EU CyberNet

The Institutional AI Dependency Test: Can Public Services Continue When AI Cannot Be Trusted?

An Evidence-Led Governance Exercise for Cyber Resilience

- Andra T. Alcalá

EU CyberNet Expert Series
No 8
2026

about

Andra T. Alcalá

Andra T. Alcalá is a PhD researcher and specialist in cybersecurity regulation, cyber resilience, AI governance and institutional accountability. She leads applied research at the Innovatech Institute of Cyber Policy & Technology, advises public and private organisations and conducts independent reviews of AI governance, cyber resilience and digital oversight. She is CEO of Genesis International Business Group and the author of forthcoming books on cyber resilience, AI governance, audit and assurance. She has been a member of the EU CyberNet Expert Pool since February 2026.

summary

When AI fails, it may not stop. It may remain available, produce convincing outputs and continue to act after an institution has lost a justified basis for relying on it. In public services, this turns a technical problem into a test of authority, accountability and continuity, particularly where agentic AI may already have accessed systems, initiated workflows or taken action. This article introduces the Institutional AI Dependency Test, a six-stage exercise for assessing whether an institution can identify its operational dependence on AI, restrict or suspend the system, trace affected outputs and actions, maintain essential services and restore use only when trust has been re-established through evidence. The test addresses the hardest case for AI governance: the system still works, but the institution can no longer safely depend on it.

AI failure does not always announce itself. A system may remain operational, continue shaping decisions and even take action after the basis for relying on it has weakened. The resulting risk is not merely technical but institutional: an organisation may become dependent on a tool whose reliability it can no longer justify. In this article, "loss of trust" refers to the suspension of justified operational reliance on an AI system's outputs or actions. As public institutions integrate AI into service delivery, internal processes and decision support, [1] the ability to recognise and manage that loss of reliance becomes a central challenge for governance, cyber resilience and public-service continuity.

The Governance Gap: Compliance Does Not Prove Resilience

An institution may have an AI policy, a risk register, an approval process and a requirement for human oversight. These are necessary foundations, but they do not demonstrate that the organisation can respond effectively when an AI system becomes unreliable. The NIST AI Risk Management Framework treats governance, accountability, security and resilience as core elements of AI risk management. [2] For high-risk AI systems, Article 14 of the EU AI Act also establishes requirements for effective human oversight. [3]

Yet controls may fail under operational pressure. Human oversight becomes a resilience measure only when the responsible person has access to the necessary information, as well as the skills, time and authority to challenge or suspend the system and activate an alternative process.

When an AI System Remains Available but Cannot Be Trusted

A system outage is usually visible. A loss of trust is not. AI systems can be exposed to adversarial manipulation, including data poisoning and evasion attacks. [4] They may also become unreliable as their data, operating environment or context changes over time, sometimes in ways that are difficult to detect. [2] They may therefore continue producing plausible outputs from compromised or outdated information.

Consider a public authority using AI to summarise cyber-threat intelligence and prioritise incidents. During a cyber event, analysts discover that the system may have processed unreliable material, but cannot determine when the problem began, which outputs were affected or whether they influenced response decisions.

The institution must decide whether to suspend the system, preserve evidence, review previous decisions and maintain the affected service. The central question is no longer whether the technology functions, but whether the institution can remain effective while operational reliance on it is suspended. This scenario serves as a running example throughout the six stages of the test.

From AI Assistance to AI Action

Institutional dependency deepens when an AI system moves from supporting decisions to executing them. Agentic AI may interact with external systems, access databases, initiate workflows, communicate with stakeholders or perform technical tasks. [5] By the time its reliability is questioned, it may already have prioritised incidents, distributed information or initiated containment measures.

The institution must therefore know what the system is authorised to do, which data and tools it can access, whether its actions are traceable and who can revoke its permissions or stop an active workflow.

Unlike conventional technology failure, an unreliable AI system may remain available, produce persuasive outputs and continue to act while the affected period remains uncertain. Restoring technical availability is therefore insufficient: the institution must identify what was affected, determine which functions can no longer be performed reliably and preserve continuity while trust is re-established.

The AI Dependency Test

Existing approaches cover elements of this problem, including trustworthy AI, human oversight, incident response and operational resilience. The Institutional AI Dependency Test integrates them around a narrower question: *can a public authority continue to exercise authority and deliver the supported service when an AI system remains available but can no longer be relied upon?* It complements, rather than replaces, AI governance audits, technical security assessments and incident-response plans.

Developed for public institutions but intentionally jurisdiction-neutral, the test can be adapted to different legal systems, institutional structures and levels of AI maturity. The institution selects one AI use case whose loss of trust could significantly affect a service, decision or operational process, and brings together relevant operational, technical, data, legal, continuity and provider representatives.

The test may be conducted as a facilitated tabletop exercise, an evidence-led review or both. Where feasible, the facilitator should be independent of direct system ownership. A focused tabletop may take two to four hours, while a combined review and exercise may require a full working day or several shorter sessions.

The framework follows six connected stages: identifying the dependency, establishing accountability, verifying controls and evidence, testing the loss of trust, maintaining the affected service and recovering and improving.

Participants should record the evidence presented, decisions taken, authority exercised, time required to act and capability gaps exposed.

Each assurance question should be rated as *demonstrated, partially demonstrated, not demonstrated or not applicable*. A capability is demonstrated only where the evidence is attributable, relevant to the assessed use case and sufficiently recent to reflect the system's current configuration, and where the required action can be performed during the exercise.

Figure 1 presents the six stages as an iterative cycle rather than a one-off checklist. Recovery and improvement may reveal new dependencies, ownership gaps or control weaknesses, requiring the institution to revisit earlier stages.

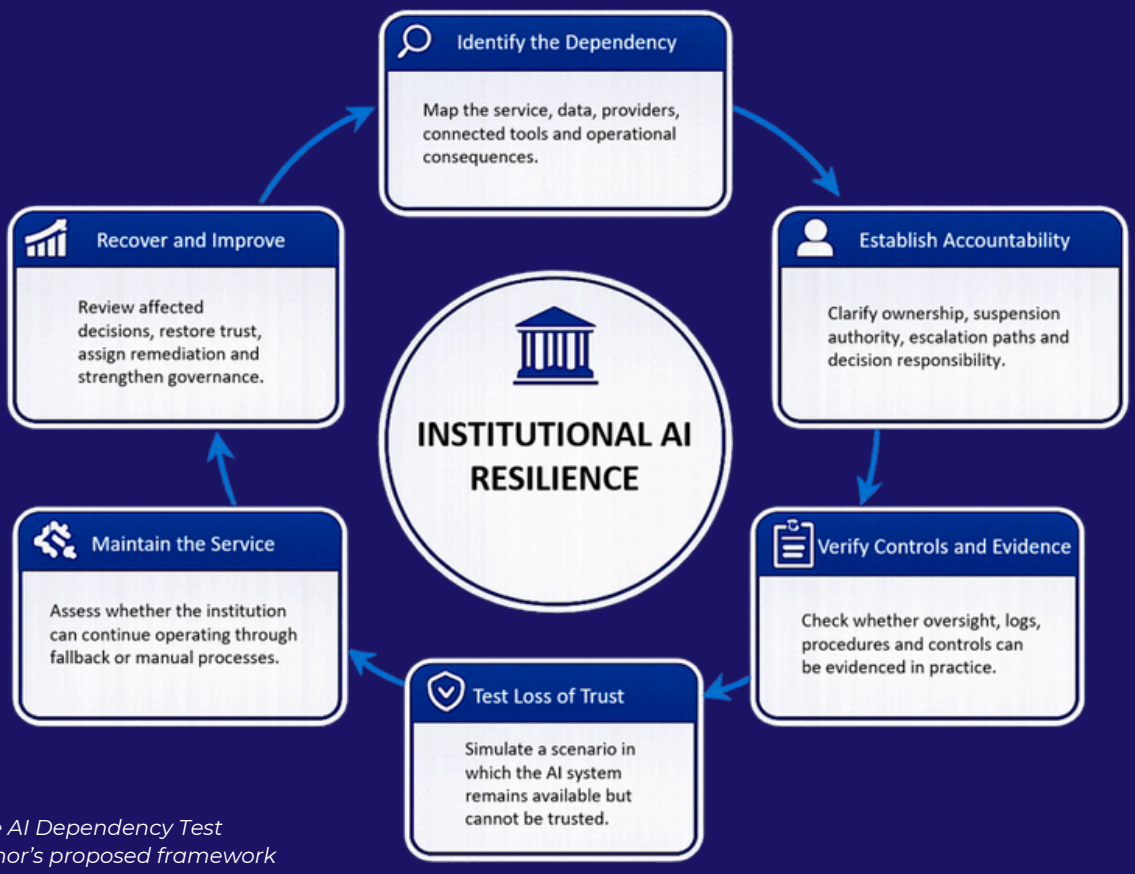

Figure 1. The AI Dependency Test
Source: Author's proposed framework

1) Identifying the Dependency

The exercise begins with one specific AI use case rather than a general discussion of AI risk. Although the NIST AI Risk Management Framework and its Playbook recommend inventorying AI systems and documenting information relevant to maintenance, incident response and third-party risk, [2, 6] an inventory alone does not reveal institutional dependence.

The institution should map the service or process supported, the people relying on it, its data and provider dependencies, connected systems and tools and the outputs, decisions or actions it influences. It should also assess the consequences of losing trust. For agentic AI, this includes the system's permissions, the actions it can initiate and whether those actions are observable and reversible. [5]

In the running scenario, this means mapping the incident-prioritisation function, threat-intelligence feeds, provider, connected response tools and the maximum tolerable period without trusted summaries.

The defining question is: *What would the institution be unable, or significantly less able, to do if this system could no longer be trusted?*

The output should be a concise dependency statement recording the supported service, influenced outputs or actions, principal dependencies, operational consequences and maximum tolerable period without the system. If the institution cannot produce this statement, the capability is not demonstrated. Box 1 provides a template for use during the later accountability, testing and continuity stages.

Template for documenting a material AI dependency

System:
Name or description of the AI system
Supported service or process:
Public service, internal function, or operational activity
Decisions or actions influenced:
What the system informs, recommends, prioritises, or initiates
Critical dependencies:
Data sources, provider, connected systems, tools, permissions
Operational consequence of loss of trust:
What the institution would be unable, or significantly less able, to do

Box 1. Example AI Dependency Statement Template
Source: Author's own elaboration

2) Establishing Accountability

Once the dependency has been identified, the institution must determine who has the authority and responsibility to act when trust in the AI system is lost. Accountability cannot be reduced to naming a system owner: operational, technical, data, legal, continuity, communications and provider functions may each hold part of the relevant knowledge or authority.

The NIST AI Risk Management Framework calls for documented and empowered roles, responsibilities and lines of communication, [2] while its Playbook recommends defined delegated authority and documented escalation and go/no-go decisions. [6]

The exercise should establish who can restrict or suspend the system, activate the fallback process, review affected outputs or actions, approve return to service, engage the provider and communicate with relevant stakeholders. These responsibilities may be distributed: the system owner may understand the technical configuration, while the service owner assesses the consequences of interruption and cybersecurity personnel investigate potential compromise.

For agentic AI, the accountability question extends to action authority. The institution must know who approved the system's permissions, who can revoke access to connected tools, who can stop an active workflow and who is responsible for reviewing actions already taken. [5]

In the running scenario, the institution would identify who may suspend the tool, return analysts to original intelligence sources,

review earlier prioritisation decisions and authorise its return to service.

The practical question is: *When the reliability of the AI system is uncertain, who has the authority to decide what stops, what continues and what must be reviewed?*

The output should be an accountability decision map covering suspension, activation of the fallback process, review of affected outputs or actions and approval for return to service. The capability is demonstrated only where the responsible people can be reached and can exercise formally delegated authority during the test.

3) Verifying Controls and Evidence

Clear accountability is not enough if the institution cannot demonstrate that its controls work.

The policies and responsibilities identified in the previous stage should therefore be examined against current evidence. The NIST AI Risk Management Framework emphasises documentation that supports informed decisions and the identification of internal controls, including those relating to third-party AI components. [2, 6] ENISA similarly approaches AI cybersecurity across foundational, AI-specific and sector-specific layers, reinforcing the need to assess the system, its operational environment and the processes surrounding it together. [7]

Generic policies or outdated records may demonstrate intent, but not operational readiness. Evidence should show not only that a control exists, but that it has operated effectively under the current system configuration, that responsible personnel know how to activate or apply it and that the resulting action is recorded. The purpose is to determine whether the institution can detect unreliable behaviour, restrict the system, trace affected outputs or actions and continue the supported service (not only to collect documents for their own sake).

Relevant evidence might include source-provenance records, anomaly alerts, access controls, audit trails and a tested manual incident-triage procedure.

At this stage, the institution should be able to answer a practical assurance question: *What evidence demonstrates that this control would remain effective if trust in the AI system were lost?*

Any gaps identified at this stage should be recorded and carried into the loss-of-trust exercise. Evidence should be classified as current and tested, current but untested, outdated or absent. A control that exists only in policy should not receive a rating of demonstrated.

4) Testing the Loss of Trust

The next stage converts the dependency statement, accountability map and control evidence into a facilitated operational scenario. Participants are informed that the AI system remains available, but credible indicators suggest that its recent outputs may be unreliable. The cause, starting point and extent of the problem should initially remain uncertain.

The facilitator may then introduce additional information in stages: evidence of manipulated or contaminated data, inconsistent outputs, indications that affected recommendations influenced decisions or confirmation that an agentic workflow has already initiated actions. Participants must decide whether to restrict or suspend the system, preserve relevant evidence, revoke permissions, identify affected outputs or actions, activate escalation procedures and inform relevant stakeholders.

In the running scenario, the facilitator might first reveal inconsistent summaries, then evidence of manipulated source material, and finally that an automated containment action has already been initiated.

NIST guidance recommends monitoring AI systems for unexpected behaviour and establishing incident-response arrangements for AI-related events. [2, 6] For agentic AI, the exercise should also test whether active workflows can be stopped, connected tools isolated and previous actions traced. [5]

Decisions, responsible persons and elapsed time should be recorded. The stage is demonstrated only where the institution can make, record and execute a decision within the required response period, identify the evidence and authority on which it was based and revise it when new information emerges. Failure indicators include uncertainty over who may restrict the system, inability to preserve or trace evidence, conflicting authority or continued reliance on outputs after trust has been suspended.

5) Maintaining the Affected Service

Suspending an AI system may reduce one risk while disrupting the service that depends on it. Before the exercise, the institution should define the critical function that must continue, the minimum volume or quality of service that must be maintained and the maximum tolerable period of disruption.

In the running scenario, analysts would revert to original intelligence feeds and manual triage while maintaining the minimum required incident-response capability.

Participants should then be required to perform or simulate the critical task using manual procedures, original records, alternative information sources, independent verification or a temporary redistribution of responsibilities. ISO 22301 provides an international management-system basis for preparing for, responding to and recovering from disruptive incidents. [8] Applied to AI dependency, this means testing whether staff can access the required information, understand the fallback procedure, sustain it for the necessary period and manage any resulting delay or backlog.

The relevant measure is not whether an alternative procedure has been documented, but whether it can preserve the critical function at the required level. This stage is not demonstrated where the fallback depends on inaccessible data, unavailable personnel, untested assumptions or continued use of outputs whose reliability is uncertain.

6) Recovering and Improving

Return to service should not occur simply because the AI system is technically available again. Before renewed use is approved, the institution should define how trust will be re-established and what evidence will support that decision.

Depending on the incident, recovery criteria may include identifying the cause of the reliability failure, verifying the integrity of relevant data and system components, reviewing affected outputs or actions, revalidating permissions, conducting independent checks and confirming that residual risk is acceptable. The person authorised to approve return to service should be identified before the exercise begins.

In the running scenario, return to service would require validation of the affected sources and data, review of previous incident prioritisation and agentic actions and documented approval.

The exercise should conclude with prioritised corrective actions, assigned owners and implementation deadlines. NIST connects AI risk management with monitoring, response and recovery, while ISO/IEC 42001 places continual improvement within the AI management system. [2, 6, 9]

Recovery is demonstrated only where the institution can identify the cause or contain the relevant uncertainty, validate the integrity of the system and its inputs, define the scope of retrospective review and obtain documented approval from the authorised decision-maker. Restoring the technology

without addressing the governance weaknesses exposed by the exercise does not constitute recovery.

From Assessment to Action

The value of the AI Dependency Test lies not in completing each stage, but in demonstrating that the institution can act effectively when trust in an AI system is lost. Results should therefore be assessed through observable capabilities rather than the existence of policies alone.

Table 1 translates these capabilities into practical assurance questions, indicative evidence and an assessment result. A rating of *demonstrated* requires both current evidence and successful performance during the exercise. *Partially demonstrated* indicates that the capability is incomplete, untested, dependent on informal arrangements or too slow to operate reliably under pressure. Traceability and agent authority are treated as cross-cutting assurance capabilities assessed across several stages rather than as separate stages.

The scorecard provides a high-level assurance structure rather than an exhaustive exercise script. Institutions may expand each area through more detailed prompts, evidence requirements and scenario injects according to the use case, legal environment, system complexity and operational criticality. During the exercise, it should be supported by a decision and evidence log recording actions taken, responsible persons, elapsed time, identified gaps and required remediation.

Area	Practical assurance question	Example evidence	Assessment result
AI dependency	Can the institution identify the service, data, providers, connected tools and operational capabilities on which the AI system depends?	AI use-case inventory and dependency statement	□ D □ PD □ ND □ N/A
Accountability	Is it clear who can restrict or suspend the system, activate the fallback process, review affected outputs or actions, and approve its return to service?	Accountability decision map, assigned roles and escalation procedures	□ D □ PD □ ND □ N/A
Control effectiveness	Is there current evidence that monitoring, human oversight, access controls and incident procedures operate in practice?	Monitoring records, control tests, access records and incident-response documentation	□ D □ PD □ ND □ N/A
Traceability	Can affected outputs, decisions and agentic actions be identified, preserved and reviewed?	System logs, decision records, audit trails and evidence-preservation procedures	□ D □ PD □ ND □ N/A
Service continuity	Can the supported service continue at the minimum acceptable level without relying on trusted AI outputs?	Tested fallback procedure, manual process and continuity records	□ D □ PD □ ND □ N/A
Agent authority	Can permissions be revoked, active workflows stopped, connected tools isolated and previous actions reviewed or reversed?	Access records, revocation controls, workflow logs and action histories	□ D □ PD □ ND □ N/A
Recovery and improvement	Are the criteria for re-establishing trust, approving return to service and addressing identified weaknesses clearly defined?	Recovery criteria, validation records, assigned corrective actions and remediation roadmap	□ D □ PD □ ND □ N/A

Table 1. AI Dependency Test Assurance Scorecard
Source: Author's own elaboration.
Note: D = demonstrated; PD = partially demonstrated; ND = not demonstrated; N/A = not applicable.

The test should not be reduced to a composite numerical score. A satisfactory outcome requires that no critical capability be rated 'not demonstrated'. Any capability rated 'partially demonstrated' should result in documented remediation actions, assigned owners and implementation deadlines. An inability to restrict the system, revoke relevant permissions, identify affected outputs or actions, maintain the minimum required service or apply defined return-to-service criteria constitutes a critical gap. Such a gap

should prevent any increase in the system's permissions, operational role or institutional dependency until it has been remediated or formally risk-accepted by the competent authority, with documented interim safeguards and a defined remediation deadline.

The test can be adapted for public authorities, national and sectoral CSIRTs, regulators, municipalities and operators of critical services. In cyber-capacity-building settings, it may support facilitated tabletop exercises, institutional maturity assessments, cross-functional training and the identification of priority capability gaps. Its application should remain proportionate to the institution's legal environment, organisational maturity, available resources and operational context. [10]

Public institutions do not need to wait for a major AI incident to discover how dependent they have become. They can begin with one material AI-supported service and test whether they can recognise uncertainty, intervene with authority and preserve essential functions when the system remains available but can no longer be trusted.

The decisive question is not whether the AI system can be restored, but whether the institution can continue to govern and deliver essential services while trust in that system is suspended.

References

[1] OECD. (2026). *Digital government outlook 2026: From foundations to transformational impact* (Chapter 4, "Adopting and governing AI in government"). https://www.oecd.org/en/publications/2026/06/digital-government-outlook_4585678e/full-report/adopting-and-governing-ai-in-government_7ef312a9.html

[2] Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[3] European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, Article 14. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[4] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (NIST AI 100-2e2025). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2025

[5] National Institute of Standards and Technology. (2025, August 5). *Lessons learned from the consortium: Tool use in agent systems*. https://www.nist.gov/news-events/news/2025/08/lessons-learned-consortium-tool-use-agent-systems

[6] National Institute of Standards and Technology. (n.d.). *NIST AI RMF Playbook*. NIST AI Resource Center. Retrieved July 20, 2026, from https://airc.nist.gov/airmf-resources/playbook/

[7] European Union Agency for Cybersecurity. (2023). *Multilayer framework for good cybersecurity practices for AI*. https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

[8] International Organization for Standardization. (2019). *ISO 22301:2019: Security and resilience-Business continuity management systems-Requirements.* https://www.iso.org/standard/75106.html

[9] International Organization for Standardization. (2023). *ISO/IEC 42001:2023: Information technology-Artificial intelligence-Management system.* https://www.iso.org/standard/42001

[10] EU CyberNet. (2023). *Operational guidance: The EU's international cooperation on cyber capacity building* (2nd ed.). European Commission, Service for Foreign Policy Instruments. https://www.eucybernet.eu/wp-content/uploads/2023/11/operational-guidance-for-the-eu-international-cooperation-on-ccb.pdf

The Institutional AI Dependency Test: Can Public Services Continue When AI Cannot Be Trusted? An Evidence-Led Governance Exercise for Cyber Resilience

Andra T. Alcalá

EU CybetNet Expert Series, No 8, 2026

© EU CyberNet 2026

EU CyberNet

Funded by the European Union

Views and opinions expressed are those of the author only and do not necessarily reflect those of the European Union or the EU CyberNet. Neither the European Union nor EU CyberNet can be held responsible for them.