators
- Quality assurance and verification/validation professionals
- Cybersecurity professionals in critical infrastructure protection

Distribution Statement

This document is intended for use by qualified professionals in critical infrastructure protection and AI system assessment. Distribution to support critical infrastructure protection and NIPP implementation objectives is authorized. Commercial use or reproduction requires coordination with Eclectic Technology.

Revision History

Version 2.8 (September 2026): Conforms the framework's own vocabulary and repairs the revision history of Version 2.7. The term error class, which entered this text in the July 2026 reconciliation and was not adopted by the vocabulary determination of 7 September 2026, is replaced at thirty sites on the error axis, including five section headings. The restored terms are not new: they are the terms this framework's Critical Infrastructure variant has carried unchanged since February 2026, and the surrounding prose in this document was never conformed away from them, so that Section 1.2 has read “the tiered framework” and “the tier-specific content” throughout the period in which its own heading said otherwise. The heading of Section 2.6 is the one formulation in the set that is composed rather than restored, because that section's subject is the orthogonality of the two axes and the variant's heading does not name it. Section 2.2 carried the single remaining use of tier on the application axis, where placement was said not to turn on the criticality tier; it now reads criticality band, so that this document uses tier for the error axis and band for the application axis and never one word for both. The Consequence-Severity Tier Model is named and not renamed: it is a designation held elsewhere and a change to it is not this version's to make. On the record side, the missing Version 2.6 entry is written below and the drafting sentence is removed from the Version 2.7 entry, both disclosed in the erratum notice above. This entry records the version and not the passes that produced it, which is the practice the Version 2.7 entry departed from and the reason its drafting sentence survived into a published record. One sentence is repaired for grammar and not for substance: the fourth objective proposed to NIST at Section 11.1 closed with a fragment, in which the clause requiring model updates and changes of operational context to trigger reassessment stood as its own sentence without a subject; it is now joined to the clause it depends on, and the requirement it states is unchanged. The fragment is present in Versions 2.5, 2.6 and 2.7 as published, and it is corrected here rather than noticed because it is an error of grammar in a proposed requirement and not an error in the requirement. One item is registered and not taken: Section 11.2 places this framework's restored tiers beside the tiered criticality models of the NERC CIP standards, which sit on the application axis, and the adjacency invites the conflation this framework exists to prevent. Version 2.8 issues with version DOI 10.5281/zenodo.22726480 under concept DOI 10.5281/zenodo.21363865, superseding v2.7 (10.5281/zenodo.22726006).

Version 2.7 (September 2026): Repairs the determinism/placement fusion at §2.4. The prior text stated that deterministic ML may be deployed autonomously even in Highest-Criticality applications, which allowed a technology class to settle a placement question; the replacement states that determinism sets the scope of the Command Broker obligation and not the placement of actions, and cites FD-BL §4.2–§4.4 and FD-LD. Second drafting pass, 12 September 2026 (predecessor draft MD5 53625b76ac9075bd73ffa43483fef3a3). The Mid-Level qualification is taken at six sites: the §2.2 bridge clause and its parenthesis; the §2.2 converse, which identified below-line with the Lowest band; §11.1’s proposed NIST requirement; §11.2’s proposed NERC requirement; §11.3’s proposed 62443 Security-Level mapping, where an autonomy prohibition keyed to a threat-capability axis would inherit the axis problem this framework exists to name; and §11.4, where the rule survives because Safety-Related and Important-to-Safety are defined by what incorrect performance could cause, and gains the derivation it rested on. §2.1 gains the reconciliation of the maloperation-framed band definitions against the take-the-max rule: both routes assign the same band and the same rigour, and only the maloperation route supports an obligation derived from the necessary existence of an above-line action. Two of the six sites were not in the recorded four and were found by reading the draft rather than the record. Third

drafting pass, 12 September 2026 (predecessor draft MD5 c7086353952c8e05716b8b1a42dbf346): the erratum notice is added to the document-control block, naming both DOI-bearing predecessors by version DOI — v2.5 at 10.5281/zenodo.21363866 and v2.6 at 10.5281/zenodo.21813691, both read from the publications store rather than recalled — and quoting the §2.4 sentence from the published text, which was read in both predecessors and found byte-identical. Neither published record is edited. Recorded and not taken: the occurrences of "error class", a term the vocabulary determination of 7 September 2026 did not adopt; conforming them is a separate revision whose subject is vocabulary. Version 2.7 issues with version DOI 10.5281/zenodo.22726006 under concept DOI 10.5281/zenodo.21363865, superseding v2.6 (10.5281/zenodo.21813691).

Version 2.6 (August 2026): Conformed the framework to the governing Foundational Definitions by reference, following the Foundational Definitions corpus reconciliation. The Bright Line passage at Section 2.2 gained an appended clause citing FD-BL for placement, and the front matter gained the suggested-citation block with the concept and version Digital Object Identifiers. No analytical content was otherwise revised. Published 5 August 2026 under version DOI 10.5281/zenodo.21813691, superseding Version 2.5 (10.5281/zenodo.21363866). This entry is written at Version 2.8 from the deposit record and the reconciliation correspondence: Version 2.6 issued without an entry of its own and Version 2.7 inherited the omission. One fact belongs with it, established by byte comparison of the February 2026 drafting text against the released Version 2.5 and recorded here because a reader cannot otherwise recover it: the bridging clause at Section 2.2, which Version 2.7 qualifies, was not present in the February drafting text and entered at the Version 2.5 release act, in the same edit that added the citation block and the identifiers. It is a clause added at deposit and not a claim the framework made from the outset.

Version 2.5 (February 2026): Improved internal transitions and readability. Added bridging language between detailed and condensed Tier 1 subsections to clarify methodological continuity. Strengthened transition from deployment architecture boundaries to provenance documentation requirements. Restructured NIST objectives in the Standards Development section for improved readability at paragraph density.

Version 2.4: Added new section on Standards Development Requirements for Tiered Assessment and Bright Line Codification, addressing the roles of NIST, NERC, ISA/IEC, and NRC in converting framework content into enforceable standards—corresponding revisions to Conclusion.

Version 2.3 (February 2026): Revised Introduction to explicitly position the tiered framework as one component of the hybrid quality assurance methodology. Strengthened emphasis on the two-layer architecture with formal verification for deterministic infrastructure components and cognitive error assessment for AI layers.

Version 2.0 (February 2026): Revised for improved clarity and flow. Enhanced transitions between sections, clarified tier distinctions, strengthened connections to quality assurance frameworks, and improved readability throughout.

Version 1.0 (February 2026): Initial framework release incorporating tiered criticality model, comprehensive testing methodologies, statistical acceptance criteria, continuous monitoring provisions, and integration with established quality assurance frameworks.

1. Introduction

Large Language Models (LLMs) deployed in industrial control systems represent a transformative technology that demands equally rigorous quality assurance. When these generative AI systems provide decision support, operational guidance, or analytical capabilities in critical infrastructure environments, their cognitive errors can range from user confusion to safety-critical failures. This framework addresses that challenge with a hybrid quality assurance methodology that applies proven verification principles to deterministic components and introduces systematic assessment approaches for cognitive elements exhibiting emergent behavior.

1.1 The Hybrid Quality Assurance Framework

The hybrid methodology recognizes a fundamental architectural distinction in GenAI deployments for critical infrastructure. Each such deployment comprises two interconnected layers that require different yet complementary verification approaches. The infrastructure layer—operating systems, communication protocols, data storage mechanisms, and input/output routines—consists of deterministic components amenable to traditional formal verification using established mathematical techniques. The cognitive layer—natural language processing, reasoning chains, knowledge synthesis, contextual interpretation—exhibits emergent behaviors that resist classical formal methods but still demand rigorous quality assurance commensurate with their deployment criticality.

Traditional formal verification methods prove the correctness of deterministic infrastructure components through mathematical analysis, model checking, and theorem proving. These techniques have matured over decades in safety-critical domains and provide confidence appropriate for the foundations upon which cognitive systems operate. The tiered framework presented in this document addresses the second component: the systematic assessment of cognitive errors in the AI layer that operates on top of this formally verified infrastructure. Together, these approaches constitute a hybrid quality assurance framework that ensures the entire system—from hardware abstraction through AI inference—maintains safety properties appropriate for critical infrastructure deployment.

1.2 The Tiered Criticality Model for Cognitive Errors

The fundamental insight underlying the tiered framework is that not all cognitive errors in AI systems carry equal risk. A generative AI system that occasionally uses awkward phrasing poses fundamentally different concerns than one that hallucinates equipment specifications or fails at multi-step reasoning in fault analysis. This observation motivates the tiered criticality model that structures cognitive error assessment, separating safety-critical failures from quality-of-service issues and compliance-driven concerns.

Three tiers organize cognitive error assessment by the kind of failure and its potential consequences for critical infrastructure deployment. Tier 1, the safety-critical tier, encompasses errors that could directly compromise system integrity, equipment safety, or protective functions. These errors require assessment rigor comparable to that used for verification and validation of safety-related software in nuclear applications or protective relay logic in power systems. Tier 2,

the quality-of-service tier, addresses errors that degrade system utility and user confidence without immediate safety implications. These errors affect operational efficiency and user acceptance but can be managed through training, procedural controls, and appropriate deployment constraints. Tier 3, the compliance-driven tier, covers errors related to regulatory requirements, ethical considerations, and social-responsibility obligations that, while essential for responsible deployment, do not directly threaten system safety.

This tiered approach enables proportional resource allocation while maintaining comprehensive coverage of the cognitive layer. Organizations can focus intensive testing on safety-critical error categories while applying appropriate but less resource-intensive assessments to quality and compliance dimensions. The selective testing enabled by this structure is particularly valuable given the breadth of potential cognitive error modes and the resource constraints most critical infrastructure operators face.

1.3 Framework Audiences and Applications

The framework serves multiple audiences with overlapping but distinct needs. Asset owners and operators (AOO) require methodologies for assessing AI systems before deployment and for monitoring them during operation, whether those systems operate in Command Broker architectures with human decision authority or in lower-criticality applications with limited autonomy. ICS and AI system suppliers need clear guidance on what assessment evidence regulators and customers will expect across different deployment architectures and criticality bands. Sector regulators need structured approaches to evaluate whether organizations have exercised appropriate due diligence in AI deployment decisions while respecting architectural boundaries established through consequence analysis. Quality assurance professionals need connections between AI-specific assessment challenges and the quality frameworks they already understand from nuclear, power systems, and industrial safety domains.

Each tier is addressed in subsequent sections, progressing from safety-critical concerns through quality-of-service issues to compliance-driven requirements. Following the tier-specific content, additional sections address compound error patterns arising from interactions among error types, statistical frameworks for establishing acceptance criteria, continuous monitoring and periodic reassessment requirements, and explicit connections to established quality assurance frameworks, including ASME NQA for nuclear applications, IEEE standards for power system protective relaying, and IEC 61508 for functional safety.

2. Application Criticality and Deployment Architecture Boundaries

The cognitive testing methodologies in this framework apply to generative AI systems across a spectrum of deployment architectures. However, application criticality fundamentally constrains which architectures are acceptable. The International Society of Automation Global Cybersecurity Alliance (ISAGCA) Supplier Working Group has established Bright Line Criteria that define clear boundaries for generative AI autonomy based on the severity of consequences. These criteria recognize that not all control functions should permit autonomous AI operation regardless of system performance.

2.1 Criticality Bands and Regulatory Context

Three criticality bands organize deployment constraints based on consequence severity and sector-specific regulatory requirements. Highest-Criticality applications involve systems where maloperation could cause catastrophic consequences. In the nuclear sector, these correspond to NRC safety-related systems that could fail, affecting reactor safety or radiation containment. In the electric sector, these are FERC/NERC CIP-002 High-Impact BES Cyber Systems, where maloperation could cause cascading grid failures affecting wide geographic areas. Other critical infrastructure sectors have analogous highest-consequence categories defined by their respective regulators.

Mid-Level Criticality applications involve systems where failures produce significant but contained consequences. Nuclear Important-to-Safety systems and FERC/NERC CIP-002 Medium Impact BES Cyber Systems exemplify this category. These systems affect operational effectiveness, equipment protection, or local reliability without threatening wide-area safety or grid stability. Failures have serious operational and economic consequences but remain bounded in scope.

Lowest-Criticality applications encompass systems in which failures have minimal consequences. Nuclear balance-of-plant systems and FERC/NERC Low Impact BES Cyber Systems represent this category. These systems affect efficiency, comfort, or non-critical auxiliary functions. While failures may create inconvenience or inefficiency, they do not threaten safety, reliability, or significant assets.

These bands are assigned by the take-the-max rule of the Consequence-Severity Tier Model: an application reaches a band through whichever consequence category is most severe, so a grid-only asset can reach the Highest band through economic & service-reliability with no safety consequence at all.

The two statements above are read together. The band definitions name maloperation because the paradigm case in each sector is a system whose incorrect operation produces the consequence that sets the band. The take-the-max rule admits a second route, in which an application reaches a band through a category — economic and service-reliability most often — that loss of service can reach without any wrong action at all. Both routes assign the same band and the same assessment rigour, and they do not support the same inferences: an obligation derived from the necessary existence of an above-line action holds on the maloperation route and does not hold on the availability route. An instrument that keys an obligation to a band therefore records which route the application took, and the pinning category the Consequence-Severity Tier Model already requires is what records it.

2.2 The Bright Line

The Bright Line establishes the fundamental constraint: an action whose incorrect performance could produce a consequence at or above the facility's severity threshold may not be executed by a generative component without a human Command Broker interposed. Where a band is entered through maloperation — where incorrect performance could produce the consequence that sets the band — at least one such action necessarily exists, and fully autonomous generative AI operation is therefore prohibited; that is the case for the Highest-Criticality and Mid-Level-Criticality categories as the sector regulators define them. Where a band is entered through loss of service or availability rather than through wrong action, no above-line action is guaranteed by

the band alone, and the prohibition is a conservative default rather than a consequence of placement. Above this line, generative AI may operate only with a human serving as Command Broker in the control loop. The AI guides and informs the operator through analysis, recommendations, and decision support, but the human makes all control decisions and executes all actions affecting the controlled system. This architectural constraint reflects the indeterministic nature of generative AI and the fundamental limits of testing-based assurance for non-deterministic systems in safety-critical applications. Placement is governed by FD-BL: the Bright Line is drawn on the consequence of the individual action, assessed per action — not on the criticality band or the technology class; this framework's band framing applies that definition and does not vary it (where the band is entered through maloperation, the prohibition on fully autonomous generative AI is the application-level expression of an action-level placement; where it is entered through loss of service, it is a conservative default and not that expression). Mode of discharge is governed by FD-BL-D1, indication integrity by FD-BL-D2, and envelope terminology by FD-EV, by reference.

Below the Bright Line, generative AI may take limited autonomous actions. Below-line actions exist in applications at every band, and a Lowest-Criticality application may contain above-line actions; the band sets the rigour of assessment and not the placement of any action. Even in this regime, deployment should incorporate appropriate bounds checking, monitoring, and override capabilities. The limited consequences of Lowest Criticality applications make experimentation and learning acceptable, supporting the development of best practices that might eventually inform deployment in higher-criticality contexts using different architectural patterns.

2.3 The Command Broker Architecture

The Command Broker architecture, as required above the Bright Line, positions the human operator as the authority who receives AI guidance but makes independent decisions. The generative AI system analyzes current conditions, historical patterns, and relevant technical information. It presents findings, identifies options, and may recommend specific actions. The interface supports natural language interaction, allowing operators to probe the AI’s reasoning, request alternative analyses, or seek clarification. However, the AI cannot directly command changes to setpoints, valve positions, or other process control parameters.

The operator, serving as Command Broker, evaluates AI recommendations alongside other information sources, including conventional human-machine interfaces, alarm systems, and their own expertise and training. The operator decides what action, if any, to take and executes it via verified conventional control interfaces. This architecture preserves human judgment and accountability while leveraging AI analytical capabilities. It also provides inherent defense-in-depth—even if the AI hallucinates a recommendation or exhibits reasoning failures, the human broker prevents flawed guidance from affecting the controlled system.

2.4 Generative AI versus Machine Learning Distinction

A critical distinction must be drawn between generative AI and other machine learning approaches. Generative AI systems based on large language models exhibit indeterministic behavior—the same input may produce different outputs, and responses cannot be exhaustively enumerated for verification. This indeterminism precludes the application of formal verification methods and necessitates the cognitive testing approaches detailed in this framework. The

extensive testing required cannot provide the same level of assurance as formal methods for deterministic systems.

Determinism sets the scope of the Command Broker obligation, not the placement of actions. A component that is deterministic within a stated input bound and formally testable as deployed — for example a neural network trained for a specific control task, whose behavior is bounded and reproducible — is outside the obligation and may operate without Command Broker interposition; a generative component is inside it (FD-BL §4.4; FD-LD). Formal verification, stability analysis, and safety-integrity calculation apply to such deterministic components, which are assessed under the existing standards ecosystem. This is a property test applied to the component as deployed, not a technology label, and the carve-out is not the line: it does not follow that deterministic ML sits below the Bright Line or that generative AI sits above it. A deterministic component can produce an action above the line, and a generative action may fall below it, so placement remains an action-by-action determination by consequence (FD-BL §4.2–§4.4). This framework's cognitive-testing methodology addresses the components inside the obligation.

2.5 Framework Scope and Application

This framework addresses the assessment of cognitive errors in generative AI across deployment architectures. The testing methodologies and acceptance criteria apply whether the system operates as decision support in a Command Broker architecture or with limited autonomy in Lowest Criticality applications. However, application criticality affects several assessment dimensions.

Acceptance criteria severity should scale with the criticality of the requirement. Generative AI in Highest Criticality Command Broker roles requires more stringent error-rate thresholds than in Lowest Criticality autonomous roles, because human operators depend on the AI for critical decisions. Testing emphasis should reflect deployment architecture—systems in Command Broker roles require particular focus on how effectively they communicate uncertainty, acknowledge limitations, and support operator decision-making. Systems with autonomous capabilities require emphasis on the effectiveness of bounds checking and behavioral predictability within authorized action spaces.

The statistical framework and continuous monitoring provisions apply across criticality bands, but monitoring intensity and reassessment frequency should increase with criticality. Red teaming is recommended at all criticality bands, as noted in the ISAGCA criteria, with particular emphasis on Highest- and Mid-Level Criticality deployments, where cognitive vulnerabilities could be exploited to influence operator decisions in safety-significant or reliability-critical scenarios.

2.6 Integration of the Tiers with the Criticality Bands

The framework's three tiers—safety-critical, quality-of-service, and compliance-driven—apply orthogonally to the ISAGCA criticality bands. This orthogonality is the point: criticality is a property of the application (how severe the worst credible consequence), the tier is a property of the assessment (what kind of failure is being tested), and the two are graded independently. A Highest-band deployment requires rigorous assessment across all three tiers, since even a

quality-of-service degradation could impair operator effectiveness in critical situations. Mid-band deployments warrant thorough Tier 1 assessment with appropriate Tier 2 and Tier 3 coverage. Lowest-band deployments might focus primarily on Tier 1 if autonomous, or accept lighter assessment if the consequences of all error types remain minimal.

The Bright Line criteria provide architectural boundaries; this framework provides the testing methodology for systems deployed within those boundaries. Together, they establish a comprehensive approach to the responsible deployment of generative AI in critical infrastructure: the ISAGCA criteria ensure appropriate architectural constraints based on the severity of consequences. In contrast, this framework ensures that systems operating within those constraints meet acceptable performance standards through rigorous cognitive assessment.

3. System Provenance Documentation Requirements

Regardless of where a generative AI system falls relative to the Bright Line—whether it operates under Command Broker constraints in higher-criticality applications or with limited autonomy in lowest-criticality deployments—credible cognitive error assessment requires a clear understanding of the system being assessed. Complete provenance documentation establishes the foundation for all subsequent testing and provides the configuration baseline against which changes are managed. While this documentation contains supplier intellectual property that requires confidentiality protections, its completeness determines whether a credible assessment is even possible.

The provenance requirements extend established software configuration management practices from nuclear quality assurance into the AI domain. Just as safety-related software requires documented version control, development procedures, and verification records, AI systems in critical infrastructure demand comparable rigor adapted to their unique characteristics. The machine learning development process differs from traditional software engineering, but the fundamental principles of configuration control, traceability, and documented verification remain applicable.

3.1 Base Model Identification

Documentation must specify the underlying large language model with sufficient precision to enable exact reproduction of the assessed configuration. This includes model architecture family, parameter count, any architectural modifications from standard implementations, and the specific version or checkpoint being deployed. For commercial foundation models, the supplier and model name with the version identifier may suffice. For custom or modified architectures, detailed specifications become necessary.

The composition of the training corpus requires documentation, even when the full corpus cannot be disclosed. At minimum, provenance should characterize data sources, approximate corpus size, temporal coverage, and domain-specific content proportions relevant to the deployment context. For safety-critical applications in power systems, the presence and quality of content covering electrical engineering, protective relaying, and relevant regulatory standards directly affect assessment confidence. A model trained predominantly on general internet text will require more extensive testing than one trained specifically for the deployment domain.

3.2 Fine-Tuning and Adaptation

Any fine-tuning or domain adaptation applied to the base model must be documented with the same configuration management rigor applied to software modifications in safety-related systems. This includes the rationale for fine-tuning decisions, the datasets used, training and hyperparameter settings, convergence criteria, and the validation approaches applied during development. Reinforcement learning from human feedback or other alignment procedures requires documenting feedback-collection methodologies, evaluator qualification criteria, and the specific behaviors being reinforced or suppressed.

The distinction between base-model capabilities and fine-tuning enhancements affects how assessment resources are allocated. Capabilities in the base model benefit from broader validation through general use of the model, while fine-tuned adaptations specific to this deployment require focused testing, since the enhancement may not have been extensively validated elsewhere.

3.3 Guardrails and Constraints

Guardrail implementations—the constraints and filtering mechanisms intended to prevent harmful or inappropriate outputs—require thorough characterization. Content filtering approaches require documentation of prohibited content categories, detection methodologies, and performance characteristics from validation testing, including false-positive and false-negative rates. Behavioral constraints, such as limitations on the types of recommendations the system can offer or on the domains where it must defer to human judgment, require explicit specification and the technical mechanisms that enforce them.

Guardrails are often the primary mechanism for managing AI system risks during operational deployment. Their reliability directly affects overall system safety, yet they are often subjected to less rigorous verification than the underlying AI model. Documentation should address how guardrails are tested, what failure modes have been considered, and whether guardrail effectiveness degrades under particular conditions.

3.4 Version Control and Configuration Management

Version control and configuration management documentation should follow practices analogous to safety-related software requirements. This includes unique version identifiers for all components—base model, fine-tuned adaptations, guardrail implementations, and integration code. Baseline specifications define the exact configuration being assessed, establishing the reference against which any changes are compared. Change control procedures govern how modifications are managed, reviewed, approved, and documented. For systems intended for long-term deployment, update and maintenance procedures must be specified, including criteria that trigger reassessment when significant modifications occur.

The configuration management challenge extends beyond the AI model itself to include integration architecture, operational data sources, and human-AI interaction workflows. Changes in any of these elements can affect system performance even without modifying the AI model. Comprehensive configuration management must therefore encompass the entire AI-enabled system, not merely the generative model component.

4. Tier 1: Safety-Critical Error Assessment

Safety-critical errors represent cognitive failures that could directly compromise system safety, equipment integrity, or protective functions. These errors require assessment rigor comparable to that used for verification and validation of Class 1E systems in nuclear applications or protective relay logic in power systems. The six error categories in Tier 1—factual errors, hallucinations, reasoning limitations, technical errors, temporal errors, and inconsistency—each require comprehensive testing with clearly defined acceptance criteria.

The common thread linking these error types is their potential to produce incorrect outputs that, if acted upon, could result in equipment damage, system instability, safety incidents, or inappropriate protective actions. Whether the generative AI operates in a Command Broker architecture, where operator decisions are influenced by AI recommendations, or in Lowest Criticality applications, where AI directly affects controlled parameters within bounds, these cognitive errors pose risks that require rigorous assessment. A factual error that provides incorrect equipment specifications, a hallucination that fabricates non-existent procedures, or a reasoning failure in fault analysis all share the characteristic that reliance on the flawed output could have safety consequences.

Assessment approaches for Tier 1 emphasize formal verification methods where applicable, extensive testing with representative operational scenarios, and statistical confidence requirements aligned with safety significance. The acceptance criteria specified for each error type reflect this emphasis, generally requiring performance levels comparable to those of safety-related systems rather than accepting commercial-grade quality when demonstrable failures are tolerable.

4.1 Factual Errors

Factual accuracy forms the foundation of reliable AI assistance in critical infrastructure. When operators or engineers query an AI system for equipment specifications, regulatory requirements, or established technical relationships, the responses must be correct. Factual errors encompass incorrect technical information, misattribution of regulatory requirements, and false statements about system specifications or operating parameters. In critical infrastructure contexts, factual accuracy extends beyond general knowledge to include domain-specific technical facts, equipment specifications, regulatory requirements, industry standards, and established operational practices.

Consider a generative AI providing incorrect impedance values for transmission line protection calculations. If an engineer relies on these values for relay coordination, the resulting settings could allow faults to persist longer than intended or cause miscoordination, with downstream devices operating before upstream protection. Similarly, misrepresenting relay coordination time intervals or misinterpreting NERC reliability standards could result in protection system configurations that fail to meet regulatory requirements or industry best practices.

4.1.1 Testing Methodology

Testing for factual errors requires comprehensive benchmark datasets covering the specific technical domain. These datasets must include verifiable facts about equipment specifications from manufacturer documentation; regulatory requirements from authoritative sources such as

NERC reliability standards or NRC regulatory guides; industry consensus standards from organizations like IEEE or ASME; and established technical relationships, such as electrical power equations or thermodynamic principles. The benchmark approach draws from acceptance testing practices in safety-related instrumentation, where known input conditions must produce specified outputs within defined tolerances.

Controlled-questioning protocols should use multiple formulations of the same factual query to assess consistency and robustness. Rephrasing questions using different technical terminology or approaching the same fact from different conceptual angles helps determine whether the system possesses robust factual knowledge or merely pattern-matches against training examples. A system that correctly answers “What is the typical operating voltage of a 345 kV transmission line?” but fails when asked “At what voltage level would you expect to find lines designated as 345 kV class?” demonstrates fragile knowledge despite the questions being essentially equivalent.

Cross-referencing against authoritative sources remains essential throughout testing. Primary sources—equipment manufacturer documentation, regulatory text, and peer-reviewed technical literature—should be prioritized over secondary interpretations. For regulatory requirements, verification should trace to the applicable standard or regulatory document rather than relying on summaries or guidance materials that may introduce interpretive errors.

4.1.2 Acceptance Criteria

Statistical acceptance criteria for factual accuracy must reflect the safety significance of different fact categories. Critical safety parameters, protective setpoint calculations, and regulatory compliance requirements warrant stringent thresholds approaching 99% accuracy or better. General background technical information might accept somewhat lower accuracy levels, though any threshold below 95% raises questions about whether the system provides sufficient value to justify deployment risks.

Error rate tracking should categorize failures by technical domain, severity, and type. Commission errors—providing incorrect information—generally pose a greater risk than omission errors, in which the system acknowledges it lacks information. The graded approach to acceptance criteria parallels the stratification of quality requirements in nuclear QA programs, in which safety-related calculations are verified more rigorously than general design information.

4.2 Hallucinations

Hallucinations are perhaps the most insidious type of error because fabricated information often appears internally consistent, technically sophisticated, and authoritative. Unlike factual errors, where incorrect information might be recognized through inconsistency with other knowledge, hallucinated content may form a coherent but entirely fictional technical narrative. An AI system might generate detailed but fabricated guidance for coordinating overcurrent relays, complete with plausible-sounding time-current curve descriptions and coordination margins that do not exist. Or it might invent NERC standards with convincing numbering schemes and requirements that sound reasonable but are completely fictional.

The risk amplification in critical infrastructure contexts stems from the gap between AI capabilities and typical users' expertise. Engineers familiar with general protective relaying

principles but lacking deep specialization might not recognize that a detailed, technical-sounding relay coordination procedure is entirely fabricated. The presentation confidence typical of large language models provides no reliable signal about content accuracy—hallucinated content often appears with the same confidence as factually correct responses.

4.2.1 Testing Methodology

The fundamental challenge in hallucination detection is distinguishing correct technical information beyond the evaluator's immediate knowledge from fabricated content. Testing must therefore emphasize systematic verification against authoritative sources rather than relying solely on evaluator expertise. Every significant technical claim, equipment specification, or regulatory citation generated by the system should be traced to a verifiable primary source. This verification requirement imposes substantial testing overhead but remains essential for safety-critical applications.

Stress testing through requests for information on edge cases reveals hallucination tendencies. Systems prone to hallucination will generate confident responses to queries about non-existent equipment models, fictional technical standards, or impossible specifications. Testing should include deliberate requests for citations, document references, or technical standard numbers that can be independently verified. The system's handling of knowledge boundaries provides critical insight: does it acknowledge limitations when appropriate, or does it fabricate authoritative-sounding but false information to maintain an appearance of competence?

Temporal boundaries present particular hallucination risks. AI systems may generate plausible technical information about time periods beyond the training data's currency, fabricating details about recent regulatory changes, newly released equipment models, or industry developments. Testing should probe how the system handles queries about recent events and whether it appropriately acknowledges knowledge cutoff limitations or, if equipped with retrieval mechanisms, correctly accesses and validates external data sources.

4.2.2 Acceptance Criteria

Acceptance criteria for hallucination must approach zero tolerance for safety-critical applications. Unlike factual errors, which may be acceptable depending on the severity of the consequences, hallucinations represent a fundamental reliability failure. Any demonstrated tendency to fabricate technical information, regulatory requirements, or equipment specifications should either disqualify the system from safety-critical deployment or require substantial mitigation through mandatory external verification of all technical claims before operational use. The mitigation approach effectively reduces AI to a hypothesis generator that requires independent verification rather than a trusted information source.

The preceding subsections on factual errors and hallucinations illustrate the full assessment methodology in detail—benchmark development, controlled-questioning protocols, cross-referencing against authoritative sources, and statistically grounded acceptance criteria. The remaining safety-critical error categories follow the same methodological pattern, adapted to each error type's specific characteristics. They are presented in condensed form to avoid repetition of the common methodology while highlighting the domain-specific testing emphases and acceptance considerations that distinguish each category.

4.3 Reasoning Limitations

Reasoning capabilities—logical deduction, causal analysis, multi-step problem solving, analogical thinking, and counterfactual reasoning—prove essential for many critical infrastructure applications. AI systems must analyze fault scenarios, trace causation through complex system interactions, evaluate protective coordination schemes, and assess implications of operational changes. Reasoning failures manifest as logical fallacies, incomplete consideration of relevant factors, inability to follow causal chains, or poor performance on multi-step analytical problems. Testing uses scenario-based assessments that mirror real analytical tasks, with acceptance criteria reflecting the criticality of reasoning for the intended application.

4.4 Technical Errors

Technical errors extend beyond content quality to encompass implementation reliability. Numerical precision in calculations, deterministic behavior, response latency under varying loads, integration reliability with operational systems, and graceful handling of edge cases all represent technical quality dimensions analogous to software defects in safety-related systems. Testing addresses numerical accuracy in domain-specific calculations, output consistency for identical inputs, performance under operational loads, integration interface reliability, and boundary condition handling. Acceptance criteria distinguish between critical failures that preclude deployment and lesser issues that can be managed through operational controls.

4.5 Temporal Errors

Temporal awareness encompasses both the currency of training data and the understanding of operational time context. Currency limitations—outdated technical information, obsolete regulatory requirements, discontinued equipment specifications—stem from knowledge frozen in time at the training cutoff. Operational context failures—misinterpreting temporal references or misunderstanding how the system state evolves—represent more fundamental limitations in temporal reasoning. Both aspects challenge the deployment of critical infrastructure, where current information and temporal awareness are often essential. Testing verifies the currency of knowledge for time-sensitive categories, assesses temporal reasoning in operational scenarios, and evaluates the system's handling of distinctions between historical and current information.

4.6 Inconsistency

Consistency—providing non-contradictory information across queries, maintaining coherent positions over extended interactions, and avoiding logical contradictions in responses—builds user trust and supports reliable decision-making. Inconsistent relay coordination recommendations across repeated identical queries, or self-contradictory technical analysis within single responses, undermine confidence in any outputs. Testing employs standardized question batteries repeated across sessions, semantically equivalent questions with different phrasing, multi-turn dialogues to check position coherence, and logical consistency analysis of complex responses. Acceptance criteria require very high consistency for safety-critical technical information, while allowing greater variation in analytical content, provided contradictions are avoided.

5. Tier 2: Quality-of-Service Error Assessment

Quality-of-service errors affect the usefulness of an AI system without directly compromising safety. While these errors do not immediately threaten equipment or personnel, they influence whether operators and engineers find the AI system valuable enough to use, whether they develop appropriate trust calibration, and whether the system enhances or impedes operational efficiency. Degraded quality of service can indirectly affect safety by eroding trust to the point where valuable capabilities are ignored, or by creating inefficiencies that increase operator workload and distraction during critical periods.

The three error categories in Tier 2—overgeneralization, input sensitivity, and ambiguity handling—each affect practical utility. Assessment rigor remains substantial but is calibrated differently from safety-critical testing, with an emphasis on usability studies, operational feedback, and iterative improvement rather than formal verification. Acceptance criteria for quality-of-service errors should account for intended use cases and user expertise, recognizing that some quality limitations can be mitigated through training, clear communication about system capabilities, or deployment constraints.

5.1 Overgeneralization

Overgeneralization manifests when AI systems provide broad, generic responses where specific, nuanced answers are appropriate. In critical infrastructure contexts, this reduces practical value by failing to account for system-specific details, equipment-specific characteristics, or context-specific constraints. An AI that responds to protective coordination questions with general principles rather than specific recommendations tailored to the actual equipment and system configuration provides limited operational value, even if those general principles are technically correct. The user already knows the general principles—they need specific guidance for their particular situation.

Testing uses queries that deliberately require detailed responses referencing specific equipment models, configurations, or scenarios. Comparative questioning reveals whether the system articulates meaningful distinctions between similar approaches or resorts to generic descriptions. Contextual variation testing assesses whether responses are appropriately tailored to different equipment types, voltage classes, operational regimes, or regulatory jurisdictions. Acceptance criteria should account for user expertise and use case—systems supporting experienced engineers require greater specificity than those providing general guidance to broader audiences.

5.2 Sensitivity to Input Variations

Input sensitivity describes how small, inconsequential changes in query formulation affect responses. Excessive sensitivity indicates fragility—good responses for queries matching training examples, but poor performance when questions are phrased differently. This brittleness frustrates users who cannot predict which formulations will work. Testing employs paraphrasing with synonyms and varied sentence structures, noise injection (including common typos and grammatical variations), and context-dependent assessment comparing standalone versus conversational presentations. Statistical characterization should show roughly continuous relationships between the magnitude of input variation and output divergence, rather than sharp discontinuities in which minor changes trigger qualitatively different responses.

5.3 Ambiguity and Misinterpretation

Ambiguity handling determines how effectively systems navigate queries admitting multiple interpretations. Technical communication often contains potential ambiguities—terms with multiple meanings, unclear pronoun references, or implicit contextual assumptions. Human experts resolve these through contextual understanding and, when necessary, clarifying questions. Testing employs deliberately ambiguous queries, contextual disambiguation assessments, and evaluations of domain-knowledge application. Acceptance criteria should emphasize metacognitive awareness—recognizing ambiguity and requesting clarification demonstrates safer behavior than confidently proceeding with potentially wrong assumptions.

6. Tier 3: Compliance-Driven Error Assessment

Compliance-driven errors address regulatory requirements, ethical obligations, and social responsibility concerns that, while not directly safety-critical in the technical sense, are essential to deployment. These error categories respond to regulatory frameworks, industry standards, and societal expectations governing AI deployment in critical infrastructure. Failures may not immediately compromise system safety, but they can expose organizations to regulatory action, liability, or the erosion of public trust.

Assessment approaches draw from compliance audit methodologies rather than technical verification, with emphasis on demonstrable adherence to established norms and documented evidence of appropriate constraints. The single primary category in Tier 3—ethical and social considerations—requires careful evaluation but resists simple testing protocols because it involves normative judgments where societal consensus remains incomplete.

6.1 Ethical and Social Considerations in Critical Infrastructure

Ethical considerations for critical infrastructure AI differ substantially from general AI ethics concerns about offensive content or demographic bias. The relevant ethical issues center on decision frameworks when human safety is at stake, appropriate boundaries for AI autonomy in safety-critical functions, liability allocation when AI recommendations contribute to adverse outcomes, and the balance between optimization objectives and conservative safety margins. These concerns require evaluation through multiple dimensions.

Safety-critical decision framework testing assesses how systems handle scenarios involving conflicting values. Testing presents tradeoff scenarios and evaluates whether reasoning demonstrates appropriate prioritization—safety considerations should generally dominate economic optimization in critical infrastructure contexts. Authority boundary testing assesses whether systems appropriately recognize the limits of their proper role and defer to questions that require professional engineering judgment or regulatory interpretation with compliance implications. Uncertainty communication testing verifies that confidence levels accurately reflect actual reliability rather than presenting all outputs with uniform confidence. Accountability testing ensures reasoning can be documented for engineering review, regulatory examination, or legal proceedings if necessary.

Acceptance criteria necessarily involve normative judgments that testing cannot fully resolve. However, certain behaviors should disqualify systems from deployment: failure to prioritize safety appropriately in critical scenarios, overconfident communication of uncertain analysis, inability to recognize proper authority boundaries, or demonstrated bias compromising impartial

technical analysis. Systems that exercise appropriate caution, acknowledge limitations, defer to human judgment when appropriate, and demonstrate balanced reasoning can proceed with documented ethical considerations and oversight frameworks in place.

7. Compound and Interaction Error Patterns

The tiered taxonomy treats individual error types as distinct categories, but real-world failures often arise from multiple error types occurring simultaneously or sequentially. A hallucination may impair reasoning, leading to factual errors, while the system maintains high apparent confidence and a consistent presentation. These compound failure modes pose a particular risk because individual error types may each fall within acceptable thresholds, yet their combination produces unacceptable outcomes.

Sequential error propagation represents one common pattern. An initial error creates false premises or an incorrect context that propagates through subsequent reasoning. For example, a factual error in equipment specifications might lead to incorrect protective coordination analysis, resulting in flawed recommendations presented with inappropriate confidence. Testing should include scenarios deliberately designed to expose the propagation potential—queries that require multiple analytical steps, where errors at any stage could cascade.

Confidence-error mismatch poses particular danger when high confidence accompanies incorrect outputs. Some systems maintain confident presentation even when hallucinating or making substantial reasoning errors. This combination proves especially problematic because user trust correlates with presentation confidence. Testing should assess whether confidence calibration degrades when multiple error types coincide. Systems that become appropriately uncertain when reasoning chains are tenuous demonstrate safer behavior than those that maintain confidence regardless of output quality.

Consistency-accuracy interaction represents another critical pattern. High consistency across repeated queries can be misleading if the system consistently produces incorrect responses. A system might score well on consistency testing by reliably producing the same wrong answer while failing accuracy testing. Acceptance criteria must avoid treating high consistency as evidence of reliability without verifying that consistent responses are also correct.

Assessment of compound errors requires integrated testing scenarios that probe multiple dimensions simultaneously. Isolating testing of individual error types may miss interaction effects. Testing protocols should include complex, multifaceted scenarios that represent realistic operational queries and enable interactions among various error types. Error correlation analysis should examine whether certain types tend to co-occur, potentially indicating systematic rather than random failures deserving particular attention in risk assessment.

8. Statistical Assessment Framework and Acceptance Criteria

Rigorous AI assessment for critical infrastructure requires statistical frameworks comparable to those applied in safety-related software verification and validation. The testing approaches generate quantitative data that must be analyzed against defined acceptance criteria to support regulatory compliance decisions. This framework must address sample size determination,

confidence interval calculation, error-rate threshold specification, and statistical significance testing, while recognizing that generative AI systems present unique challenges compared to traditional deterministic software.

Sample size determination adapts statistical quality assurance principles to the non-deterministic behavior of AI. Generative systems require larger sample sizes than deterministic software because response variability can require multiple trials even for identical inputs. Sample size calculation should account for the expected error rate, the minimum detectable difference considered significant, and the desired confidence level. For safety-critical applications, confidence levels should match those used in protective relay testing or safety-related software validation, typically 95% or higher.

Error rate thresholds must be established for each category based on safety significance and operational context. Safety-critical errors warrant stringent thresholds analogous to protective relay misoperation rates or safety system failure rates. Factual error rates for safety-critical information might require thresholds below 1 percent, while hallucination tolerance should approach zero for deployment without mandatory external verification. Quality-of-service errors can tolerate higher rates because the consequences are less severe. Compliance assessments may emphasize policy adherence and appropriate behavior patterns rather than pure error counting.

Confidence interval calculation provides essential context for error rate measurements. Point estimates based on finite testing carry uncertainty that must be quantified. Confidence intervals indicate whether measured error rates provide sufficient assurance that acceptance criteria are met. Test coverage adequacy ensures that testing probes the relevant operational space. Coverage metrics should track the proportion of relevant technical topics, equipment categories, operational conditions, and query types included in testing, with requirements graded by tier.

Pass-fail criteria must be clearly specified before testing to avoid retrospective rationalization. For regulatory acceptance, unambiguous criteria prove essential. Some categories admit binary determination—the system either maintains acceptable consistency or does not. Others require threshold-based criteria, in which measured error rates must fall below specified limits with adequate statistical confidence. Criteria should distinguish among conditions that preclude deployment, those that require mitigation measures, and those that support unrestricted deployment for intended use cases.

9. Continuous Monitoring and Periodic Reassessment

Generative AI systems exhibit behaviors that can evolve through intentional updates, gradual operational performance drift, or changing operational contexts that shift the relationship between the deployment environment and the training data. This temporal variability distinguishes AI systems from traditional industrial control software, in which deterministic behavior remains stable unless explicitly modified. Effective quality assurance requires continuous monitoring frameworks and periodic reassessment protocols to detect performance changes and trigger appropriate responses when degradation occurs or the operational context shifts beyond the assessed configuration.

Operational monitoring should track key performance indicators aligned with assessed error categories. Automated logging enables statistical analysis of operational error rates against

acceptance-testing baselines. Significant increases in factual errors, inconsistencies, or other monitored metrics should trigger an investigation and, if warranted, intervention. User feedback mechanisms provide qualitative performance data complementing automated metrics. Operators and engineers can identify problems automated monitoring might miss—subtle misunderstandings, inappropriate responses to unusual scenarios, or degraded utility in specific contexts.

Drift detection methodologies identify gradual performance degradation that may not trigger threshold-based alarms but represents meaningful change relative to validated baselines. Statistical process control techniques can track trends in error rates and flag concerning patterns before reaching unacceptable levels. This is particularly important for systems that continue to learn from operational data, though even static, deployed models may exhibit apparent drift as the operational environment changes relative to the characteristics of the training data.

Model update protocols establish procedures for evaluating and qualifying new versions before operational deployment. Updates may address identified deficiencies, incorporate new capabilities, or reflect improvements to the vendor model. Regardless of the motivation, updates represent configuration changes that may affect all assessed error categories. Update protocols should specify reassessment requirements ranging from abbreviated testing for minor updates to comprehensive reassessment for major version changes, with clear criteria defining what constitutes a minor versus a major change and when full reassessment is mandatory.

Periodic reassessment schedules establish regular intervals for comprehensive retesting independent of identified performance issues or planned updates. The interval should reflect system safety significance, observed stability, and operational context volatility. Safety-critical applications may warrant annual reassessment, whereas less critical applications may accept longer intervals. Periodic reassessment validates that monitored performance metrics accurately reflect overall system quality, identifies issues that operational monitoring may miss, and provides documented evidence of ongoing due diligence for regulatory purposes.

10. Integration with Established Quality Assurance Frameworks

The tiered assessment framework builds upon and extends established quality assurance frameworks familiar to critical infrastructure operators and regulators, including the ISAGCA Bright Line Criteria for deployment architecture boundaries. Rather than requiring entirely novel quality paradigms, the framework maps AI-specific error categories and testing approaches onto proven quality assurance principles from nuclear safety, industrial control systems, and protective relay coordination. This mapping facilitates adoption by enabling organizations to leverage existing quality expertise while addressing unique characteristics of generative AI systems.

The ASME NQA family of standards, particularly NQA-1, provides foundational principles directly applicable to generative AI assessment. The NQA-1 concept of graded quality requirements—where verification rigor scales with safety significance—directly informs the tiered criticality model. Tier 1 assessment aligns with NQA-1 requirements for safety-related software and requires comprehensive verification and validation, supported by formal documentation. Tier 2 assessment aligns with augmented quality requirements for important-to-

safety but not safety-related functions. Tier 3 assessment addresses commercial-grade dedication concepts, ensuring systems meet specified requirements even when not developed under full safety-related quality programs.

Nuclear Regulatory Commission guidance on software quality assurance, particularly Branch Technical Position 7-14 and regulatory guides on digital system qualification, offers precedent for addressing non-deterministic system behavior in safety-critical applications. While traditional software determinism differs fundamentally from generative AI stochasticity, the regulatory frameworks for addressing software common-cause failure potential, diversity requirements, and defense-in-depth principles translate to AI deployment strategies. The concept of verified, diverse actuation systems—where independent methods provide backup when primary systems fail—maps to AI deployment architectures in which generative AI provides analysis support. In contrast, deterministic systems provide the highest level of protection.

NERC reliability standards for critical infrastructure protection and power system protection provide additional framework parallels. NERC CIP standards establish tiered criticality models for cyber assets based on their impact on the reliability of the bulk electric system. This criticality-based approach to security requirements directly parallels the tiered assessment model. Similarly, NERC protection system standards specify verification testing requirements for protective relays, informing generative AI testing protocols. The concepts of settings verification, operational testing, and periodic retesting from protective relay standards map to AI qualification testing, operational monitoring, and periodic reassessment, respectively.

IEEE standards for software verification and validation, particularly IEEE 1012, provide structured approaches to confirmation testing that inform the assessment methodology for generative AI. The IEEE 1012 concept of integrity levels—where verification tasks are selected based on the consequence of failure—parallels tier-based assessment. The standard's emphasis on requirements traceability, test coverage measurement, and acceptance criteria specification all apply to AI testing. However, adaptation is required to address AI-specific characteristics such as non-determinism and the absence of traditional requirements specifications.

The framework integration does not simply rebrand AI assessment with familiar terminology; it establishes genuine connections between established quality principles and AI-specific requirements. Safety significance-based grading, comprehensive verification of safety-critical functions, defense-in-depth through diverse approaches, systematic documentation and configuration management, and lifecycle quality assurance from initial qualification through operational monitoring are proven principles that naturally extend to generative AI deployment. By explicitly connecting AI assessment to these established frameworks, the tiered model provides critical infrastructure organizations with clear guidance on how AI quality assurance fits within existing programs while addressing the technology's unique characteristics.

11. Standards Development Requirements for Tiered Assessment and Bright Line Codification

The tiered assessment methodology and bright-line construct documented in this framework address an emerging gap in the standards landscape: no existing standard comprehensively governs the deployment of generative AI in critical-infrastructure control systems. While

established standards and regulatory frameworks provide foundational principles, they were developed before large language models were considered for integration into safety-critical and reliability-critical environments. Closing this gap requires deliberate action by standards-developing organizations to codify tiered assessment requirements and bright-line criteria into enforceable standards that span sectors and regulatory regimes. The companion document, A Study of AI Risks (Vendor Agnostic), provides essential context for these recommendations, mapping bright-line criteria to sector-specific regulatory bases—Section 215 of the Federal Power Act for the electric sector, Title 10 of the Code of Federal Regulations for nuclear, and corresponding authorities across the remaining critical infrastructure sectors—and documenting the supplier community's consensus that industry self-governance, while necessary, cannot substitute for standards carrying regulatory authority.

Standards-developing organizations occupy distinct roles in the critical infrastructure ecosystem, and the pathway to codification differs accordingly. Some organizations, such as the National Institute of Standards and Technology, develop reference frameworks and special publications that become de facto requirements through regulatory adoption, procurement mandates, and executive direction. Others, such as the North American Electric Reliability Corporation, develop standards with direct regulatory enforceability within their sector. Still others, such as the International Society of Automation, develop consensus standards that are voluntarily adopted or referenced in regulatory instruments. Each pathway serves a different function in establishing the tiered approach as normative practice, and this section addresses the role each should play.

11.1 NIST: Establishing the Cross-Sector Reference Standard

The National Institute of Standards and Technology should incorporate the tiered criticality model and the bright-line construct into its AI standards portfolio as a cross-sector reference for all 16 critical infrastructure sectors defined in the National Infrastructure Protection Plan. The NIST AI Risk Management Framework provides the most natural integration point. Still, the scope of the requirement extends beyond risk management into testing, evaluation, verification, and validation, areas that may warrant a dedicated special publication.

A NIST reference standard should accomplish four objectives.

First, the standard should codify the bright line as a normative requirement keyed to the consequence of the action rather than to the band as such: no generative component may execute an action whose incorrect performance could produce a consequence at or above the facility's severity threshold without a human Command Broker in the control loop. Where an application's band is set by the severity of what its maloperation could produce — the case for the highest-criticality and mid-level-criticality categories as the sector regulators define them — at least one such action necessarily exists and a prohibition on fully autonomous generative operation follows. Where a sector's classification admits an application into a band through loss of service rather than through wrong action, the standard should require the placement analysis rather than infer the prohibition from the band. The standard should define the bright line in terms of consequence severity rather than sector-specific regulatory categories, enabling each sector to map it to its own criticality classification scheme. The nuclear sector maps highest criticality to NRC Safety-Related designations; the electric sector maps it to FERC/NERC CIP-002 High Impact BES Cyber Systems; other sectors map it to their respective highest-consequence

categories. The NIST reference standard provides the unifying principle while accommodating sector-specific implementation.

Second, the standard should establish the three-tier cognitive error taxonomy as the required assessment structure for any generative AI system deployed in critical infrastructure. The safety-critical, quality-of-service, and compliance-driven tiers provide a consistent vocabulary and assessment architecture across sectors. The standard should require that all generative AI deployments above the bright line undergo a comprehensive Tier 1 assessment, and that the assessment rigor scale with application criticality, consistent with the graded quality requirements familiar from NQA-1 and IEC 61508.

Third, the standard should establish minimum statistical acceptance criteria for cognitive error rates in safety-critical applications, providing quantitative benchmarks that sector-specific regulators can adopt or tighten, depending on their regulatory frameworks. The acceptance criteria specified in this framework, including near-zero hallucination tolerance, factual accuracy thresholds exceeding ninety-nine percent for critical safety parameters, and confidence interval requirements aligned with safety-related system verification practices, should serve as the NIST reference baseline.

Fourth, the standard should require continuous monitoring and periodic reassessment as lifecycle quality assurance obligations, establishing the principle that initial qualification alone is insufficient for generative AI systems exhibiting non-deterministic behavior. The NIST reference should specify that monitoring intensity and reassessment frequency scale with application criticality, and that model updates, or significant changes in the operational context, trigger a formal reassessment against the tiered criteria.

11.2 NERC: Codifying the Bright Line in Reliability Standards

The North American Electric Reliability Corporation occupies a distinctive position as both a regulator with FERC-delegated enforcement authority and a standards-developing organization whose products carry the force of federal law upon FERC approval. NERC should develop a reliability standard, or a set of requirements within existing CIP standards, that mandates the use of a bright-line and tiered cognitive error assessment for generative AI deployed in Bulk Electric System operations.

NERC's existing CIP-002 categorization of BES Cyber Systems as High, Medium, or Low Impact provides a ready-made mapping for the bright line, and the mapping holds where the classification rests on the consequence of maloperation. A NERC standard should require that no generative AI system operate autonomously in any High or Medium Impact BES Cyber System whose impact rating rests on what its incorrect operation could cause — the cascading-outage and wide-area-instability cases the categorization is built around. Where a system reaches its rating on service reliability or availability rather than on maloperation consequence, the standard should require the placement analysis action by action rather than take the prohibition from the rating. Any generative AI that provides decision support, analytical capabilities, or operational guidance within High or Medium Impact environments must operate within a Command Broker architecture for its above-line actions, where qualified personnel retain full decision authority and execute those actions through verified conventional control interfaces. This requirement codifies, within an enforceable reliability standard, the principle that the indeterministic nature of

generative AI precludes autonomous operation in systems whose maloperation could cause cascading outages, wide-area instability, or significant reliability events.

The NERC standard should further require that any generative AI deployed within BES Cyber Systems undergo cognitive error assessment consistent with the tiered methodology documented in this framework, or a methodology of equivalent rigor accepted by the applicable Regional Entity. Responsible Entities would be required to maintain provenance documentation, demonstrate compliance with tier-appropriate acceptance criteria, implement continuous monitoring, and conduct periodic reassessments at intervals scaled to the criticality classification. This requirement structure aligns with existing CIP compliance obligations and integrates seamlessly into NERC's compliance monitoring and enforcement program.

NERC's standards development process, which involves industry-representative drafting teams, successive comment and ballot periods, and FERC regulatory review, ensures that the resulting standard reflects operational realities, supplier capabilities, and regulatory objectives. The process also provides the formal notice-and-comment opportunity that gives the resulting standard legal defensibility. NERC should initiate a Standards Authorization Request to begin this process, drawing on the framework's technical content and the regulatory context established in the companion AI Risks study.

11.3 ISA and IEC: Extending the Bright Line to Global Industrial Standards

The International Society of Automation and the International Electrotechnical Commission should incorporate the bright-line and tiered assessment into the ISA/IEC 62443 series for industrial automation and control system security, as well as into the IEC 61508 functional safety framework. The ISAGCA Supplier Working Group's bright-line criteria, which inform this framework, originated within the ISA ecosystem and represent supplier-community consensus on the boundaries of generative AI deployment. Formalizing these criteria within ISA/IEC standards would extend their reach beyond the supplier working group to the full global community of ICS practitioners, integrators, and regulators who reference 62443 and 61508.

Within 62443, the tiered approach maps to the existing Security Level concept, where security requirements scale with the severity of consequences. A technical report or standard addition addressing AI-specific requirements could establish that assessment rigour for generative AI scales with the target Security Level, while autonomy remains governed by the placement of each action on its consequence. Security Level is a threat-capability axis, and an autonomy prohibition keyed to it would inherit the axis problem this framework exists to name: the Security Level should set how demanding the assessment is, not whether an action may proceed without a broker. Within 61508, the Safety Integrity Level framework provides a closer mapping, because a safety function is defined by what its incorrect performance could cause: generative AI should be prohibited from autonomous safety functions at SIL 3 and SIL 4, and cognitive error assessment requirements should scale with the application's SIL.

11.4 NRC: Regulatory Guidance for Nuclear Sector Applications

The Nuclear Regulatory Commission, while not a voluntary standards-developing organization, issues regulatory guides and branch technical positions that function as de facto standards for the nuclear industry. NRC should update its digital instrumentation and controls guidance to address

generative AI, incorporating the bright line principle that no generative AI system shall operate autonomously in any Safety-Related application or Important-to-Safety system. These classifications rest on what incorrect performance could cause, so the prohibition follows from the placement of the actions those systems perform and not from the classification label. The existing regulatory framework for software quality assurance in nuclear applications, including Branch Technical Position 7-14, Regulatory Guides 1.152, 1.168, 1.169, and 1.173, and the NQA-1 quality assurance requirements referenced therein, provides the foundation for AI-specific guidance.

The NRC's licensing and compliance infrastructure offers a model for how other sector regulators might adopt the tiered framework. The existing graded approach to quality assurance, where Safety-Related systems receive the most rigorous verification and Balance-of-Plant systems receive proportionally less, directly parallels the tiered assessment structure. Applicants seeking to deploy generative AI in nuclear facilities must demonstrate compliance with tiered cognitive error assessment criteria as part of their licensing basis, with the rigor of the assessment scaled to the safety classification of the supported system.

11.5 The Imperative for Coordinated SDO Action

The urgency of coordinated standards development reflects the pace of AI deployment relative to the pace of standards processes. Generative AI products are being marketed to critical infrastructure operators today, and deployment decisions are being made in an environment where no sector has enforceable standards governing AI autonomy in control systems. The AI Risks study documents that suppliers themselves recognize the need for a bright line and are developing criteria through the ISAGCA working group. Still, supplier self-governance, however responsible, cannot substitute for standards with regulatory backing.

The consequences of inaction are predictable from historical precedent. When transformative technologies are deployed in critical infrastructure without adequate standards frameworks, the results include inconsistent implementation, preventable incidents, and reactive regulatory responses that impose greater costs and disruption than proactive standards development would have required. The transition from analog to digital control systems in the nuclear and electric sectors during the 1990s and 2000s demonstrated both the difficulty of developing standards for new technology and the necessity of doing so before, rather than after, widespread deployment.

Coordinated SDO action should follow a deliberate sequence. NIST should first establish a cross-sector reference standard to provide a common foundation. Sector-specific SDOs and regulators should then develop implementation requirements that map the NIST reference to their particular regulatory bases, criticality classifications, and compliance structures. This layered approach—a reference standard providing unifying principles, with sector-specific standards providing implementation detail—mirrors the successful model established by the NIST Cybersecurity Framework and its sector-specific adaptations.

Standards-developing organizations should also coordinate on the distinction between generative AI and deterministic machine learning that both this framework and the AI Risks study emphasize. Standards must be precise about their scope: the bright-line and cognitive-error assessment methodologies apply to generative AI systems based on large language models and similar indeterministic architectures. Deterministic machine learning systems designed for

closed-loop control remain amenable to traditional verification and validation methods and are not subject to the bright line prohibition on autonomous operation. Imprecise standards that conflate generative AI with all forms of artificial intelligence would unnecessarily constrain beneficial ML applications and may fail to address the specific risks of generative systems.

12. Conclusion and Implementation Guidance

This tiered assessment framework establishes a comprehensive yet practical approach to evaluating generative AI systems for deployment in critical infrastructure. By organizing cognitive error types into safety-critical, quality-of-service, and compliance-driven tiers, the framework enables proportional allocation of assessment resources while maintaining coverage across all relevant risk dimensions. Integration with ISAGCA Bright Line Criteria ensures that testing rigor is applied within appropriate deployment architecture boundaries determined by application criticality. The structure balances rigor with practicality, recognizing that perfect AI systems are unattainable while establishing clear performance criteria for different risk contexts.

The framework's value lies not merely in its technical content but in how it connects AI-specific challenges to established quality assurance principles that critical infrastructure organizations already understand and practice. Organizations implementing this framework can leverage existing verification and validation expertise, quality assurance programs, and regulatory compliance processes rather than building entirely new capabilities. The explicit connections to NQA standards, NRC guidance, NERC reliability standards, and IEEE verification standards provide familiar touchpoints for professionals accustomed to rigorous quality requirements in safety-critical domains.

Implementation should proceed systematically, beginning with thorough provenance documentation that establishes exactly which system is being assessed. Organizations should then determine which error categories are most relevant to their specific deployment context and use case. Not every application requires comprehensive testing across all categories—the tiered structure explicitly supports selective application based on risk significance. However, all safety-critical error categories warrant serious consideration for any deployment where AI outputs could influence safety-related decisions or protective actions.

The statistical framework, continuous monitoring provisions, and periodic reassessment requirements address the temporal dynamics of AI systems in operational deployment. Initial qualification represents only the beginning of quality assurance—ongoing monitoring and periodic revalidation ensure that performance remains within acceptable bounds as systems evolve and operational contexts change. Organizations should establish these lifecycle quality assurance processes before initial deployment rather than attempting to add them retroactively.

As generative AI systems transition from experimental applications to operational roles in critical infrastructure, frameworks such as this become essential for managing the unique risks these technologies present while enabling their beneficial capabilities. The standards development requirements articulated in Section 11 establish a pathway to convert this framework's technical content into enforceable requirements across critical infrastructure sectors, beginning with NIST cross-sector reference standards and extending to sector-specific regulatory instruments. The approach documented here provides a foundation for responsible AI

deployment that protects the integrity of critical infrastructure while supporting technological advancement. Organizations implementing this framework contribute not only to their own operational excellence but to the development of industry-wide best practices for AI deployment in safety-critical environments.

The framework remains a living document, expected to evolve as experience with AI deployment in critical infrastructure accumulates and the technology advances. Feedback from practitioners, regulators, and researchers will inform future refinements. Organizations implementing the framework are encouraged to document their experiences, share lessons learned through appropriate channels, and contribute to the collective understanding of effective quality assurance for AI in critical infrastructure applications.