Governance is a Runtime Dependency: An Evidence Map and Operational AI Readiness Matrix for Disaster Management

Sandip Patel
Independent Researcher Frisco, Texas USA
ORCID: 0009-0003-5229-6181

Deependra Singh Rawat
Independent Researcher
Highlands Ranch, Colorado, USA
ORCID: 0009-0005-3943-963X

Bhavin Gandecha
Independent Researcher
Monroe Twp, New Jersey, USA
ORCID: 0009-0001-0728-8305

ABSTRACT
Disaster-management AI is commonly governed before deployment through ethics reviews, data-sharing agreements, and standards checklists. This paper argues that design-time governance is necessary but incomplete because decisive failures occur while systems run: a prediction must be overridden, a model rolled back, a decision audited, an incident escalated, or a system kept safe when connectivity degrades. We reframe governance as a runtime dependency and synthesize a purposively screened 80-record set of standards, policy documents, surveys, primary studies, and technical sources. Ten runtime controls are coded through 232 non-exclusive source-control assignments. Coverage averages 23.2 sources per control (median 21.5; range 6 to 43); the five most-covered controls account for 80.2% of assignments, whereas the five least-covered controls account for 19.8%. A six-family standards crosswalk contains support in 31 of 60 cells (51.7%), with each control represented by two to four instrument families and no family covering all controls. Auditability, co-production, and human override are prominent, while ownership, rollback, incident reporting, and degraded-mode operation remain comparatively sparse. The contribution is an Operational AI Readiness Matrix with pass/fail questions, evidence requirements, and four maturity levels. An analytical evaluation checks completeness, traceability, maturity discrimination, non-compensation, and scenario coverage, while explicitly recognizing that the matrix has not yet been validated against deployment outcomes. The result is a practical instrument for testing whether disaster-AI systems remain governable during drills and live events, not merely documented before use.

General Terms
Artificial Intelligence, Disaster Management, Responsible AI, Information Systems, Risk Management, Human Oversight.

Keywords
Runtime governance; disaster AI; AI readiness matrix; auditability; human override; model rollback; degraded-mode operation; standards crosswalk.

1. INTRODUCTION
When an AI system for disaster management is procured, governance is often treated as a design-time activity. An ethics review is completed, a data-sharing agreement is signed, a fairness review may be commissioned, and a checklist may be marked complete. These actions matter, but they do not guarantee that the system can be governed during an emergency.

The operational moments that most require governance occur after deployment. A flood model may produce an implausible forecast while a duty officer has minutes to decide whether to override it. A damage-assessment model may silently degrade after a novel hazard and require rollback to a known-good version. A cloud-dependent component may become unreachable when local connectivity fails. A biased alert or missed warning may need to be reported and investigated before the next event. These situations are runtime governance events.

This paper therefore reframes governance as a runtime dependency: a set of controls that must function continuously while the disaster-AI system is operating. The reframing changes what readiness means. A deployment is not ready because governance was documented; it is ready when governance controls have been exercised and can produce evidence under stress.

The research question is: which runtime governance controls are specified or evidenced by existing standards and policy, where are the crosswalk gaps, and how can those controls be operationalized into an assessment instrument for disaster-AI systems? The answer is presented as an evidence map, a standards crosswalk, and an Operational AI Readiness Matrix.

2. RELATED WORK AND POSITIONING
Operational constraints are visible in work on on-device inference [1], AI-incident schemas [2], disaster data systems [3], multilingual safety [4], emergency-management uses of AI [5], Indigenous data sovereignty [6, 7], legal perspectives on disaster risk reduction [8], the political economy of disasters [9], and warning-to-action failures [10]. Together these sources show that readiness depends on technical operation, institutional authority, affected communities, and field conditions.

Formal governance and risk literature includes the EU AI regulatory framework and its human-oversight provision [11, 12], responsible-AI guidance for disaster risk management [13], public-sector accountability frameworks [14], taxonomies of AI incidents [15], disaster misinformation research [16], early-warning failure analysis [17], and international standards work for responsible AI in disaster management [18]. These sources define obligations, but they do not consistently specify how a response team demonstrates that a control works during an event.

Interoperability and continuity are addressed through OGC readiness guidance [19], crisis-resilience roadmaps [20], edge-cloud surveys [21], public-health risk communication and community engagement [22], humanitarian case studies [23],

multilingual edge models [24], and distributed disaster-service architectures [25]. This stream makes connectivity, partner exchange, local execution, and language coverage part of operational readiness rather than secondary implementation details.

General AI risk controls appear in the NIST AI Risk Management Framework and its generative-AI profile [26, 27], OGC disaster-interoperability work [28], human-in-the-loop research [29], risk-mitigation taxonomies [30], technical disaster applications such as FloodNet and climate-risk early warning [31, 32], analysis of data cascades [33], transdisciplinary interoperability and co-production [34], and responsible-AI working-group guidance [35]. These contributions motivate controls for monitoring, intervention, evidence, and cross-organizational coordination.

Participatory and humanitarian-AI work emphasizes power, engagement, and field learning [36, 37]. Emergency-interoperability guidance [38], hazard and loss tracking initiatives [39, 40], rights-based AI policy [41], and multimodal damage-assessment research [42] further show that disaster-AI governance spans data exchange, accountability, equity, and model operation. The literature is therefore rich in principles and domain requirements but fragmented across standards communities.

The gap addressed here is not the absence of governance principles. It is their operational expression at runtime. A policy may require human oversight, but a response team still needs to know who is on call, whether the override path works inside the decision window, whether the action is logged, and whether the procedure was tested. A model-change policy may exist, but the operational question is whether a known-good version can be restored during an incident without waiting for a full rebuild.

This paper contributes by translating distributed governance obligations into runtime controls, readiness questions, maturity levels, and evidence requirements. It does not propose a new prediction model, scheduler, or performance benchmark. Its novelty is an integrated assessment instrument for AI-enabled disaster systems whose harms arise from interactions among models, data, operators, communities, partner agencies, and infrastructure.

3. METHOD

3.1 Corpus Construction

This work is a structured scoping evidence review and framework synthesis, not a completed PRISMA systematic review. A broad topical search identified candidate literature across disaster AI, governance, standards, interoperability, human oversight, incident reporting, multilingual access, data sovereignty, and resilient infrastructure. A targeted strengthening pass then focused on instrument families and operational failure patterns relevant to the ten runtime controls.

The coding set contained 80 purposively screened records. The 42 references listed in this article are the sources cited directly in the narrative; the full source-by-control codebook is not reproduced in the article. Table 1 must therefore be read as a descriptive summary of the 80-record coding set rather than reconstructed from the narrative reference list alone. Inclusion prioritized standards or policy instruments and sources that directly addressed at least one runtime governance control.

3.2 Coding Approach and Validity Boundary

Each source was coded against ownership, data sovereignty, interoperability, auditability, human override, co-production, model update and rollback, incident reporting, multilingual access, and degraded-mode operation. Coding was non-exclusive: one source could support multiple controls. The 80 records therefore produced 232 source-control assignments, and counts across Table 1 do not sum to 80.

Coding was exploratory and performed by one coder. No independent recoding or inter-rater reliability statistic was produced. Counts are descriptive of this screened set and should not be interpreted as field prevalence. Table 2 is likewise the authors' coding of support found in the screened evidence, not an official conformance assessment by OGC, NIST, ITU, UNDRR/WMO, WHO, or the European Union.

4. EVIDENCE MAP

Table 1 reports coded coverage of runtime controls in the screened set. A high count means that multiple sources touched a control. A low count means the control was sparsely evidenced in this screen, not that the control is unimportant in practice.

Table 1. Corpus-evidenced coverage of runtime controls in the screened set (N=80).

Runtime control	Sources	Share
Auditability	43	54%
Co-production	40	50%
Human override	38	48%
Interoperability	35	44%
Multilingual access	30	38%
Degraded-mode operation	13	16%
Data sovereignty	11	14%
Incident reporting	9	11%
Ownership	7	9%
Model update / rollback	6	8%

4.1 Quantitative Evidence Analysis

The ten controls contain 232 non-exclusive source-control assignments. Mean coverage is 23.2 sources, the median is 21.5, and the range extends from 6 for model update and rollback to 43 for auditability. The range of 37 and the gap between multilingual access (30) and degraded-mode operation (13) show a pronounced two-tier distribution within this coding set.

The three most-covered controls account for 121 of 232 assignments (52.2%). The top five account for 186 (80.2%), leaving 46 assignments (19.8%) across the bottom five. This concentration indicates that the literature more often discusses visibility, participation, intervention, exchange, and language than the operational ability to assign stop authority, report incidents, enforce data restrictions, restore a known-good model, or continue safely under degraded infrastructure.

These statistics describe attention in the screened corpus, not control importance or implementation quality. Adjacent ranks within each tier could change under independent recoding, especially among the low-count controls. The larger separation

between the two tiers is more stable as a descriptive pattern, but it also requires confirmation with a published codebook and multiple coders.

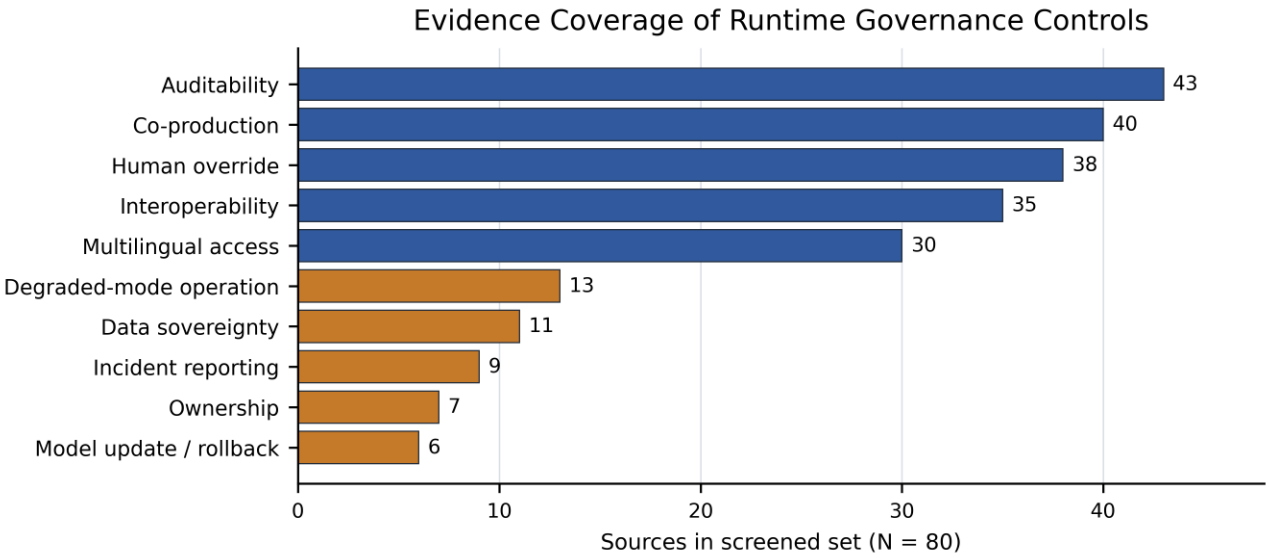

Figure 1. Evidence coverage is concentrated in five controls. Colors distinguish the upper and lower halves of the ranked distribution; sources could be coded to multiple controls.

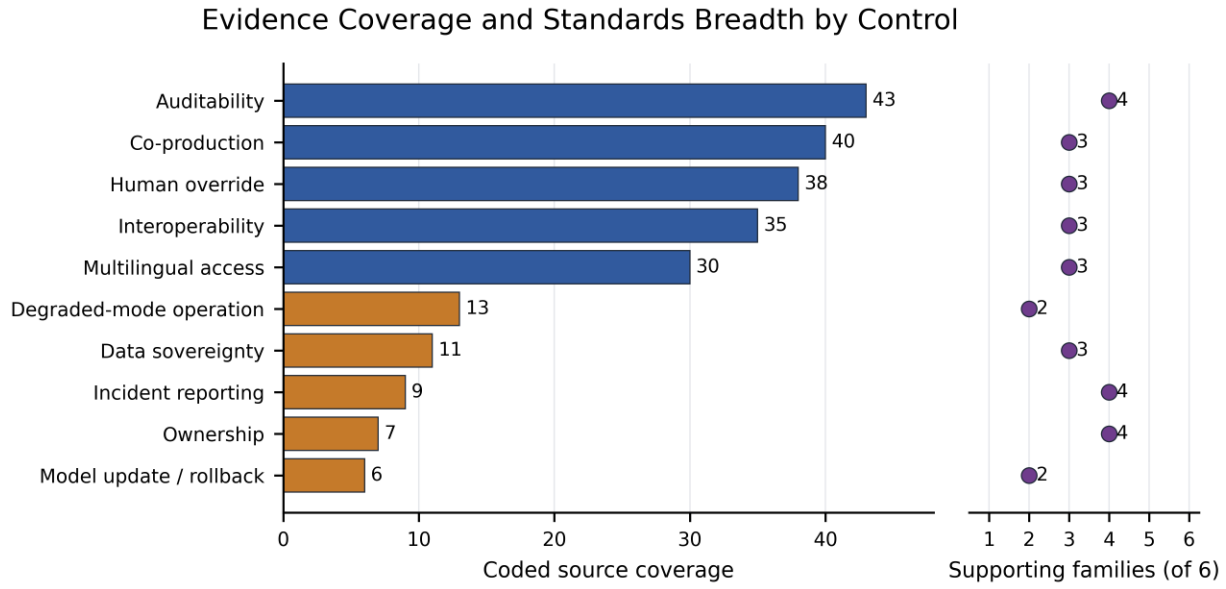

Figure 2. Evidence coverage and standards breadth are shown for each control. The left panel reports coded source coverage and the right panel reports supporting instrument families; the rank association is weak (Spearman rho = 0.15, n = 10) and no inferential claim is made.

The evidence map shows a practical imbalance. Auditability, co-production, human override, interoperability, and multilingual access are comparatively visible. Model update and rollback, ownership, incident reporting, data sovereignty, and degraded-mode operation are less evidenced even though they determine whether a system can be stopped, corrected, explained, or kept safe during an emergency. Incident taxonomies and mitigation studies reinforce the need to connect observed failures to operational correction paths [2, 15, 30].

Coverage and specification are different. A source may mention a control without specifying how a team executes it, while a standard may specify one narrow obligation but omit adjacent disaster conditions. The next crosswalk therefore evaluates the breadth of coded support across instrument families, and Figure 2 compares that breadth with narrative evidence volume.

5. STANDARDS CROSSWALK

Table 2 compacts the authors' crosswalk across six instrument families. A check mark indicates that support for the control was found in the screened evidence. Blank cells indicate no support found in this screen; neither checks nor blanks are official conformance determinations. The ITU and WHO columns are particularly dependent on secondary standards and

public-health sources in the screened set because a primary instrument-by-instrument audit was outside the review scope.

Table 2. Author-coded standards crosswalk for runtime governance controls.

Control	OGC	NIST	ITU	UNDRR/WMO	WHO	EU AI Act
Ownership		✓		✓	✓	✓
Data sovereignty		✓		✓		✓
Interoperability	✓		✓	✓		
Auditability		✓	✓		✓	✓
Human override		✓			✓	✓
Co-production	✓			✓	✓	
Model update / rollback		✓				✓
Incident reporting		✓		✓	✓	✓
Multilingual access			✓	✓	✓	
Degraded-mode operation	✓		✓			

5.1 Crosswalk Breadth Analysis

The crosswalk contains 31 supported cells out of 60 possible control-family combinations (51.7%). Each control is supported by two to four families, with a mean breadth of 3.1 and median of 3.0. Ownership, auditability, and incident reporting each appear in four families; rollback and degraded-mode operation appear in only two. No control is supported by all six families, which confirms that composition is required.

By column, the coded support totals are OGC 3, NIST 6, ITU 4, UNDRR/WMO 6, WHO 6, and EU AI Act 6. These totals measure breadth in the authors' screen, not legal force, technical depth, or implementation quality. For example, a legally binding obligation and a voluntary risk-management recommendation both count as one supported cell, so the crosswalk should guide source selection rather than certify compliance.

Across the ten controls, evidence count and instrument-family breadth have a weak descriptive rank association (Spearman rho = 0.15, n = 10). With only ten controls and exploratory coding, no p-value or population inference is warranted. The practical result is that narrative attention and standards coverage should be reviewed separately: ownership and incident reporting have low corpus counts but breadth across four families, while rollback and degraded mode are weak on both dimensions.

The crosswalk suggests three findings. First, no single instrument family covers all ten controls. NIST and EU AI Act sources provide strong anchors for accountability, human oversight, monitoring, record keeping, and model-change governance, but they do not by themselves address all disaster-specific interoperability and degraded-mode concerns. [11, 12, 26, 27]

Second, interoperability and degraded-mode operation are mainly anchored in geospatial and telecommunications traditions. This matters because disaster-AI teams that use only an AI-governance framework may miss continuity, partner data exchange, and offline-operation requirements. [18, 19, 21, 25, 28, 34, 38]

Third, equity-oriented controls such as co-production, multilingual access, and data sovereignty are often grounded in humanitarian, public-health, risk-reduction, and Indigenous-data-sovereignty work. These bodies of work are sometimes treated as policy-adjacent rather than technical, but runtime readiness depends on them. [4, 6, 7, 22, 24, 35-37, 41]

6. OPERATIONAL AI READINESS MATRIX

The Operational AI Readiness Matrix converts the evidence map and crosswalk into an assessment instrument. The matrix is designed for procurement reviews, tabletop exercises, technical drills, and post-event lessons learned. Each control has a runtime requirement and a pass/fail readiness question.

Table 3. Operational AI Readiness Matrix for disaster-AI systems.

Control	Runtime requirement	Readiness question
Ownership	Named role accountable for live decisions and stop authority.	Is there an on-call owner who can halt the system now?
Data sovereignty	Runtime use respects subject and community restrictions, including Indigenous data rights.	Can a data-use restriction be honored during operation?
Interoperability	System exchanges data or model outputs with partner agencies during an event.	Can standards-based data be exchanged mid-event?
Auditability	Consequential decisions are logged with enough context to reconstruct them.	Can any output be reconstructed after the fact?
Human override	Operator can countermand or halt the system within the decision deadline.	Can a human override a decision in time?
Co-production	Responders and affected communities can shape operation, not only design.	Is there a live correction channel during use?
Model update / rollback	Models can be updated or reverted safely during operation.	Can a known-good model be restored within minutes?
Incident reporting	Failures are reported, escalated, reviewed, and fed back into improvement.	Did the last exercise produce a logged incident report?
Multilingual access	System operates for the required languages and dialects of affected populations.	Are top local languages supported at required quality?
Degraded-mode operation	Safe behavior is defined when connectivity, compute, or data degrade.	Has degraded mode been tested without cloud dependence?

6.1 Maturity Levels

L0 - Absent: the control is not addressed and the runtime dependency is unmanaged.

L1 - Documented: the control exists in a policy, role definition, plan, or design artifact, but has not been exercised.

L2 - Exercised: the control has been tested in a drill or exercise and produced evidence, such as an override log, rollback record, restored model version, or incident report.

L3 - Operational: the control has functioned during a real event, and the resulting evidence has been reviewed and fed back into improvement.

The minimum readiness target for life-safety deployment should be L2. A documented but untested rollback plan, incident channel, or degraded mode should not be treated as ready merely because it exists in design material.

7. CONTROL-BY-CONTROL SYNTHESIS

Ownership

Ownership is foundational because every other runtime control requires an accountable actor. Without a named owner who can halt or restrict the system during a live event, override, incident reporting, and rollback procedures become ambiguous. The readiness question intentionally asks whether the owner can act now, not whether an owner appears in a governance plan. [14, 26]

Model update and rollback

Rollback is the least-evidenced control in the screened set but one of the most operationally important. Disaster conditions can differ sharply from training and validation conditions; a recent update may therefore be unsafe during an unusual event. Runtime readiness requires a known-good model, a tested rollback path, and evidence that the rollback can be executed within the response window. [1, 21, 25, 30]

Incident reporting

Incident reporting turns failures into learning. A policy that says incidents should be reported is not enough. The operational test is whether a channel exists, whether responders know how to use it, whether reports are logged, and whether reports trigger review and remediation. [2, 15, 30]

Degraded-mode operation

Disasters often degrade the infrastructure that AI systems assume. Connectivity, cloud access, compute, and fresh data may become unavailable exactly when the system is needed. A runtime-ready system has a safe degraded mode, including reduced functionality, cached data, partner data exchange, or manual fallback. [1, 21, 25, 38]

Auditability

Auditability is comparatively well evidenced, but the runtime question is concrete: can the team reconstruct why a specific output was produced, what data and model version were used, who saw the output, and what action followed? [14, 26, 27]

Human override

Human override is meaningful only if it can occur inside the decision window. The mechanism must be reachable, understood by operators, and logged so that overrides become part of the accountability trail. [12, 29]

Co-production

Co-production connects technical operation with field reality. Responders and communities can surface errors, language gaps, local context, missing data, and misinformation that centralized design teams may not observe. [22, 34-37]

Multilingual access and data sovereignty

Equity controls are runtime controls. A system that cannot support the languages of the affected population or honor data-use restrictions during an event may fail the people it is intended to serve, even if its model performance is strong in benchmark conditions. [4, 6, 7, 24, 41]

8. IMPLEMENTATION WORKFLOW

A practical readiness assessment can be run in four steps. First, identify the disaster-AI use case, supported decisions, affected populations, partner agencies, infrastructure assumptions, and data flows. Second, score each of the ten controls at L0 through L3 using the readiness questions. Third, require evidence for each claimed level, such as drill records, logs, screenshots, role assignments, model registry entries, incident tickets, multilingual quality checks, or partner exchange tests. Fourth, remediate the weakest runtime dependencies before deployment or before expanding scope.

The output should be a profile rather than a single score. Averaging controls would allow strong auditability to hide absent rollback or absent degraded mode. The binding risk is often the weakest control. For example, if ownership is L0, then override, rollback, and incident reporting cannot be fully trusted even if written procedures exist. If auditability is L1, incident reporting cannot reliably reconstruct failures. If interoperability is weak, degraded-mode operation may be unrealistic.

Procurement teams can use the matrix as a requirement specification. Emergency-management teams can use it during exercises. Technical teams can use it to connect governance requirements with model registries, logging, monitoring, access controls, rollback pipelines, and offline operation. Policy teams can use it to identify which standards families must be composed for a complete governance posture.

9. EVIDENCE REQUIREMENTS FOR ASSESSMENT

The readiness matrix is most useful when each maturity claim is tied to evidence. Table 4 gives examples of evidence that an assessor may request. These examples are intentionally operational: they ask for artifacts that prove the control has been used, not simply artifacts that describe an intended process.

Table 4. Example evidence artifacts for readiness assessment.

Control group	Evidence examples	Assessment concern
Ownership and authority	On-call roster; named decision owner; stop-authority record; escalation tree tested in an exercise.	Can a person with authority act during the response window?
Auditability and incident reporting	Decision logs; model version records; incident tickets; post-exercise review notes; corrective-action tracker.	Can failures be reconstructed, escalated, and improved?
Human override and rollback	Override log; rollback drill transcript; model registry entry;	Can the system be corrected quickly without a full rebuild?

	deployment pipeline record; known-good model package.	
Interoperability and degraded mode	Partner exchange test; offline or reduced-mode runbook; cached data validation; network-failure exercise evidence.	Can the system continue safely when infrastructure degrades?
Equity controls	Language coverage test; community feedback log; data-use restriction test; responder correction channel output.	Can affected populations influence and access the system during use?

Evidence requirements also reduce ambiguity between L1 and L2. A policy document may justify L1, but it does not justify L2. L2 requires proof that the control was exercised. For instance, rollback is L2 only when a team can show that a model was actually reverted in a drill and that the result was verified. Incident reporting is L2 only when the channel produced a logged report. Degraded mode is L2 only when degraded behavior was tested under a realistic constraint, such as missing connectivity or stale data.

This evidence-first interpretation is important because disaster decisions are time-sensitive. During an event, teams do not have time to discover that a runbook is outdated, that the owner is unavailable, that the rollback package is incomplete, or that multilingual messages cannot be generated with acceptable quality. The assessment should therefore force proof before deployment, not confidence during deployment.

The matrix can be implemented as a lightweight review packet. Each control receives a maturity level, a short rationale, a link or reference to evidence, and a remediation owner. The review packet should be revisited after exercises and real events because runtime readiness is not static. Model changes, partner changes, staffing changes, language needs, and infrastructure dependencies can all lower readiness if the controls are not re-tested.

10. PROCUREMENT AND GOVERNANCE USE CASES

The matrix can be applied at acquisition time by converting the ten controls into vendor or project requirements. A procurement team can ask vendors to provide evidence of exercised override, rollback, audit, incident reporting, interoperability, and degraded-mode capabilities. This avoids a common pattern in which vendors describe governance principles but provide little proof that the controls operate under realistic disaster conditions.

For internal governance teams, the matrix can serve as a bridge between responsible-AI policy and engineering evidence. Many organizations already maintain risk registers, model inventories, security reviews, and incident-management processes. The runtime-readiness view links those artifacts to disaster-specific operating questions: who can stop the system, which model version is safe, which partner data formats are supported, which local languages are required, and how the system behaves when cloud access is unavailable.

For emergency-management teams, the matrix supports tabletop and functional exercises. Exercise designers can inject failures that test each control: an implausible forecast to test override, a bad model update to test rollback, missing connectivity to test degraded mode, a partner-agency data request to test interoperability, and a language-access gap to test multilingual operation. The output of the exercise becomes readiness evidence.

Table 5. Example assessment prompts for exercise design.

Exercise inject	Control tested	Expected evidence
Model produces a high-confidence but locally implausible alert.	Human override; ownership; auditability.	Override action, owner decision, time-to-override, and audit log.
New model version degrades on a novel hazard scenario.	Model update / rollback; incident reporting.	Rollback record, known-good version restoration, and incident ticket.
Network access to a cloud service is unavailable.	Degraded-mode operation; interoperability.	Reduced-mode run, partner data exchange, and documented safe limits.
Affected area requires languages not covered by default messages.	Multilingual access; co-production.	Language test output, responder feedback, and corrected communication path.
Community data-use restriction changes during response.	Data sovereignty; ownership.	Runtime restriction enforcement and accountable approval record.

11. DEPLOYMENT SCENARIOS AND CONTROL INTERACTIONS

11.1 Early Warning and Human Override

In an early-warning setting, the system may combine weather data, sensor feeds, geospatial risk layers, and historical event patterns to generate alerts. The runtime-governance risk is not only that the model may be wrong, but that the organization may not be able to act responsibly when the model appears wrong. A warning can be over-triggered, under-triggered, delayed, or routed to the wrong audience. Each case requires an accountable owner, a human-override path, audit logs, and an incident channel. [10, 17, 32]

The matrix treats the override as a tested operational dependency. An operator should know where the override mechanism is, what authority is required, how the override is documented, whether the override changes downstream notifications, and how the action will be reviewed after the event. If any of these elements are missing, the system may have human oversight in principle but not in operation.

The interaction between auditability and override is especially important. An override without auditability can become arbitrary. Auditability without override can become a historical

record of a preventable failure. Runtime governance requires both: the ability to intervene and the ability to explain why the intervention occurred.

11.2 Damage Assessment and Model Rollback

Damage-assessment systems often use imagery, field reports, social-media signals, or infrastructure data to estimate severity and allocate response resources. These systems are exposed to distribution shift because each disaster has local features: construction patterns, landscape, weather, sensor availability, image quality, and reporting behavior may differ from the conditions used during model development. [31, 42]

In this scenario, rollback is a governance control, not only a software-release convenience. If a newly updated model performs poorly during a live response, the team needs a known-good version, a versioned model artifact, a tested deployment path, and a decision owner who can approve restoration. Without those elements, the system may continue to influence allocation even after field responders identify its outputs as unreliable.

Incident reporting also interacts with rollback. A bad update should generate a report that records the model version, input context, observed failure, corrective action, and review outcome. This report should feed into future validation and procurement requirements. The point is not to punish model failure; the point is to make failure visible and correctable.

11.3 Connectivity Failure and Degraded Mode

Connectivity failure is a defining condition of many disasters. Roads, cell towers, power, cloud reaches, and data feeds may degrade or become intermittent. An AI system that assumes stable cloud connectivity may be unavailable at precisely the moment when faster situational awareness is needed. Degraded-mode operation therefore becomes a safety control.

A degraded mode should define what the system will do, what it will not do, and how users will know the difference. For example, the system may shift to cached geospatial layers, delayed synchronization, a smaller local model, manual data-entry mode, or a read-only decision-support mode. The most important governance feature is explicit limitation: responders need to know when the system is operating with stale data, reduced confidence, or unavailable partner feeds.

Interoperability supports degraded mode because partner agencies may become the fallback source of data or validation. If the system cannot exchange data cleanly before a disaster, it is unlikely to become interoperable during one. The readiness matrix therefore asks for prior evidence of partner exchange and degraded-mode testing.

11.4 Equity, Language, and Community Correction

Equity failures in disaster AI may appear at runtime even when a model appears technically accurate. A warning may not be available in the languages used by affected populations. Input data may underrepresent informal settlements, remote communities, people with limited connectivity, or groups with lower visibility in official data. A community may restrict how its data is used during an event. These are operational conditions, not abstract policy concerns. [4, 6, 7, 24, 33, 36, 41]

The matrix therefore treats multilingual access, data sovereignty, and co-production as runtime controls. Multilingual access requires more than translation at design time; it requires operational quality checks for the languages and dialects needed in the affected area. Data sovereignty requires the ability to honor use restrictions while data is flowing, not only when the system is configured. Co-production requires a live channel through which responders and affected communities can correct system outputs and surface missing context.

These controls also protect technical quality. Community correction can identify data cascades, stale assumptions, mislabeled locations, or misleading automated classifications. Language access can determine whether an alert is actionable. Data-sovereignty controls can determine whether the system retains legitimacy with communities whose data is being used.

12. DESIGN PRINCIPLES FOR THE READINESS MATRIX

The matrix follows five design principles. First, it is evidence-based: each maturity claim should be supported by an artifact from policy, technical systems, exercises, or real events. Second, it is runtime-oriented: the questions are written around actions that must function during operation. Third, it is control-specific: readiness is not collapsed into a single score because one missing control can dominate risk. Fourth, it is standards-composable: different control families may require different standards communities. Fifth, it is equity-integrated: language, data rights, and community correction are built into the same matrix as auditability and rollback.

The matrix is deliberately simple because disaster response is complex. A readiness instrument that requires excessive interpretation may not be used during procurement or exercises. The pass/fail question gives assessors a concrete starting point. The maturity level adds nuance without turning the instrument into a heavyweight certification scheme. The evidence requirement prevents self-attestation from becoming the only proof of readiness.

The matrix also separates readiness from performance. A highly accurate model may still be unready if it cannot be overridden, audited, or rolled back. A less sophisticated system may be safer in some contexts if its limits are clear, its owner is reachable, and its fallback modes are tested. This distinction matters because disaster-AI procurement can overemphasize predictive performance while underweighting operational governability.

Table 6. Design principles embedded in the readiness matrix.

Principle	Purpose	Operational implication
Evidence-based	Avoid unsupported readiness claims.	Each L2 or L3 claim needs logs, exercise records, or event evidence.
Runtime-oriented	Focus on controls that must function during use.	Questions ask whether the team can act now, not only document intent.

Control-specific	Prevent averaging from hiding critical gaps.	Weak ownership or rollback remains visible even if auditability is strong.
Standards-composable	Recognize that no single framework covers all controls.	AI, geospatial, telecom, humanitarian, and equity evidence must be combined.
Equity-integrated	Treat language, community correction, and data rights as operational controls.	Affected populations are part of runtime readiness, not an appendix.

13. THREATS TO ADOPTION

Several adoption barriers should be expected. The first is documentation bias: teams may have extensive design documents but limited evidence that controls were exercised. The matrix intentionally exposes this gap, which may create friction during early assessments. The second barrier is ownership ambiguity across agencies. Disaster response often involves multiple organizations, and no single party may believe it owns all AI-system decisions. The third barrier is technical fragmentation. Model registries, logging systems, incident tools, and emergency-management workflows may not be connected. [33]

A fourth barrier is false confidence created by high model performance in laboratory or benchmark settings. Benchmarks may not include degraded connectivity, multilingual communication, local data restrictions, or time-constrained override. A fifth barrier is procurement language that asks for responsible-AI principles but does not ask for exercised evidence. Vendors and internal teams may therefore optimize for documentation rather than runtime proof.

These barriers are addressable. Procurement language can require L2 evidence for high-risk controls. Exercises can be designed to test governance failures as well as operational scenarios. Technical teams can map model registries, audit logs, and incident systems to the readiness questions. Governance teams can require that post-event reviews update the maturity profile. The matrix is useful precisely because it turns vague governance aspirations into testable conditions.

14. DISCUSSION

The core implication is that governance should be tested like availability, latency, and failover. A disaster-AI system should not simply claim human oversight; it should demonstrate that an operator can override a consequential recommendation before the decision deadline. It should not simply claim model governance; it should demonstrate that a known-good model can be restored. It should not simply claim accountability; it should show who can stop the system and how the decision will be logged.

The framework also clarifies why runtime governance is interdisciplinary. AI risk management contributes accountability and monitoring. Geospatial and telecommunications standards contribute exchange and continuity. Humanitarian and public-health sources contribute trust, community engagement, multilingual communication, and equity. Disaster-AI readiness requires composition across these domains.

The Global South and resource-constrained environments are not edge cases for this framework. They are strong tests of it. Intermittent connectivity, linguistic diversity, local data rights, limited institutional capacity, and multi-agency response structures make runtime governance more important, not less. A system that can pass the matrix in these settings is likely to be more robust elsewhere.

15. ANALYTICAL EVALUATION

Internal completeness and traceability were checked across the artifact. All ten controls appear in the evidence map, standards crosswalk, readiness matrix, and pass/fail questions. The five exercise injects in Table 5 collectively test every control, with ownership appearing twice. Tables 3 through 5 also connect each control to a requirement, an evidence artifact, and an exercise observation. This is structural traceability within the proposed instrument, not evidence that an external assessor can reproduce the coding.

The four maturity levels create their clearest distinction between L1, documented, and L2, exercised: a policy alone cannot satisfy an evidence claim that requires a drill artifact. The distinction between L2 and L3 is less controlled because it partly measures exposure to a real event. For that reason, L2 is the proposed minimum target for life-safety use. Readiness remains a profile rather than an average, so a strong control cannot compensate numerically for absent ownership, rollback, incident reporting, or degraded mode.

Scenario coverage is broad but not exhaustive. The four deployment scenarios and five exercise injects address early warning, damage assessment, connectivity loss, language access, community correction, and data restrictions. They do not yet cover slow-onset hazards, compound or cascading failures, contested authority across agencies, or sustained operation over long events. No real system was scored, no exercise was run for this study, no independent assessor applied the matrix, and no relationship with deployment outcomes was tested. The evaluation therefore supports internal coherence and testability, not empirical effectiveness.

16. LIMITATIONS AND FUTURE WORK

This study has six limitations. First, it is a structured scoping review rather than a complete multi-database PRISMA review. Second, the purposively screened 80-record set cannot establish field prevalence, and the full source-by-control codebook is not reproduced. Third, coding was single-coder and exploratory, so adjacent ranks and crosswalk cells require independent review. Fourth, Table 2 is the authors' interpretation of screened evidence rather than official conformance guidance; the ITU and WHO columns in particular need future confirmation against named primary instruments. Fifth, maturity levels and the non-compensatory profile are conceptual and lack validated thresholds. Sixth, the matrix has not been tested against assessor agreement, exercise performance, live-event outcomes, or implementation cost.

Future work should proceed in stages. The first stage is to publish a source-by-control codebook, verify the crosswalk against primary instruments, and re-code a sample with multiple reviewers to report agreement. The second is a prospective tabletop study in which independent assessors

apply the matrix to the same systems and record evidence availability, scoring disagreements, time to decision, and remediation actions. The third is longitudinal validation of the falsifiable hypothesis that systems reaching L2 on ownership, rollback, incident reporting, and degraded mode experience fewer unresolved runtime failures. Further work should formalize non-compensatory decision rules, develop profiles for floods, wildfires, earthquakes, public-health emergencies, and compound hazards, and integrate evidence links into model registries, incident platforms, procurement workflows, and governance dashboards.

17. CONCLUSION

Disaster-AI governance is often completed before deployment, but its most consequential moments occur while a system is running. The evidence map identifies a marked concentration: five controls account for 80.2% of the 232 coded assignments, while ownership, rollback, incident reporting, data sovereignty, and degraded-mode operation share only 19.8%. The crosswalk is also incomplete, with support in 31 of 60 cells and only two to four instrument families per control. These results show why no single governance or technical framework is sufficient for runtime readiness.

The Operational AI Readiness Matrix translates this fragmented evidence into ten control-specific questions, four maturity levels, and concrete evidence expectations. Its practical standard is deliberately demanding: a disaster-AI system is not ready because an owner, override, rollback plan, incident channel, language policy, or fallback mode is documented. For life-safety use, those controls should be exercised and should produce evidence before deployment. A profile preserves weak controls that an average score would hide.

The matrix is a design-science artifact, not a validated certification scheme. The immediate future scope is to publish the coding artifact, verify primary-instrument mappings, measure inter-assessor agreement, and test the matrix in tabletop exercises. Longer-term studies should examine whether L2 readiness predicts faster correction and fewer unresolved failures during real events, refine sector-specific thresholds, and connect runtime evidence to operational governance systems. These steps would determine whether the proposed matrix can move from an internally coherent assessment instrument to a reliable basis for procurement and deployment.

18. ACKNOWLEDGMENTS

The authors acknowledge the standards bodies, policy organizations, emergency-management researchers, and Indigenous-data-sovereignty community whose published work informed this synthesis.

19. REFERENCES

[1] Abstreiter, M., Tarkoma, S., and Morabito, R. 2025. Sometimes painful but promising: Feasibility and trade-offs of on-device language model inference.

[2] Agarwal, A. and Nene, M. J. 2025. Standardised schema and taxonomy for AI incident databases in critical digital infrastructure.

[3] Amin, S. and Goldstein, M. 2008. Data against natural disasters: Establishing effective systems for relief, recovery, and reconstruction. World Bank.

[4] Banerjee, S., Hazra, R., and Mukherjee, A. 2026. Bridging the multilingual safety divide: Efficient, culturally-aware alignment for Global South languages.

[5] Blice, A. et al. 2025. An inquiry into artificial intelligence as related to emergency management. Texas A&M University.

[6] California Natural Resources Agency. n.d. Indigenous data sovereignty for tribal stewardship.

[7] CARE Principles for Indigenous Data Governance. 2019. Research Data Alliance International Indigenous Data Sovereignty Interest Group.

[8] Chun, K. P. et al. 2025. Transforming disaster risk reduction with AI and big data: Legal and interdisciplinary perspectives.

[9] Cohen, C. and Werker, E. 2008. The political economy of natural disasters. Harvard Business School.

[10] Da Costa, J., Ebert, E., Hoffmann, D., Cloke, H. L., and Neumann, J. 2025. Signals without action: A value chain analysis of Luxembourg's 2021 flood disaster.

[11] European Commission. 2024. AI Act: Regulatory framework for artificial intelligence.

[12] European Union. n.d. Article 14: Human oversight. EU Artificial Intelligence Act.

[13] GFDRR, Deltares, and University of Toronto. n.d. Responsible artificial intelligence for disaster risk management.

[14] GAO. 2021. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities.

[15] Hadan, H., Mogavi, R. H., Zhang-Kennedy, L., and Nacke, L. E. 2025. Who is responsible when AI fails? Mapping causes, entities, and consequences of AI privacy and ethical incidents.

[16] Hilberts, S., Govers, M., Petelos, E., and Evers, S. 2025. The impact of misinformation on social media in the context of natural disasters: Narrative review.

[17] Intergovernmental Oceanographic Commission and UNDRR. 2019. Limitations and challenges of early warning systems: Palu-Donggala tsunami case study.

[18] Kuglitsch, M. and Xoplaki, E. 2025. International standards for responsible AI in disaster management.

[19] Lavender, S. and Lavender, A. 2024. OGC Disaster Pilot: Provider Readiness Guide. Open Geospatial Consortium.

[20] Lee, C.-C. et al. 2026. Roadmap toward responsible AI in crisis resilience and management.

[21] Liu, J. et al. 2026. Edge-cloud collaborative computing on distributed intelligence and model optimization: A survey.

[22] Mahl, D. et al. 2025. Responsible artificial intelligence in public health: Delphi study on risk communication, community engagement and infodemic management.

[23] Mandefro, M. T. 2026. Harnessing AI for humanitarian impact: Lessons and insights from 11 case studies.

[24] Miranda, L. J. V., Hu, S., Reichart, R., and Korhonen, A. 2026. Multilinguality at the edge: Developing language models for the Global South.

[25] Nguyen, V.-C., Dinh, N.-T., and Kim, Y. 2018. A distributed NFV-enabled edge-cloud architecture for ICN-based disaster management services.

[26] NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

[27] NIST. 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1.

[28] Open Geospatial Consortium. 2018. Disasters Concept Development Study and interoperability resources.

[29] Pohler, J., Flegel, N., Mentler, T., and Van Laerhoven, K. 2025. Keeping the human in the loop: Are autonomous decisions inevitable?

[30] Popchanovska, E. et al. 2026. When AI fails, what works? A data-driven taxonomy of real-world risk mitigation strategies.

[31] Rahnemoonfar, M. et al. 2020. FloodNet: A high resolution aerial imagery dataset for post-flood scene understanding.

[32] Reichstein, M. et al. 2025. Early warning of complex climate risk with integrated artificial intelligence.

[33] Sambasivan, N. et al. 2021. Data cascades in high-stakes AI. CHI Conference on Human Factors in Computing Systems.

[34] Schroter, K. et al. 2025. Fostering interoperability of data, models, communication, and governance for disaster resilience through transdisciplinary knowledge co-production.

[35] Soden, R., Wagenaar, D., and Tijssen, A. n.d. Responsible AI for disaster risk management: Working group summary. World Bank.

[36] Suge, S. et al. 2026. From experimentation to engagement: Participatory AI and power in forced displacement and humanitarian crises.

[37] Tanui, M., Tjalve, M., Weber, D., Grieder, E., and Parkinson, K. M. 2026. Humanitarian AI: Lessons learned, trends and opportunities for 2026.

[38] Trakas, A. n.d. Interoperability: A key requirement for emergency and disaster management. UN-SPIDER.

[39] UNDRR. 2023. New system for tracking hazardous events and losses and damages.

[40] UNDP, WMO, and UNDRR. 2022. Statement on tracking hazardous events and disaster losses and damages.

[41] White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights: Algorithmic discrimination protections.

[42] Zhu, R. and Lan, X. 2026. Multimodal building damage assessment method fusing adaptive attention mechanism and state-space modeling. Sensors.