CONCEPTUAL PRODUCT RESEARCH / CYBERSECURITY

A Human-Governed Agentic Approach to Cyber Resilience

A conceptual architecture for evidence-led security landscape assessment across enterprise IT and operational environments

CONCEPTUAL PUBLICATION · VERSION 1.0

Jami Kiran Kumar

November 2026 · Independent conceptual manuscript
Not peer reviewed · No customer systems assessed

Abstract

Abstract. Organizations operate security controls across fragmented tools, inventories and recovery processes. This conceptual paper describes CyberResilience AI, a proposed workspace connecting authorized evidence, asset and service context, bounded specialist-agent analysis, human review and improvement tracking. The design covers enterprise network domains and recognizes safety constraints in operational technology. It maps the workflow to NIST Cybersecurity Framework 2.0 and discusses relevance to EU NIS2 and DORA contexts without asserting compliance. Provenance, confidence, scope controls and approval gates are treated as core system properties. This is a design exposition, not an empirical evaluation: the deployed application is an interactive browser-based demonstration using local sample records; no customer environment was tested. Validation requirements and a research agenda are proposed.

Keywords: cyber resilience; agentic systems; security assessment; evidence provenance; human oversight; operational technology; risk governance

1. Introduction

Cyber risk decisions increasingly span infrastructure, identities, suppliers, operational technology, business services and recovery capabilities. A conventional finding list may communicate severity while obscuring whether an asset is in scope, whether a result is current, what service depends on it, and who owns the next decision. The practical gap is therefore one of context and workflow as well as detection.

This paper presents a conceptual design for CyberResilience AI: an evidence-led workspace intended to help organizations review their security landscape and resilience posture. The design preserves distinctions among observed facts, imported findings, inferred risk, simulated scenarios and unknowns. It places agent assistance inside a governed process, with people retaining authority over scope and consequential actions.

The contribution is a product architecture and validation agenda, not a claim of technical novelty established through experiments. The current public implementation is a front-end demonstration: local browser storage, example records, deterministic assessment rules, declared network inventory and authorized CSV finding import. It does not perform live discovery or scanning, connect production telemetry, execute penetration tests or provide a configured production multi-agent inference service.

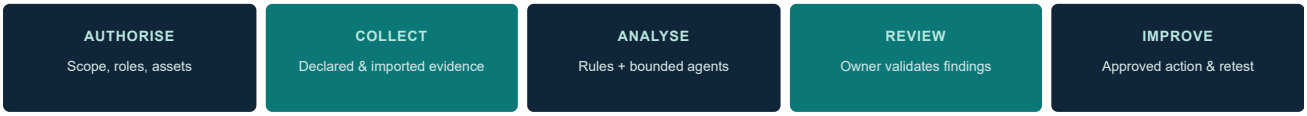
2. Background and design context

NIST CSF 2.0 organizes cybersecurity outcomes across Govern, Identify, Protect, Detect, Respond and Recover, intended to operate as a connected risk-management lifecycle [1]. Its outcome-based format can be adapted to different organizational contexts. This is a useful backbone for a platform linking governance and evidence to security and recovery decisions.

For OT, cybersecurity actions must account for mission, safety, reliability and operational constraints. CISA and partner organizations describe asset inventory and taxonomy as foundational for OT owners and operators to prioritize defenses and response [2]. For EU organizations, NIS2 establishes risk-management and incident-reporting obligations for applicable entities; DORA defines ICT risk and digital operational resilience requirements for covered financial entities [3,4]. A software workspace can help structure relevant information, but cannot determine applicability, provide legal advice or certify conformity.

3. System model

The conceptual workflow comprises five stages: authorization, evidence collection, bounded analysis, human review and improvement. Each stage produces an auditable record carried forward rather than collapsed into an opaque score.



Stage	Inputs and outputs	Control question
Authorization	User, role, tenant, assets, time window and permitted operations.	Is this within the explicitly approved scope?
Evidence collection	Declared inventories, owner records, control evidence and imported results.	What is the source, timestamp, method and coverage?
Bounded analysis	Rule-based checks and specialist-agent summaries or prioritization.	Which statements are facts, inferences or recommendations?
Human review	Validation, assignment, acceptance, exception or rejection with rationale.	Who is accountable for the decision?
Improvement	Approved remediation, exercise, recovery task and reassessment.	What changed, and what evidence verifies improvement?

A normalized finding can be represented as (asset, service, observation, status, source, timestamp, confidence, severity, business context, owner, recommendation, approval state, closure evidence). Status should distinguish at least confirmed, suspected, simulated, remediated and evidence gap. These are design proposals; the demo implements only a subset in local records.

4. Specialist agents and governance

The proposed agent model consists of bounded analytical roles coordinated by an orchestration service. Agents receive only the records needed for their task and return structured outputs with supporting record identifiers. Example roles include asset reconciliation, vulnerability triage, threat-context review, service-impact analysis, framework evidence organization, incident/recovery planning and executive synthesis.

Agent authority should be asymmetric: agents may summarize, compare, identify missing evidence, propose questions and draft recommendations; humans authorize scan scope, approve changes, accept risk, trigger containment or communicate externally. A policy layer should enforce role permissions, tool allowlists,

data minimization, rate limits, approval gates and an append-only audit trail. Tool execution should be isolated, and each integration should have documented permissions and a revocation path.

A model output is not itself a verified observation. Retain input records, source identifiers, model and prompt versions where applicable, output, uncertainty, reviewer action and resulting task. Retrieval should be constrained to tenant-authorized data. Evaluation should test evidence attribution, unsupported claims, prompt injection through imported content, cross-tenant leakage, unsafe tool invocation and consistency with incomplete records.

5. LAN/WAN and OT coverage

The design treats network coverage as declared domains: LAN, WAN, intranet, switching, routing, firewalls, wireless, remote access, network monitoring, identity, cloud/application, email/DNS, endpoint, backup and OT/IoT. Coverage should be shown as assessed, partial, not assessed or unknown, with evidence basis and date. Missing inventory must never appear as a clean bill of health.

An evidence-led review can flag an overdue firewall configuration review, missing owner, unknown firmware or absent segmentation evidence. These are prompts, not proof that a control is defective. Confirming ACL effectiveness, discovering hidden devices, validating CVEs or proving segmentation requires appropriate authorized source data or a separately scoped test.

For industrial control systems, default to passive inventory, configuration evidence and isolated simulations unless an operator approves a specific active test. Coordinate with safety and operations leads and account for vendor restrictions and outage windows. An agent must not autonomously probe or alter production OT. This reflects the operational sensitivity in OT asset-inventory guidance [2].

6. Risk prioritization and resilience measures

A transparent prioritization model should combine technical severity with exploit evidence, exposure, service criticality, dependencies, compensating controls, confidence and freshness. Disclose factors and weights and allow organizations to align them with risk appetite. Keep inputs visible so analysts can challenge the rank; do not manufacture precision when evidence is sparse.

- **Coverage:** assets with a known owner, service, criticality and current evidence.
- **Exposure handling:** findings by severity and validation state, age and accountable owner.
- **Control evidence:** freshness and review state for identity, configuration, segmentation, logging and data protection.
- **Resilience:** recovery exercise completion and measured restoration against approved objectives.
- **Decision quality:** validation, correction, accepted risk, false positives and action closure.

A composite score must not imply that an organization is “secure” or “hack-proof.” Disclose scope, source dates, model version, exclusions, unknowns and confidence. Separate posture indicators from operational service-level measures and establish baselines before interpreting trends.

7. Sector use cases

Sector templates can help users declare important assets and dependencies without substituting for sector expertise. Supermarket assessments might connect point-of-sale and payment systems, store connectivity and refrigeration monitoring to continuity requirements. Blood banks and hospitals can map storage monitoring, clinical and laboratory systems, medical devices, suppliers and validated restoration steps. Power-grid operators can prioritize passive asset visibility, remote access, segmentation evidence and safe recovery planning. Banks and trading organizations can connect ICT risk, critical services, market connectivity, transaction integrity, third-party dependencies and resilience exercises.

These examples illustrate requirements, not tested product performance. No output should claim it will prevent a particular hack or guarantee availability. Security results depend on the environment, evidence, controls, operations and attacker behavior.

8. Evaluation agenda

A credible production evaluation should be staged and independently reviewable:

Evaluation	Proposed method	Success evidence
Evidence integrity	Seed known confirmed, suspected, simulated and gap records; compare labels and source links.	No status conflation; traceability to source records.
Prioritization utility	Qualified reviewers rank blinded findings with and without service context.	Ranking agreement, decision time and rationale quality.
Agent reliability	Test incomplete, stale, contradictory and adversarial evidence.	Unsupported assertion rate, abstention and injection resistance.
Authorization enforcement	Exercise identity, role, tenant, tool allowlist and approval controls.	Prohibited actions blocked and audited.
OT safety	Review with operators; verify passive defaults and explicit test approvals.	Documented hazards, constraints and safe failure behavior.
Operational value	Prospective customer-approved pilot; control group where practical.	Predefined evidence coverage, validated closure and recovery outcomes.

No such controlled evaluation is claimed here. The demo is not proof of agent efficacy, detection performance, compliance or breach prevention.

9. Limitations and conclusion

The design depends on reliable, timely and authorized evidence. Connectors introduce credential, data-quality and supply-chain risks. Language models can produce unsupported content; deterministic rules can miss context; scores can create false confidence. Human review can also fail without adequate evidence, authority and time. The application is not a replacement for asset discovery, vulnerability scanners, security operations platforms, incident responders, auditors or recovery testing.

CyberResilience AI offers a coherent product direction: connect evidence and service context; make uncertainty explicit; use bounded analysis to help teams prioritize; and record human decisions through remediation and recovery. Its value must be demonstrated through secure production engineering, customer-

authorized pilots and measured evaluation. Until then, the public application remains a product demonstration.

Ethics, competing interests and publication status

Author: Jami Kiran Kumar. The author is associated with the CyberResilience AI product concept. This manuscript is an independent conceptual product publication, not peer reviewed and not based on new empirical research or customer-deployment data. No customer systems were assessed for this paper. All capabilities beyond the explicitly described demonstration are proposed design direction.

References

[1] NIST, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29 (2024).. https://doi.org/10.6028/NIST.CSWP.29

[2] CISA et al., Foundations for OT Cybersecurity: Asset Inventory Guidance for Owners and Operators (2025).. https://www.cisa.gov/resources-tools/resources/foundations-ot-cybersecurity-asset-inventory-guidance

[3] Directive (EU) 2022/2555 (NIS2), Official Journal of the European Union.. https://eur-lex.europa.eu/eli/dir/2022/2555/oj

[4] Regulation (EU) 2022/2554 (Digital Operational Resilience Act), Official Journal of the European Union.. https://eur-lex.europa.eu/eli/reg/2022/2554/oj

Publication note: Conceptual product and architecture paper. Not peer reviewed. Not based on empirical research or customer testing. Standards and legal references provide context and are not legal or compliance advice.