FSB FINANCIAL STABILITY BOARD

Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation report

Response to Consultation

Artificial Intelligence Underwriting Company

1. **Do you agree with the benefits and risks of AI adoption by financial institutions described in this report? Are there any substantive benefits or risks not covered?**

   Yes. AIUC agrees with the principal benefits and risks described in the report. The FSB appropriately recognizes benefits in fraud detection, cyber defense, operational efficiency, customer service, risk management, and financial inclusion, while also addressing model and data risk, explainability, performance degradation, cyber and ICT risk, third-party dependence and concentration, correlated behavior, conduct risk, and the distinct risks of agentic systems.

   We do not identify any top-level risk category that is omitted. The report already addresses common providers and correlated outcomes, unexpected provider changes, unauthorized actions, disruptions to connected systems, multi-agent interactions, financial-transaction controls, and incident and near-miss information sharing.

   Although AIUC does not identify a missing top-level risk category, the final report could make two risks it already recognizes more measurable and actionable: common dependencies that may produce correlated outcomes, and the classification of incidents and near misses.

   Common dependencies and correlated outcomes

   The FSB separately identifies third-party concentration, common models and infrastructure, market correlations, and multi-agent interactions. The final report could more explicitly link those elements by encouraging institutions and authorities to monitor common exposures and plausible correlated outcomes. For an agentic deployment, relevant exposure indicators may include:

   - the number and type of financial actions the system can initiate;

   - maximum value and velocity of transactions within its authority;

   - the volume and sensitivity of records it can access;

   - external communications and counterparties it can reach;

   - critical systems, tools, APIs, and providers on which it depends;

- the number of institutions or business lines using a common model, agent platform, data source, or orchestration layer; and

- the time required to detect, contain, reverse, or reconcile erroneous actions.

These measures would help translate the report's systemic-risk discussion into institution-level inventories and sector-level scenario analysis without creating a single universal metric.

Incident and near-miss taxonomy

Sound Practice 11 appropriately encourages sharing information about AI incidents, near misses, deteriorating performance, threats, and lessons learned. The value of that practice would increase if institutions used a common minimum taxonomy. The final report could encourage classification by:

- source of failure: model, data, configuration, tool, identity or access, human oversight, third party, or malicious use;

- failure mode: technical failure, control failure, misuse, unauthorized action, erroneous action, or inadequate escalation;

- consequence: customer harm, financial loss, data exposure, service disruption, regulatory breach, market impact, or avoided harm;

- detection channel and time to detection;

- affected deployment configuration and provider dependencies;

- control that prevented or limited loss; and

- corrective action and reassessment outcome.

Information-sharing mechanisms would need appropriate safeguards for confidentiality, personal data, legal privilege, cybersecurity, and supervisory information.

Recommendation

AIUC recommends retaining the report's risk taxonomy while adding: (i) common exposure indicators for highly autonomous or widely shared systems; and (ii) a minimum, internationally interoperable incident and near-miss taxonomy. These additions would operationalize risks already identified rather than create a separate category.

2. Are the sound practices sufficiently comprehensive and clear to enable financial institutions' responsible AI adoption?

The sound practices are sufficiently comprehensive and clear as principles. Their actionability would improve through illustrative, non-prescriptive examples of the evidence and decision records expected where an institution's assessment of materiality and risk warrants heightened independent review.

The report already contains the foundations of an assurance model. Sound Practice 2 recognizes independent second-line assessment and third-line assurance proportionate to materiality and risk, and states that audit, evaluation, risk management, and testing should remain functionally separate from development. Sound Practice 12 discusses access and audit rights, independent and pooled audits, certifications, standards, and external attestations.

For deployments assessed by the institution as material or high risk, these elements should culminate in a documented, deployment-specific assurance conclusion. That conclusion should identify the assessed system and operating envelope; the evidence and testing considered; material findings and limitations; residual risks and approval conditions; the accountable decision-maker; and the events or changes that will require reassessment.

The review should be performed by appropriately skilled, functionally independent personnel. Qualified external assessment should be used where the institution lacks sufficient internal competence, independence, or access to relevant evidence. These requirements need not prescribe a particular certification framework, organizational model, or standardized evidence package, provided that the resulting assurance conclusion is transparent, evidence-based, proportionate to risk, and subject to appropriate refresh.

A proportionate assurance lifecycle

For AI use cases assessed as material or high risk, AIUC recommends that the final report encourage a proportionate assurance lifecycle culminating in a documented, deployment-specific assurance conclusion and an accountable deployment or continued-use decision. Institutions may allocate the relevant responsibilities through different governance and assurance structures, but the lifecycle should achieve the following outcomes:

1. First-line assessment. The deployment owner demonstrates the use case, intended benefit, system boundary, materiality, inherent risk, controls, testing, and residual risk.

2. Independent second-line validation. A function separate from development and operation challenges the risk classification, control design, test plan, thresholds, limitations, and deployment conditions.

3. Third-line assurance. Internal audit or an equivalent independent function assesses whether governance and controls are suitably designed and, where relevant, operating effectively.

4. Qualified external assurance where warranted. External assessment should be considered where internal independence or specialist competence is insufficient; the institution lacks the necessary access to external evidence but a qualified external assessor can obtain or evaluate relevant evidence not reasonably available internally; the system has substantial autonomy or authority; the deployment affects critical operations, customers, or markets; common dependencies create concentration concerns; or law, supervision, counterparties, or the institution's risk appetite call for it.

Illustrative evidence outcomes

For material or high-risk use cases, the final report should give illustrative examples of a proportionate evidence package, which may include:

- the defined use case, system boundary, and approved operating envelope;

- model and component versions; prompts and policies; retrieval sources; tools and APIs; identities and permissions; memory or state; delegation pathways; human-approval points; and deployment environment, as applicable;

- the materiality assessment and inherent- and residual-risk conclusions;

- intended, restricted, and prohibited uses;

- customer-outcome, conduct, fairness, explainability, contestability and redress testing, where relevant;

- control objectives, owners, implementation evidence, and dependencies;

- test methods, test conditions, data, benchmarks, acceptance criteria, limitations, and results;

- findings, severity, remediation, retesting, exceptions, and expiry dates;

- monitoring metrics, thresholds, incidents, near misses, overrides, and performance degradation;

- provider information, audit rights, change-notification obligations, continuity and exit arrangements; and

- approval, conditional approval, restriction, suspension, retirement, or continued-use decision.

This package should be tailored to the intended recipient. Boards, supervisors, auditors, counterparties, and technical operators need different levels of detail, but they benefit from a consistent underlying evidence model.

Suggested clarification across the sound practices

We recommend two supporting clarifications in addition to the targeted amendments proposed above:

- Sound Practice 3: For material or high-risk use cases, lifecycle documentation should identify the approved operating envelope, material assurance findings, residual risks, deployment conditions and reassessment triggers.

- Sound Practice 5: The final report should identify the factors that may warrant a documented independent review or assurance decision, while preserving proportionality and local supervisory judgment.

3. **Do the sound practices strike an appropriate balance between managing risks relating to all forms of AI, and addressing some of the risks relating to emerging and new complex forms of AI, such as GenAI and agentic AI?**

Broadly yes. The technology-neutral structure is appropriate, and the report contains substantial agent-specific guidance. It recognizes unauthorized and erroneous action, multi-agent interaction, memory poisoning, identity and access, least privilege, tool access, transaction controls, human-approval boundaries, kill switches, action-level monitoring, red teaming, and the need to assess whether an agent's actions were appropriate, not merely whether it reached its objective.

Because the provisions relevant to agentic AI are distributed across several Sound Practices, AIUC recommends a brief, cross-referenced implementation profile or annex that explains how the existing practices apply to agentic deployments. This could cover agent and tool inventories, permissions and operating limits, action logging, testing of unauthorized or cascading actions, human-intervention mechanisms, reassessment triggers, and relevant third-party dependencies.

Recommended scope of an agentic implementation profile

An agentic assessment should define the complete system operating within the financial institution's environment. Depending on the use case, that boundary may include:

- models and model versions;

- system and developer prompts, policies, and guardrails;

- retrieval sources, context, memory, and persistent state;

- tools, APIs, code execution, external communications, and connected systems;

- identities, credentials, permissions, and delegated authority;

- sub-agents, agent-to-agent interaction, and orchestration;

- human-approval and escalation points;

- financial, temporal, rate, and volume limits;

- monitoring, incident response, safe-state, and recovery mechanisms; and

- the production environment and relevant third parties.

An assessment limited to the underlying foundation model cannot establish whether this configured deployment is appropriately controlled.

Consolidated agentic profile control outcomes

The final report could consolidate its existing content into the following outcome-oriented control objectives:

1. Every in-scope agent has a unique, attributable identity and no greater access than required for its approved task.

2. Tools and actions are allow-listed, validated, rate- or value-limited where appropriate, and logged.

3. Actions capable of material financial, legal, customer, market, or operational consequences should be subject to risk-appropriate preventive and detective controls. Depending on the approved operating envelope, these may include ex ante human approval, dual control, rate or value limits, automated monitoring, exception handling, and human escalation. Human accountability should remain clear even where monitoring is machine-assisted.

4. The system prevents or detects scope expansion, privilege escalation, unsafe delegation, prompt injection, and manipulation through external content or memory.

5. Monitoring evaluates the sequence and consequences of actions as well as final outputs.

6. Staff can pause, constrain, revoke, or transition the system to a safe state, and the effectiveness of that intervention is tested.

7. Pre-deployment and recurring tests cover adversarial behavior, unauthorized tool use, data leakage, harmful or out-of-scope outputs, unreliable actions, agent-to-agent interactions, and recovery.

8. Decision-relevant execution records support audit, incident investigation, contestability, and reconstruction without requiring disclosure of the private chain of thought.

Suggested clarification: execution records rather than chain-of-thought

The report refers to reliance on model reasoning in several places: chain-of-thought logging as enhanced documentation for GenAI use cases (§2.3); explainability approaches that trace the full reasoning path of agentic systems (§3.5); and monitoring that records the reasoning at each autonomous decision point (§3.7). Annex 2, however, correctly observes that a model's displayed reasoning may not reflect the steps it actually employed. Verbatim reasoning traces can therefore provide false assurance about how an output or action was produced. Their capture and retention may also create privacy, security, legal-privilege, and proprietary-information risks.

AIUC recommends that the final report harmonize these references by encouraging auditable execution records in place of reliance on internal reasoning: inputs and outputs; retrieved content; tool and API invocations with parameters and results; the identities and permissions exercised; policy-gate outcomes; human approvals, overrides, and exceptions; and the resulting actions and their consequences. Model-generated reasoning summaries may usefully supplement these records, but should be treated as unverified explanations rather than authoritative accounts of system behavior. This approach preserves the report's objectives — explainability, effective oversight, and reconstruction of incidents — while relying on records that are verifiable.

4. Are the sound practices sufficiently flexible to accommodate and address newer types of AI and responsible adoption over time?

Yes. The lifecycle, proportionality, and technology-neutral framing are sufficiently flexible. The report already calls for periodic and event-triggered reassessments of materiality and risk, change management, version control, rollback where available, repeated testing, monitoring of provider updates, more intensive assessment of dynamically changing systems, and restriction or retirement where risks cannot be controlled.

The final report would be stronger if it defined the concept of material change and linked it expressly to targeted reassessment.

Material-change triggers

A non-exhaustive list should include changes to:

- the underlying model, model version, model provider, fine-tuning, or safety settings;

- system prompts, policies, routing, retrieval sources, or guardrails;

- data sources, data rights, data sensitivity, or production-data distribution;

- tools, APIs, connected systems, code-execution capability, or external communications;

- identities, credentials, permissions, transaction authority, or approval thresholds;

- memory, persistence, state, autonomy, planning horizon, delegation, or multi-agent interaction;

- monitoring, logging, incident, recovery, or human-oversight controls;

- hosting, deployment environment, supply chain, or critical third party;

- scale, geography, customer population, business line, critical-operation status, or intended use; and

- the threat environment, a material incident or near miss, performance degradation, or a newly identified capability.

The institution should map a change to the affected risks, controls, tests, documentation, and approvals. Targeted reassessment is generally preferable to full reassessment where the unaffected boundary can be demonstrated. A change that moves the system outside its approved operating envelope should cause the institution to revisit the deployment decision and assurance conclusion before expanded use.

Evidence and status over time

The report could encourage an assurance record to identify:

- the in-scope configuration and approved operating envelope;

- the version of the assessment criteria;

- the date and scope of testing;

- open findings and conditions;

- monitoring obligations and reassessment triggers;

- the accountable owner; and

- whether the system is approved, conditionally approved, restricted, suspended, retired, or due for reassessment.

Machine-readable formats may improve reuse and comparison, but should remain an implementation option rather than a prescribed FSB outcome.

5. **Do the case studies in this report sufficiently highlight how different types of financial institutions can benefit from responsible AI adoption? Are there additional case studies for inclusion in the report? If so, please provide such case studies, particularly for nonbanks.**

The case studies are useful and demonstrate a range of benefits and governance approaches. The report includes an insurer operational-efficiency example and an insurance use case in the explainability annex. Nevertheless, the main operational cases are weighted toward large or internationally active banks and groups.

The final report would benefit from detailed risk-and-control examples involving a payment provider, fintech, asset manager, financial-market infrastructure, smaller institution, or outsourced service provider. At least one case should involve a third-party agent with production access to sensitive data, customer communications, operational systems, or financial transactions. Another should show a deployment that was restricted, delayed, or rejected because evidence or controls were insufficient.

AIUC is not submitting a specific case study for inclusion in the report at this time.

6. **Do the case studies in this report provide actionable insights for financial institutions in their responsible AI adoption?**

The case studies provide useful insights, and the report already includes concrete testing dimensions, ex ante thresholds, a no-deploy example, ongoing monitoring techniques, and annexed testing examples. Their actionability would improve if each followed a common structure linking the relevant risk, control objective, implementation, testing, findings, remediation, residual risk, and deployment decision. For cases reflecting an actual deployment, the evidence should demonstrate both how the relevant controls were designed and whether they operated effectively in the deployed environment. Depending on the control and use case, this may include approved procedures, system configurations, activity logs, demonstrations, test results, remediation records, and monitoring reports. The final report need not prescribe a fixed evidence taxonomy, but each case should clearly show what evidence supported the institution’s decision to deploy, restrict, or decline the use case.

Recommended case-study template

Each case should identify:

1. institution type and use case;

2. intended benefit and decision to be made;

3. deployed-system boundary and operating envelope;

4. materiality, inherent risk, and principal harm scenarios;

5. accountable owner and independent challenger;

6. control objective and control implementation;

7. evidence produced;

8. test method, data or scenario, and acceptance criterion;

9. findings and severity;

10. remediation, compensating controls, and retesting;

11. residual risk and approval, restriction, or no-deploy decision; and

12. production monitoring, incident triggers, and material-change reassessment.

This structure would make transparent the evidential link between a stated principle and a deployment decision. It would also make examples more comparable across banks and nonbanks without prescribing identical controls.

Hypothetical illustration

Risk: An agent attempts an unauthorized refund or invokes the wrong payment tool.

Control objective: The agent can use only approved tools, within the authority and transaction limits assigned to its identity and use case.

Implementation evidence: Tool authorization and validation configuration (AIUC-1 D003.1), rate or transaction limits (D003.2), tool-call records (D003.3), and, where selected as a supplemental control, human approval for sensitive actions (D003.4).

Independent test evidence: A tool-call assessment attempts unauthorized actions, restricted-data access, and decisions outside the intended scope (D004.1).

Decision evidence: The report records the acceptance criteria, findings, remediation and retest, residual risk, approval conditions, and monitoring threshold. The institution, not the assessor or tool provider, decides whether the deployment falls within risk appetite.

7. Are the definitions in the glossary clear and aligned with industry sound practices, including recent developments in AI?

The glossary is generally clear. It would benefit from distinguishing an individual AI agent from the broader agentic system in which it operates and from defining several concepts already used or implied in the body of the report. It should also identify common factors for assessing materiality and criticality while avoiding universal numerical thresholds. Relevant factors may include effects on critical operations, customers, legal obligations, transaction authority, sensitive-data access, reversibility, autonomy, interconnectedness, and common third-party dependencies. The final report should also distinguish open-source from open-weight models, consistent with the more nuanced discussion in the third-party case study.

Proposed definitions

AI agent. An AI-enabled software system or component that pursues an externally specified goal and, with a degree of autonomy, can select, sequence, and execute actions, including through tools or interaction with a virtual or physical environment.

Agentic system. The complete socio-technical system through which one or more AI agents operate. It may include models, orchestration, prompts and policies, data and retrieval, memory and state, tools and interfaces, identities and permissions, other agents, human oversight, and the deployment environment.

Basis for the distinction: The proposed definition of an AI agent is broadly consistent with NIST's description of agents as programs that interact with their environment and take self-directed actions toward an externally specified goal. Defining an agentic system separately reflects emerging international usage that treats an agent as a component of a broader socio-technical system rather than using the two terms interchangeably.

Autonomy. The degree to which an AI system can select and perform actions without contemporaneous human approval, considering the range, duration, reversibility, and potential consequences of those actions.

Materiality. The significance of an AI use case to a financial institution's viability, critical operations, or ability to meet key legal and regulatory obligations, including towards customers. Materiality and risk are distinct concepts, although they may be assessed within a single process.

Criticality. The extent to which the disruption, failure, misuse, or unavailability of an AI use case or shared dependency could materially impair critical operations, customer outcomes, markets, or the functioning or resilience of one or more financial institutions.

Operating envelope. The authorized goals, tasks, users, data, tools, systems, environments, counterparties, actions, limits, approval requirements, monitoring conditions, and escalation rules within which an AI use case is approved to operate.

Material change. A change reasonably capable of altering the materiality or inherent or residual risk of an AI use case, the effectiveness of a relevant control, the approved operating envelope, or a prior validation or assurance conclusion.

AI incident. An event in which the development, provision, or use of AI causes or materially contributes to customer or market harm, financial loss, legal or regulatory breach, data compromise, operational disruption, unauthorized action, or material control failure.

AI near miss. An event that could reasonably have resulted in an AI incident but did not because of timely detection or intervention, a functioning control, or other circumstances.

Open-weight model. A model for which trained parameters are made available under specified terms, but for which source code, training data, or other information required to reproduce or fully evaluate the model may not be available.

Open-source AI system. An AI system made available under terms that permit relevant use, study, modification, and redistribution, with access to the components and information required by the applicable open-source definition. The availability of model weights alone should not automatically be described as open source.

Deployment-specific assurance. An objective, evidence-based assessment by a competent function independent of the development and operation being assessed, evaluating whether relevant controls for a defined AI use case and deployed configuration are suitably designed and, where applicable, operating effectively within an approved operating envelope. Such assurance does not guarantee safety or transfer accountability from the financial institution.

Assurance and certification

The glossary should define deployment-specific assurance only if the final report substantively uses the concept. If included, the definition should preserve four elements: a defined scope; assessment by a competent and appropriately independent function; consideration of evidence, findings, and limitations; and continuing accountability of the financial institution.

“Certification” should not be treated as synonymous with assurance. Certification ordinarily indicates conformity with specified criteria, whereas deployment-specific assurance evaluates whether the controls applicable to a defined use case and operating environment are suitably designed and, where relevant, operating effectively. Neither term should imply a guarantee of safety or transfer accountability from the financial institution.

8. Are there any comments you would like to make on this report? Please fill in.

AIUC strongly supports the FSB's proposed sound practices. The executive summary of our response recommends three targeted amendments, which we reproduce here for completeness; these are the amendments referenced in our answer to Question 2.

AIUC discloses that it has a commercial interest in the adoption of third-party assurance; precisely for that reason, our recommendations are outcome-based and vendor-neutral. Certifications and external assessments should inform, and never substitute for, institutional and supervisory judgment; concentration in a small number of assurance providers or testing methodologies can itself become a source of correlated false confidence.

Suggested addition to Sound Practice 2 (Governance and accountability):

"For material or high-risk AI use cases, independent challenge should extend to performance testing and to evidence of control effectiveness supporting initial deployment and continued operation decisions, consistent with Sound Practice 9."

Suggested addition to Sound Practice 9 (Performance management):

"For AI use cases assessed as high materiality or high risk, performance testing and the effectiveness of relevant controls should be reviewed by appropriately skilled, functionally independent personnel who are separate from development and day-to-day operations. Depending on the institution’s capabilities and the nature of the use case, this review may be performed by an internal control function or a qualified external assessor.

The review should consider the configured use case in its deployment environment, including relevant data connections, tools, permissions, human-oversight arrangements, and operating limits. Its scope, evidence, material findings, limitations, residual risks, approval conditions, and reassessment triggers should be documented proportionally."

Suggested addition to Sound Practice 12 (Third-party AI risk management):

"Provider-level certifications, standards, attestations, model cards and system cards may inform due diligence, but do not by themselves demonstrate that a specific AI use case is appropriately configured and controlled in the financial institution’s operating environment. Financial institutions should apply proportionate use-case-level validation and, where appropriate, independent review. Such review supplements, and does not transfer or reduce, the financial institution’s accountability."

AIUC has additionally provided supplementary material by email to fsb@fsb.org: the complete formatted response, and the requirements and controls of the AIUC-1 standard (July 15, 2026 release), reproduced from AIUC's public registry.

AIUC

Response to the Financial Stability Board Consultation

Sound Practices for Responsible Adoption of Artificial Intelligence (AI)

RESPONDENT	Artificial Intelligence Underwriting Company (AIUC)
JURISDICTION	United States of America
CONTACT	Tom Fehring <tom@aiuc.com>
CONTRIBUTORS	Tom Fehring†, Tyler Pearson, Emil Bender Lassen†, Rajiv Dattani†
CONSULTATION DEADLINE	22 July 2026
DATE SUBMITTED	22 July 2026
PUBLICATION STATUS	Public

† *denotes AIUC affiliation*

AIUC response to the consultation

About AIUC and statement of interest

Artificial Intelligence Underwriting Company (AIUC) develops AIUC-1, a private assurance standard for the security, safety, reliability, accountability, data and privacy, and societal-risk controls of agentic AI systems. AIUC-1 applies to identified agentic systems and combines control evidence with technical evaluation. Under the current certification model, AIUC-1-accredited external auditors assess applicable controls; AIUC conducts recurring technical evaluations; and AIUC issues the certificate after reviewing the assessment materials.

AIUC-1 illustrates how selected principles can be translated into control activities and auditable evidence for deployed agentic systems. It is not a substitute for a financial institution's board governance, enterprise risk management framework, materiality assessment, model risk management, operational resilience program, legal obligations, or accountability for its own deployments.

AIUC discloses that it has a commercial interest in the adoption of third-party assurance; precisely for that reason, our recommendation is outcome-based and vendor-neutral. Where independent assurance is warranted, a financial institution should be able to use an appropriate framework or an equivalent method, provided that the scope, assessor competence, independence and conflicts of interest, testing methods, findings, limitations, and refresh conditions are transparent and commensurate with the risk. Certifications and external assessments should inform, and never substitute for, institutional and supervisory judgment; concentration in a small number of assurance providers or testing methodologies can itself become a source of correlated false confidence.

Executive summary

AIUC strongly supports the FSB's proposed sound practices. The report is comprehensive at the level appropriate for a cross-border, technology-neutral menu of practices. It correctly treats materiality and risk as distinct but related concepts; recognizes that agentic performance must be assessed by examining actions as well as outcomes; and identifies the limitations of provider-level transparency and assurance when a third-party model, system, or agent is integrated into a financial institution's environment.

The central opportunity is to make the existing principles more operational for material or high-risk deployments. The final report should:

1. **Clarify the assurance pathway.** State when independent challenge should culminate in a documented assurance conclusion, what evidence should support that conclusion, and when qualified external assessment is appropriate.
2. **Assess the deployed system.** Make clear that provider governance certifications and model-level evaluations are inputs to due diligence, but do not establish that a particular deployment is appropriately configured and controlled in its actual financial environment.
3. **Illustrate evidence outcomes.** Encourage traceability from materiality and risk to control objective, control implementation, test method, acceptance criterion, finding, remediation, residual risk, approval, and continued-use decision, while preserving proportionality.

4. **Prefer verifiable execution records to model-generated explanations.** Clarify that financial institutions should not treat verbatim chain-of-thought as necessary or authoritative evidence of system behaviour. For relevant use cases, oversight should rely on decision-relevant records, including inputs and outputs, retrieved sources, tool and API activity, identities and permissions, policy-gate outcomes, human interventions, and resulting actions and consequences. Model-generated reasoning summaries may supplement, but should not replace, these records.
5. **Connect monitoring to change-triggered reassessment.** Identify a non-exhaustive set of material changes that should cause the institution to revisit affected controls, testing, approval, and assurance status.
6. **Consolidate the agentic provisions.** Bring the report's extensive but dispersed agentic AI content into a short agentic deployment profile or annex covering system boundaries, identities, permissions, tools, action limits, approval points, logs, testing, monitoring, and intervention, and clarify that oversight should rely on auditable execution records rather than model-reported reasoning.

Recommended targeted amendments

The FSB report already establishes the principal components of an assurance framework. Sound Practice 2 addresses functional independence and effective challenge; Sound Practice 9 addresses proportionate performance assessment, testing and ongoing monitoring; and Sound Practice 12 addresses third-party certifications, standards, attestations and compensating controls.

AIUC therefore does not recommend creating an additional sound practice or prescribing a particular certification framework. Instead, we recommend three targeted amendments that connect the existing provisions and clarify the distinction between assurance over a provider's governance and review of a specific AI use case in its deployed environment.

Sound Practice 2: Governance and accountability

Suggested addition:

For material or high-risk AI use cases, independent challenge should extend to performance testing and to evidence of control effectiveness supporting initial deployment and continued operation decisions, consistent with Sound Practice 9.

This cross-reference would make clear that the independence principles in Sound Practice 2 apply not only to governance processes, but also to the evidence supporting consequential deployment decisions.

Sound Practice 9: Performance management

Suggested addition:

For AI use cases assessed as high materiality or high risk, performance testing and the effectiveness of relevant controls should be reviewed by appropriately skilled, functionally independent personnel who are separate from development and day-to-day operations. Depending on the institution's capabilities and the nature of the use case, this review may be performed by an internal control function or a qualified external assessor.

The review should consider the configured use case in its deployment environment, including relevant data connections, tools, permissions, human-oversight arrangements, and operating limits. Its scope, evidence,

material findings, limitations, residual risks, approval conditions, and reassessment triggers should be documented proportionally.

Placing this language in Sound Practice 9 would link testing to a documented deployment or continued-use decision. Independent review could be conducted internally where adequate independence and expertise exist, with external assessment used where the institution lacks the necessary competence, independence, or access.

Sound Practice 12: Third-party AI risk management

Suggested addition:

Provider-level certifications, standards, attestations, model cards and system cards may inform due diligence, but do not by themselves demonstrate that a specific AI use case is appropriately configured and controlled in the financial institution's operating environment. Financial institutions should apply proportionate use-case-level validation and, where appropriate, independent review. Such review supplements, and does not transfer or reduce, the financial institution's accountability.

This clarification would build directly on the report's observation that existing transparency and assurance tools may not show how a model, system or agent can be integrated safely into a financial institution's environment.

Deployment-specific review, however, addresses only one layer of third-party AI risk. AIUC recommends distinguishing two complementary layers. At the upstream and systemic layer, authorities and industry should work toward consistent, sector-relevant provider disclosures, change notifications, and pooled assurance mechanisms. At the deployment layer, each financial institution should assess its configured use case, document material information gaps and compensating controls, and remain accountable for customer and regulatory outcomes. Neither layer substitutes for the other.

Overall effect of the proposed amendments

Taken together, these amendments ensure that material or high-risk deployments are supported by appropriately independent review of the configured use case—not solely by evidence about the underlying model or provider. The detailed responses to the consultation questions below explain why these targeted refinements would improve the clarity and actionability of the FSB's existing framework.

1. Benefits and risks

> **Question 1: Do you agree with the benefits and risks of AI adoption by financial institutions described in this report? Are there any substantive benefits or risks not covered?**

Response

Yes. AIUC agrees with the principal benefits and risks described in the report. The FSB appropriately recognizes benefits in fraud detection, cyber defense, operational efficiency, customer service, risk management, and financial inclusion, while also addressing model and data risk, explainability, performance degradation, cyber and ICT risk, third-party dependence and concentration, correlated behavior, conduct risk, and the distinct risks of agentic systems.

We do not identify any top-level risk category that is omitted. The report already addresses common providers and correlated outcomes, unexpected provider changes, unauthorized actions, disruptions to connected systems, multi-agent interactions, financial-transaction controls, and incident and near-miss information sharing.

Although AIUC does not identify a missing top-level risk category, the final report could make two risks it already recognizes more measurable and actionable: common dependencies that may produce correlated outcomes, and the classification of incidents and near misses.

Common dependencies and correlated outcomes

The FSB separately identifies third-party concentration, common models and infrastructure, market correlations, and multi-agent interactions. The final report could more explicitly link those elements by encouraging institutions and authorities to monitor common exposures and plausible correlated outcomes. For an agentic deployment, relevant exposure indicators may include:

- the number and type of financial actions the system can initiate;
- maximum value and velocity of transactions within its authority;
- the volume and sensitivity of records it can access;
- external communications and counterparties it can reach;
- critical systems, tools, APIs, and providers on which it depends;
- the number of institutions or business lines using a common model, agent platform, data source, or orchestration layer; and
- the time required to detect, contain, reverse, or reconcile erroneous actions.

These measures would help translate the report's systemic-risk discussion into institution-level inventories and sector-level scenario analysis without creating a single universal metric.

Incident and near-miss taxonomy

Sound Practice 11 appropriately encourages sharing information about AI incidents, near misses, deteriorating performance, threats, and lessons learned. The value of that practice would increase if institutions used a common minimum taxonomy. The final report could encourage classification by:

- source of failure: model, data, configuration, tool, identity or access, human oversight, third party, or malicious use;
- failure mode: technical failure, control failure, misuse, unauthorized action, erroneous action, or inadequate escalation;
- consequence: customer harm, financial loss, data exposure, service disruption, regulatory breach, market impact, or avoided harm;
- detection channel and time to detection;
- affected deployment configuration and provider dependencies;
- control that prevented or limited loss; and
- corrective action and reassessment outcome.

Information-sharing mechanisms would need appropriate safeguards for confidentiality, personal data, legal privilege, cybersecurity, and supervisory information.

Recommendation

AIUC recommends retaining the report's risk taxonomy while adding: (i) common exposure indicators for highly autonomous or widely shared systems; and (ii) a minimum, internationally interoperable incident and near-miss taxonomy. These additions would operationalize risks already identified rather than create a separate category.

2. Comprehensiveness and clarity of the sound practices

> **Question 2: Are the sound practices sufficiently comprehensive and clear to enable financial institutions' responsible AI adoption?**

Response

The sound practices are sufficiently comprehensive and clear as principles. Their actionability would improve through illustrative, non-prescriptive examples of the evidence and decision records expected where an institution’s assessment of materiality and risk warrants heightened independent review.

The report already contains the foundations of an assurance model. Sound Practice 2 recognizes independent second-line assessment and third-line assurance proportionate to materiality and risk, and states that audit, evaluation, risk management, and testing should remain functionally separate from development. Sound Practice 12 discusses access and audit rights, independent and pooled audits, certifications, standards, and external attestations.

For deployments assessed by the institution as material or high risk, these elements should culminate in a documented, deployment-specific assurance conclusion. That conclusion should identify the assessed system and operating envelope; the evidence and testing considered; material findings and limitations; residual risks and approval conditions; the accountable decision-maker; and the events or changes that will require reassessment.

The review should be performed by appropriately skilled, functionally independent personnel. Qualified external assessment should be used where the institution lacks sufficient internal competence, independence, or access to relevant evidence. These requirements need not prescribe a particular certification framework, organizational model, or standardized evidence package, provided that the resulting assurance conclusion is transparent, evidence-based, proportionate to risk, and subject to appropriate refresh.

A proportionate assurance lifecycle

For AI use cases assessed as material or high risk, AIUC recommends that the final report encourage a proportionate assurance lifecycle culminating in a documented, deployment-specific assurance conclusion and an accountable deployment or continued-use decision. Institutions may allocate the relevant responsibilities through different governance and assurance structures, but the lifecycle should achieve the following outcomes:

1. **First-line assessment.** The deployment owner demonstrates the use case, intended benefit, system boundary, materiality, inherent risk, controls, testing, and residual risk.
2. **Independent second-line validation.** A function separate from development and operation challenges the risk classification, control design, test plan, thresholds, limitations, and deployment conditions.

3. **Third-line assurance.** Internal audit or an equivalent independent function assesses whether governance and controls are suitably designed and, where relevant, operating effectively.
4. **Qualified external assurance where warranted.** External assessment should be considered where internal independence or specialist competence is insufficient; the institution lacks the necessary access to external evidence but a qualified external assessor can obtain or evaluate relevant evidence not reasonably available internally; the system has substantial autonomy or authority; the deployment affects critical operations, customers, or markets; common dependencies create concentration concerns; or law, supervision, counterparties, or the institution's risk appetite call for it.

Illustrative evidence outcomes

For material or high-risk use cases, the final report should give illustrative examples of a proportionate evidence package, which may include:

- the defined use case, system boundary, and approved operating envelope;
- model and component versions; prompts and policies; retrieval sources; tools and APIs; identities and permissions; memory or state; delegation pathways; human-approval points; and deployment environment, as applicable;
- the materiality assessment and inherent- and residual-risk conclusions;
- intended, restricted, and prohibited uses;
- customer-outcome, conduct, fairness, explainability, contestability and redress testing, where relevant;
- control objectives, owners, implementation evidence, and dependencies;
- test methods, test conditions, data, benchmarks, acceptance criteria, limitations, and results;
- findings, severity, remediation, retesting, exceptions, and expiry dates;
- monitoring metrics, thresholds, incidents, near misses, overrides, and performance degradation;
- provider information, audit rights, change-notification obligations, continuity and exit arrangements; and
- approval, conditional approval, restriction, suspension, retirement, or continued-use decision.

This package should be tailored to the intended recipient. Boards, supervisors, auditors, counterparties, and technical operators need different levels of detail, but they benefit from a consistent underlying evidence model.

Suggested clarification across the sound practices

We recommend two supporting clarifications in addition to the targeted amendments proposed above:

- **Sound Practice 3:** For material or high-risk use cases, lifecycle documentation should identify the approved operating envelope, material assurance findings, residual risks, deployment conditions and reassessment triggers.
- **Sound Practice 5:** The final report should identify the factors that may warrant a documented independent review or assurance decision, while preserving proportionality and local supervisory judgment.

3. Balance between all AI and emerging forms such as GenAI and agentic AI

> **Question 3: Do the sound practices strike an appropriate balance between managing risks relating to all forms of AI, and addressing some of the risks relating to emerging and new complex forms of AI, such as GenAI and agentic AI?**

Response

Broadly yes. The technology-neutral structure is appropriate, and the report contains substantial agent-specific guidance. It recognizes unauthorized and erroneous action, multi-agent interaction, memory poisoning, identity and access, least privilege, tool access, transaction controls, human-approval boundaries, kill switches, action-level monitoring, red teaming, and the need to assess whether an agent's actions were appropriate, not merely whether it reached its objective.

Because the provisions relevant to agentic AI are distributed across several Sound Practices, AIUC recommends a brief, cross-referenced implementation profile or annex that explains how the existing practices apply to agentic deployments. This could cover agent and tool inventories, permissions and operating limits, action logging, testing of unauthorized or cascading actions, human-intervention mechanisms, reassessment triggers, and relevant third-party dependencies.

Recommended scope of an agentic implementation profile

An agentic assessment should define the complete system operating within the financial institution's environment. Depending on the use case, that boundary may include:

- models and model versions;
- system and developer prompts, policies, and guardrails;
- retrieval sources, context, memory, and persistent state;
- tools, APIs, code execution, external communications, and connected systems;
- identities, credentials, permissions, and delegated authority;
- sub-agents, agent-to-agent interaction, and orchestration;
- human-approval and escalation points;
- financial, temporal, rate, and volume limits;
- monitoring, incident response, safe-state, and recovery mechanisms; and
- the production environment and relevant third parties.

An assessment limited to the underlying foundation model cannot establish whether this configured deployment is appropriately controlled.

Consolidated agentic profile control outcomes

The final report could consolidate its existing content into the following outcome-oriented control objectives:

1. Every in-scope agent has a unique, attributable identity and no greater access than required for its approved task.
2. Tools and actions are allow-listed, validated, rate- or value-limited where appropriate, and logged.

3. Actions capable of material financial, legal, customer, market, or operational consequences should be subject to risk-appropriate preventive and detective controls. Depending on the approved operating envelope, these may include ex ante human approval, dual control, rate or value limits, automated monitoring, exception handling, and human escalation. Human accountability should remain clear even where monitoring is machine-assisted.
4. The system prevents or detects scope expansion, privilege escalation, unsafe delegation, prompt injection, and manipulation through external content or memory.
5. Monitoring evaluates the sequence and consequences of actions as well as final outputs.
6. Staff can pause, constrain, revoke, or transition the system to a safe state, and the effectiveness of that intervention is tested.
7. Pre-deployment and recurring tests cover adversarial behavior, unauthorized tool use, data leakage, harmful or out-of-scope outputs, unreliable actions, agent-to-agent interactions, and recovery.
8. Decision-relevant execution records support audit, incident investigation, contestability, and reconstruction without requiring disclosure of the private chain of thought.

Suggested clarification: execution records rather than chain-of-thought

The report refers to reliance on model reasoning in several places: chain-of-thought logging as enhanced documentation for GenAI use cases (§2.3); explainability approaches that trace the full reasoning path of agentic systems (§3.5); and monitoring that records the reasoning at each autonomous decision point (§3.7). Annex 2, however, correctly observes that a model's displayed reasoning may not reflect the steps it actually employed. Verbatim reasoning traces can therefore provide false assurance about how an output or action was produced. Their capture and retention may also create privacy, security, legal-privilege, and proprietary-information risks.

AIUC recommends that the final report harmonize these references by encouraging auditable execution records in place of reliance on internal reasoning: inputs and outputs; retrieved content; tool and API invocations with parameters and results; the identities and permissions exercised; policy-gate outcomes; human approvals, overrides, and exceptions; and the resulting actions and their consequences. Model-generated reasoning summaries may usefully supplement these records, but should be treated as unverified explanations rather than authoritative accounts of system behavior. This approach preserves the report's objectives — explainability, effective oversight, and reconstruction of incidents — while relying on records that are verifiable.

4. Flexibility for newer forms of AI

> **Question 4: Are the sound practices sufficiently flexible to accommodate and address newer types of AI and responsible adoption over time?**

Response

Yes. The lifecycle, proportionality, and technology-neutral framing are sufficiently flexible. The report already calls for periodic and event-triggered reassessments of materiality and risk, change management, version control, rollback where available, repeated testing, monitoring of provider updates, more intensive assessment of dynamically changing systems, and restriction or retirement where risks cannot be controlled.

The final report would be stronger if it defined the concept of material change and linked it expressly to targeted reassessment.

Material-change triggers

A non-exhaustive list should include changes to:

- the underlying model, model version, model provider, fine-tuning, or safety settings;
- system prompts, policies, routing, retrieval sources, or guardrails;
- data sources, data rights, data sensitivity, or production-data distribution;
- tools, APIs, connected systems, code-execution capability, or external communications;
- identities, credentials, permissions, transaction authority, or approval thresholds;
- memory, persistence, state, autonomy, planning horizon, delegation, or multi-agent interaction;
- monitoring, logging, incident, recovery, or human-oversight controls;
- hosting, deployment environment, supply chain, or critical third party;
- scale, geography, customer population, business line, critical-operation status, or intended use; and
- the threat environment, a material incident or near miss, performance degradation, or a newly identified capability.

The institution should map a change to the affected risks, controls, tests, documentation, and approvals. Targeted reassessment is generally preferable to full reassessment where the unaffected boundary can be demonstrated. A change that moves the system outside its approved operating envelope should cause the institution to revisit the deployment decision and assurance conclusion before expanded use.

Evidence and status over time

The report could encourage an assurance record to identify:

- the in-scope configuration and approved operating envelope;
- the version of the assessment criteria;
- the date and scope of testing;
- open findings and conditions;
- monitoring obligations and reassessment triggers;
- the accountable owner; and
- whether the system is approved, conditionally approved, restricted, suspended, retired, or due for reassessment.

Machine-readable formats may improve reuse and comparison, but should remain an implementation option rather than a prescribed FSB outcome.

5. Breadth of the case studies

> **Question 5: Do the case studies in this report sufficiently highlight how different types of financial institutions can benefit from responsible AI adoption? Are there additional case studies for inclusion in the report? If so, please provide such case studies, particularly for nonbanks.**

Response

The case studies are useful and demonstrate a range of benefits and governance approaches. The report includes an insurer operational-efficiency example and an insurance use case in the explainability annex. Nevertheless, the main operational cases are weighted toward large or internationally active banks and groups.

The final report would benefit from detailed risk-and-control examples involving a payment provider, fintech, asset manager, financial-market infrastructure, smaller institution, or outsourced service provider. At least one case should involve a third-party agent with production access to sensitive data, customer communications, operational systems, or financial transactions. Another should show a deployment that was restricted, delayed, or rejected because evidence or controls were insufficient.

AIUC is not submitting a specific case study for inclusion in the report at this time.

6. Actionability of the case studies

> **Question 6: Do the case studies in this report provide actionable insights for financial institutions in their responsible AI adoption?**

Response

The case studies provide useful insights, and the report already includes concrete testing dimensions, ex ante thresholds, a no-deploy example, ongoing monitoring techniques, and annexed testing examples. Their actionability would improve if each followed a common structure linking the relevant risk, control objective, implementation, testing, findings, remediation, residual risk, and deployment decision. For cases reflecting an actual deployment, the evidence should demonstrate both how the relevant controls were designed and whether they operated effectively in the deployed environment. Depending on the control and use case, this may include approved procedures, system configurations, activity logs, demonstrations, test results, remediation records, and monitoring reports. The final report need not prescribe a fixed evidence taxonomy, but each case should clearly show what evidence supported the institution's decision to deploy, restrict, or decline the use case.

Recommended case-study template

Each case should identify:

1. institution type and use case;
2. intended benefit and decision to be made;
3. deployed-system boundary and operating envelope;
4. materiality, inherent risk, and principal harm scenarios;
5. accountable owner and independent challenger;
6. control objective and control implementation;
7. evidence produced;
8. test method, data or scenario, and acceptance criterion;
9. findings and severity;
10. remediation, compensating controls, and retesting;
11. residual risk and approval, restriction, or no-deploy decision; and
12. production monitoring, incident triggers, and material-change reassessment.

This structure would make transparent the evidential link between a stated principle and a deployment decision. It would also make examples more comparable across banks and nonbanks without prescribing identical controls.

Hypothetical illustration

Risk: An agent attempts an unauthorized refund or invokes the wrong payment tool.

Control objective: The agent can use only approved tools, within the authority and transaction limits assigned to its identity and use case.

Implementation evidence: Tool authorization and validation configuration (AIUC-1 D003.1), rate or transaction limits (D003.2), tool-call records (D003.3), and, where selected as a supplemental control, human approval for sensitive actions (D003.4).

Independent test evidence: A tool-call assessment attempts unauthorized actions, restricted-data access, and decisions outside the intended scope (D004.1).

Decision evidence: The report records the acceptance criteria, findings, remediation and retest, residual risk, approval conditions, and monitoring threshold. The institution, not the assessor or tool provider, decides whether the deployment falls within risk appetite.

7. Glossary definitions

> **Question 7: Are the definitions in the glossary clear and aligned with industry sound practices, including recent developments in AI?**

Response

The glossary is generally clear. It would benefit from distinguishing an individual AI agent from the broader agentic system in which it operates and from defining several concepts already used or implied in the body of the report. It should also identify common factors for assessing materiality and criticality while avoiding universal numerical thresholds. Relevant factors may include effects on critical operations, customers, legal obligations, transaction authority, sensitive-data access, reversibility, autonomy, interconnectedness, and common third-party dependencies. The final report should also distinguish open-source from open-weight models, consistent with the more nuanced discussion in the third-party case study.

Proposed definitions

AI agent. An AI-enabled software system or component that pursues an externally specified goal and, with a degree of autonomy, can select, sequence, and execute actions, including through tools or interaction with a virtual or physical environment.

Agentic system. The complete socio-technical system through which one or more AI agents operate. It may include models, orchestration, prompts and policies, data and retrieval, memory and state, tools and interfaces, identities and permissions, other agents, human oversight, and the deployment environment.

> **Basis for the distinction:** The proposed definition of an AI agent is broadly consistent with NIST's description of agents as programs that interact with their environment and take self-directed actions toward an externally specified goal. Defining an agentic system separately reflects emerging international usage that treats an agent as a component of a broader socio-technical system rather than using the two terms interchangeably.

Autonomy. The degree to which an AI system can select and perform actions without contemporaneous human approval, considering the range, duration, reversibility, and potential consequences of those actions.

Materiality. The significance of an AI use case to a financial institution's viability, critical operations, or ability to meet key legal and regulatory obligations, including towards customers. Materiality and risk are distinct concepts, although they may be assessed within a single process.

Criticality. The extent to which the disruption, failure, misuse, or unavailability of an AI use case or shared dependency could materially impair critical operations, customer outcomes, markets, or the functioning or resilience of one or more financial institutions.

Operating envelope. The authorized goals, tasks, users, data, tools, systems, environments, counterparties, actions, limits, approval requirements, monitoring conditions, and escalation rules within which an AI use case is approved to operate.

Material change. A change reasonably capable of altering the materiality or inherent or residual risk of an AI use case, the effectiveness of a relevant control, the approved operating envelope, or a prior validation or assurance conclusion.

AI incident. An event in which the development, provision, or use of AI causes or materially contributes to customer or market harm, financial loss, legal or regulatory breach, data compromise, operational disruption, unauthorized action, or material control failure.

AI near miss. An event that could reasonably have resulted in an AI incident but did not because of timely detection or intervention, a functioning control, or other circumstances.

Open-weight model. A model for which trained parameters are made available under specified terms, but for which source code, training data, or other information required to reproduce or fully evaluate the model may not be available.

Open-source AI system. An AI system made available under terms that permit relevant use, study, modification, and redistribution, with access to the components and information required by the applicable open-source definition. The availability of model weights alone should not automatically be described as open source.

Deployment-specific assurance. An objective, evidence-based assessment by a competent function independent of the development and operation being assessed, evaluating whether relevant controls for a defined AI use case and deployed configuration are suitably designed and, where applicable, operating effectively within an approved operating envelope. Such assurance does not guarantee safety or transfer accountability from the financial institution.

Assurance and certification

The glossary should define deployment-specific assurance only if the final report substantively uses the concept. If included, the definition should preserve four elements: a defined scope; assessment by a competent and appropriately independent function; consideration of evidence, findings, and limitations; and continuing accountability of the financial institution.

“Certification” should not be treated as synonymous with assurance. Certification ordinarily indicates conformity with specified criteria, whereas deployment-specific assurance evaluates whether the controls applicable to a defined use case and operating environment are suitably designed and, where relevant, operating effectively. Neither term should imply a guarantee of safety or transfer accountability from the financial institution.

Conclusion

The FSB has assembled a thoughtful and detailed account of responsible AI adoption in financial services. Its treatment of agentic AI, materiality, testing, human oversight, third-party risk, and information sharing provides a strong foundation.

The most valuable refinement would be to integrate those components into a proportionate assurance lifecycle for material or high-risk deployments. A financial institution should be able to demonstrate which system was assessed, which operating envelope was approved, which risks and controls were evaluated, which evidence and thresholds were used, which limitations and residual risks remain, and which changes require renewed assessment. Provider certifications can contribute to that record, but they cannot substitute for evaluating the configured deployment in its actual financial environment.

AIUC would welcome the opportunity for further engagement on outcome-based, framework-neutral implementation materials that preserve institutional accountability.

Appendix A. Illustrative mapping of AIUC-1 to the FSB sound practices

This appendix identifies selected AIUC-1 controls that illustrate how portions of the FSB practices can be expressed as auditable evidence.

Sound Practice 1: Strategic direction and oversight

Illustrative AIUC-1 evidence: risk taxonomy and review (C001.1-C001.2); change approvals (E004.1); internal review (E008.1); acceptable-use policy (E010.1); regulatory compliance review (E012.1); and quality objectives and risk management when optional E013 is selected (E013.1).

Sound Practice 2: Governance and accountability

Illustrative AIUC-1 evidence: change-approval policy and records (E004.1); failure plans (E001.1-E003.1); internal review (E008.1); logging and storage (E015.1 and E015.3); and optional quality-management evidence (E013.2-E013.3).

Sound Practice 3: Incorporation of AI risks into risk management

Illustrative AIUC-1 evidence: risk taxonomy (C001.1-C001.2); pre-deployment testing and approval (C002.1); change approval (E004.1); logging (E015.1 and E015.3); and, when selected, AI risk monitoring and system-transparency documentation (C008.1 and E017.1).

Sound Practice 4: Organisational adaptability

Illustrative AIUC-1 evidence: risk-taxonomy review (C001.2); internal review (E008.1); regulatory-compliance review (E012.1); change approval (E004.1); recurring technical evaluation; and annual reassessment.

Sound Practice 5: Materiality and risk assessment

Illustrative AIUC-1 evidence: risk taxonomy and review (C001.1-C001.2); pre-deployment testing and approval (C002.1); agent-specific risk detection and response (C005.1); and scoping by capabilities, architecture, and deployment context.

Sound Practice 6: Selection

Illustrative AIUC-1 evidence: vendor due diligence (E006.1); pre-deployment testing and approval (C002.1); model-selection change approval (E004.1); data-storage assessment (E005.1); regulatory review (E012.1); and provider IP protections (A007.1).

Sound Practice 7: Data governance

Illustrative AIUC-1 evidence: input and output data policies (A001.1-A002.2); data-access scoping (A003.1); confidentiality and customer isolation (A004-A006); IP and secrets controls (A007-A008); storage security (E005.1); processing locations (E011.1-E011.2); and log storage and integrity (E015.3-E015.4).

Sound Practice 8: Explainability and transparency

Illustrative AIUC-1 evidence: citations and source attribution (D001.2); uncertainty labels (D001.3); activity and agent-execution logging (E015.1-E015.2); AI interaction and content disclosures (E016.1-E016.5); and optional model cards, system documentation, and shared-responsibility evidence (E017.1-E017.3).

Sound Practice 9: Performance management

Illustrative AIUC-1 evidence: pre-deployment test records (C002.1); adversarial testing (B001.1); external tests for harmful output, scope, customer-defined risk, hallucinations, and tool calls (C010.1-C012.1, D002.1, and D004.1); groundedness safeguards (D001.1); and relevant activity logs (E015.1 and E015.3).

Sound Practice 10: Human oversight

Illustrative AIUC-1 evidence: high-risk-output detection and review when optional C007 is selected; intervention mechanisms when optional C009 is selected; supplemental human-review workflows (C005.2 and C007.3); human approval for sensitive tools (D003.4); and execution records (E015.2).

Sound Practice 11: Cyber and ICT risk management

Illustrative AIUC-1 evidence: adversarial testing (B001.1); endpoint protection and penetration testing (B004.1-B004.4); service restrictions and monitoring (B006.1-B006.2); access control (B007.1-B007.2); interface authentication and transport security (B008.1-B008.2); secure generated-code controls (B010); secrets controls (A008); output sanitization (C006); tool controls and testing (D003-D004); breach planning (E001.1); and logging (E015).

Sound Practice 12: Third-party AI risk management

Illustrative AIUC-1 evidence: vendor due diligence (E006.1); third-party access monitoring (E009.1-E009.2); processing locations (E011); optional platform and deployer responsibility documentation (E017.3); and external evaluation of the deployed system (B001.1, C010.1-C012.1, D002.1, and D004.1).

Appendix B. Sources

1. Financial Stability Board, *Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation report*, 10 June 2026, https://www.fsb.org/uploads/P100626.pdf.
2. AIUC-1, *Full Evidence List*, current public evidence catalog, https://www.aiuc-1.com/evidence.
3. AIUC-1, *Changelog*, 15 July 2026 release, https://www.aiuc-1.com/changelog.
4. AIUC-1, *Scoping the AIUC-1 audit*, https://www.aiuc-1.com/scoping.
5. AIUC-1, *About the AIUC-1 certificate*, https://www.aiuc-1.com/learn/certificate.
6. AIUC-1, *Accredited AIUC-1 auditors*, https://www.aiuc-1.com/accredited-auditors.
7. AIUC-1, *Re-certification: Maintaining AIUC-1*, https://www.aiuc-1.com/re-certification-maintaining-aiuc-1.
8. National Institute of Standards and Technology, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2025, March 2025, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf.
9. National Institute of Standards and Technology, *AI Agent Standards Initiative*, updated 20 April 2026, https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative.
10. OECD.AI, *Can we create a clear understanding of what agentic AI is and does?*, 3 March 2026, https://oecd.ai/en/wonk/what-agentic-ai-is-and-does.

AIUC-1

The Standard for AI Agent Security, Safety, and Reliability
Requirements and Controls — July 15, 2026 release

About this document. This document reproduces the requirements and controls of the AIUC-1 standard (July 15, 2026 release) from AIUC's public registry, and is provided as supplemental material to AIUC's response to the Financial Stability Board's consultation on *Sound Practices for Responsible Adoption of Artificial Intelligence* (submitted 22 July 2026). The authoritative, current registry — including evidence items, mandatory and supplemental designations, and requirement applicability under the AIUC-1 scoping methodology — is maintained at **aiuc-1.com/evidence**, with versioned change history at **github.com/aiunderwriting/AIUC-1-Changelog** and **aiuc-1.com/changelog**. Contact: Tom Fehring, AIUC — tom@aiuc.com.

Part 1. Requirements

A001: Establish input data policy

Establish and communicate AI input data policies covering how customer data is used for model training, inference processing, data retention periods, and customer data rights.

A002: Establish output data policy

Establish AI output ownership, usage, opt-in/out and deletion policies to customers and communicate these policies.

A003: Limit AI agent data access

Implement safeguards to limit AI agent data access based on task, user role, agent role and context.

A004: Protect IP & trade secrets

Implement safeguards or technical controls to prevent AI systems from leaking company intellectual property or confidential information.

A005: Prevent cross-customer data exposure

Implement safeguards to prevent cross-customer data exposure.

A006: Prevent PII leakage

Establish safeguards to prevent personal data leakage through AI outputs and logs.

A007: Prevent IP violations

Implement safeguards and technical controls to prevent AI outputs from violating copyrights, trademarks, or other third-party intellectual property rights.

A008: Prevent leakage of credentials and secrets

Implement safeguards to detect and prevent leakage of secrets in AI system inputs, outputs, logs, and credential storage.

B001: Third-party testing of adversarial robustness

Implement adversarial testing program to validate system resilience against adversarial inputs and prompt injection attempts in line with adversarial threat taxonomy.

B002: Detect adversarial input

Implement monitoring capabilities to detect and enable responding to adversarial inputs and prompt injection attempts.

B003: Manage public release of technical details

Implement controls to prevent over-disclosure of technical information about AI systems and organizational details that could enable adversarial targeting.

B004: Prevent AI endpoint scraping

Implement safeguards to prevent probing or scraping of external AI endpoints.

B005: Implement real-time input filtering

Implement real-time input filtering using automated moderation tools.

B006: Prevent unauthorized AI agent actions

Implement safeguards to prevent AI agents from performing actions beyond intended scope and authorized privileges.

B007: Enforce user access privileges to AI systems

Establish and maintain user access controls and admin privileges for AI systems in line with policy.

B008: Protect AI system deployment environment

Implement security measures for AI system deployment environments including encryption, access controls and authorization.

B009: Limit output over-exposure

Implement output limitations and obfuscation techniques to safeguard against information leakage.

B010: Promote secure patterns in generated code

Implement safeguards to promote secure patterns and prevent known vulnerabilities in generated code.

C001: Define AI risk taxonomy

Establish a risk taxonomy based on system capabilities and deployment context.

C002: Conduct pre-deployment testing

Conduct internal testing of AI systems prior to deployment across risk categories for system changes requiring formal review or approval.

C003: Prevent harmful outputs

Implement safeguards or technical controls to prevent harmful outputs including distressed outputs, angry responses, high-risk advice, offensive content, bias, and deception.

C004: Prevent out-of-scope outputs

Implement safeguards or technical controls to prevent out-of-scope outputs (e.g. political discussion, healthcare advice).

C005: Prevent agent-specific high risk outputs

Implement safeguards or technical controls to prevent agent-specific high-risk outputs as defined in risk taxonomy.

C006: Prevent output vulnerabilities

Implement safeguards to prevent security vulnerabilities in outputs from impacting users.

C007: Flag high risk outputs for human review

Implement an alerting system that flags high-risk outputs for human review.

C008: Monitor AI risk categories

Implement monitoring of AI systems across risk categories.

C009: Enable real-time feedback and intervention

Implement mechanisms to enable real-time user feedback collection, intervention and actioning mechanisms.

C010: Third-party testing for harmful outputs

Appoint expert third parties to evaluate system robustness to harmful outputs including distressed outputs, angry responses, high-risk advice, offensive content, bias, and deception at least every 3 months.

C011: Third-party testing for out-of-scope outputs

Appoint expert third parties to evaluate system robustness to out-of-scope outputs at least every 3 months (e.g. political discussion, healthcare advice).

C012: Third-party testing for customer-defined risk

Appoint expert third-parties to evaluate system robustness to additional high-risk outputs as defined in risk taxonomy at least every 3 months.

D001: Prevent hallucinated outputs

Implement safeguards or technical controls to prevent hallucinated outputs.

D002: Third-party testing for hallucinations

Appoint expert third-parties to evaluate hallucinated outputs at least every 3 months.

D003: Restrict unsafe tool calls

Implement safeguards or technical controls to prevent tool calls in AI systems from executing unauthorized actions, accessing restricted information, or making decisions beyond their intended scope.

D004: Third-party testing of tool calls

Appoint expert third-parties to evaluate tool calls in AI systems, including executing unauthorized actions, accessing restricted information, or making decisions beyond their intended scope at least every 3 months.

E001: AI failure plan for security breaches

Document AI failure plan for AI privacy and security breaches assigning accountable owners and establishing notification and remediation with third-party support as needed (e.g. legal, PR, insurers).

E002: AI failure plan for harmful outputs

Document AI failure plan for harmful AI outputs that cause significant customer harm assigning accountable owners and establishing remediation with third-party support as needed (e.g. legal, PR, insurers).

E003: AI failure plan for hallucinations

Document AI failure plan for hallucinated AI outputs that cause substantial customer financial loss assigning accountable owners and establishing remediation with third-party support as needed (e.g. legal, PR, insurers).

E004: Assign accountability

Document which AI system changes across the development & deployment lifecycle require formal review or approval, assign a lead accountable for each, and document their approval with supporting evidence.

E005: Document data storage security

Document data storage security practices considering data sensitivity, regulatory requirements, security controls, and operational needs.

E006: Conduct vendor due diligence

Establish AI vendor due diligence processes for foundation and upstream model providers covering data handling, PII controls, security and compliance.

E007: [Retired] Document system change approvals

Merged with E004 - see changelog (Q1 2026 update).

E008: Review internal processes

Establish regular internal reviews of key processes and document review records and approvals.

E009: Monitor third-party access

Implement systems to monitor and log third-party API connections, sessions, and data access.

E010: Establish AI acceptable use policy

Establish and implement an AI acceptable use policy.

E011: Record processing locations

Document AI data processing locations.

E012: Document regulatory compliance

Document applicable AI laws and standards, required data protections, and strategies for compliance.

E013: Implement quality management system

Establish a quality management system for AI systems proportionate to the size of the organization.

E014: [Retired] Share transparency reports

Merged with E017 - see changelog (Q1 2026 update).

E015: Log AI system activity

Maintain logs of AI system processes, actions, and agent outputs where permitted to support incident investigation, auditing, and explanation of AI system behavior.

E016: Implement AI disclosure mechanisms

Implement clear disclosure mechanisms to inform users when they are interacting with AI systems rather than humans.

E017: Document system transparency policy

Establish a system transparency policy and maintain a repository of model cards, datasheets, and interpretability reports for major systems.

F001: Prevent AI cyber misuse

Implement or document guardrails to prevent AI-enabled misuse for cyber attacks and exploitation.

F002: Prevent catastrophic misuse

Implement or document guardrails to prevent AI-enabled catastrophic system misuse (chemical / bio / radio / nuclear).

Part 2. Controls

A001: Establish input data policy

Control shoulds

- Defining and communicating input data usage policies. Including specifying how customer data is used for inference and model training, establishing data retention periods, and documenting customer data rights.
- Implementing technical controls to enforce data retention and deletion policies. For example, automating data deletion based on retention schedules, using secure removal mechanisms, and managing data lifecycles.

Control mays

- Documenting processes for handling end-user data subject rights. For example, handling requests for opt-in/opt-out rights, access, portability, or deletion of input data.

A002: Establish output data policy

Control shoulds

- Establishing output ownership and usage rights policies. For example, specifying customer ownership of AI-generated outputs versus AI inputs, defining permitted uses of outputs (commercial use, redistribution, modification), documenting usage restrictions or limitations, and clarifying how ownership applies to different output types or use cases.
- Disclosing opt-in/opt-out and deletion policies for AI outputs. For example, documenting how customers can opt out of output storage or reuse, explaining deletion request processes, specifying retention periods and data handling practices, and clarifying how customers can control or revoke permissions for their outputs.

Control mays

- Implementing technical controls to enforce AI output opt-in/opt-out and deletion policies. For example, automating customer preference enforcement through consent management configuration, processing opt-out and deletion requests within defined workflows, and validating that opted-out outputs are excluded from storage and downstream reuse.

A003: Limit AI agent data access

Control shoulds

- Configuring data access limits to reduce data and privacy exposure. For example, limiting data access to task-relevant information based on context, implementing scoping based on user roles or workflow requirements, and avoiding persistent or out-of-scope data access.

Control mays

- Enabling agent identity management. For example, assigning each agent a unique, cryptographically verifiable identity; supporting standard identity federation protocols (e.g., OAuth 2.0, OIDC) for enterprise IAM integration; publishing agent cards declaring each agent's capabilities, tools, and permission scopes.
- Enabling agent access and governance through permission-ready architecture. For example, exposing per-agent permission scopes mappable to enterprise roles; supporting just-in-time permissions scoped to specific subtasks; preventing silent inheritance of elevated permissions from orchestrators or parent agents; enforcing segregation of duties across systems; and integrating data loss prevention controls on agent actions and tool calls.

A004: Protect IP & trade secrets

Control shoulds

- Providing user guidance on protecting confidential information. For example, instructing employees not to input trade secrets, proprietary code, or confidential business information into AI systems, communicating data handling policies for AI tool usage, or establishing clear guidelines on what information can and cannot be shared with AI agents.

Control mays

- Leveraging foundation model provider protections. For example, using providers with zero data retention policies, requiring contractual commitments that inputs are not used for training, selecting models with enhanced privacy guarantees for sensitive use cases.
- Implementing technical controls to detect proprietary information in outputs.
- Establishing output monitoring for high-risk IP scenarios. For example, logging AI responses that accessed confidential data sources, implementing human review workflows for outputs flagged as potentially containing sensitive information.

A005: Prevent cross-customer data exposure

Control shoulds

- Establishing explicit consent and disclosure for combined data usage. For example, informing customers when their data will be combined with competitor data, disclosing data anonymization and abstraction policies, providing opt-out mechanisms.
- Implementing customer data isolation controls. For example, enforcing strict logical and physical separation of customer data, applying tenant-specific encryption, validating data flow boundaries in shared infrastructure, establishing technical barriers between customer datasets during training.

Control mays

- Implementing specific privacy-enhancing technologies (PETs) to reduce competitive exposure.

A006: Prevent PII leakage

Control shoulds

- Implementing safeguards to prevent personal data leakage through AI system outputs and logs. For example, filtering prompts and outputs for personal identifiers before storage or display, implementing automated PII detection and redaction in system logs, preventing retention of outputs containing sensitive personal information, or blocking responses that would expose personal identifiers.

Control mays

- Integrating with existing data loss prevention (DLP) systems to monitor and block outputs containing personal data in violation of policy.

A007: Prevent IP violations

Control shoulds

- Documenting foundation model provider IP protections which may serve as primary infringement safeguards. For example, indemnification clauses or copyright/trademark guardrails.

Control mays

- Establishing supplementary content filtering mechanisms where provider protections have gaps or limitations. For example, detecting copyrighted material in outputs, implementing trademark screening.
- Implementing user guidance and guardrails to reduce IP risk. For example, usage policies that explain prohibited content types, user warnings in product, restricting output generation in known infringement domains.
- Implementing restrictions in AI acceptable use policy.

A008: Prevent leakage of credentials and secrets

Control shoulds

- Implementing safeguards to detect credentials in user inputs. For example, scanning user prompts and pasted content for API keys, access tokens, private keys, and connection strings using pattern-matching or entropy-based detection; defining handling procedures when detected such as warning the user, refusing to persist, or flagging to the deployer.
- Implementing safeguards to keep secrets out of generated code artifacts. For example, guiding the model via system prompts to reference environment variables or secret-management tools rather than hardcoding credentials, post-generation scanning of output files for common credential patterns, or blocking/flagging generations that contain detected secrets before they are written to disk.
- Implementing safeguards to securely store user-provided credentials to enable agent-connected services. For example, storing OAuth tokens, API keys, and connection strings in a dedicated secret manager, encrypting credentials at rest, scoping credential access per tool and per session, or fetching credentials only at the point of tool invocation rather than persisting them in the agent's context window.

Control mays

- Implementing user-facing warnings when potential secrets are detected in user inputs. For example, alerting users when credentials are detected in their prompts.
- Implementing safeguards to prevent secrets from being retained in platform logs, conversation history, and stored artifacts. For example, redacting detected credential patterns before log storage, masking secrets in conversation history displayed to users or support staff, or applying scrubbing functions to prompts and outputs before persistence.

B001: Third-party testing of adversarial robustness

Control shoulds

- Establishing a taxonomy for adversarial risks. For example, drawing on NIST's AI 100-2e2023 attack classifications and aligning these to system architecture and use cases.
- Conducting comprehensive adversarial testing at least quarterly. For example, performing structured red-teaming, prompt injection assessments, jailbreaking attempts, adversarial perturbation testing, semantic manipulation, and simulated malicious tool

invocations.
- Maintaining secure testing documentation. For example, recording test cases, methods, outcomes, and system behaviors with restricted access controls, implementing secure storage for sensitive testing materials.
- Establishing improvement processes based on findings. For example, assigning owners and remediation timelines based on test severity, tracking fixes through risk registers or issue management systems, documenting updates to safeguards and procedures.

Control mays

- Aligning adversarial testing with broader security testing programs. For example, integrating AI-specific test cases into broader penetration testing, sharing threat models across red/blue teams, aligning test cycles with security audit and compliance calendars.

B002: Detect adversarial input

Control shoulds

- Establishing detection and alerting. For example, implementing monitoring for prompt injection patterns, jailbreak techniques, adversarial input attempts, and exceeding rate limits, configuring alerts and threat notifications for suspicious activities.
- Implementing incident logging and response procedures. For example, logging suspected adversarial attacks with relevant context, escalating to designated personnel based on severity, and documenting response actions in a centralized system.
- Maintaining detection effectiveness through quarterly reviews. For example, updating detection rules based on emerging adversarial techniques, analyzing incident patterns and documenting system improvements.

Control mays

- Implementing adversarial input detection prior to AI model processing where feasible. For example, using pre-processing filters to flag likely threats before model processing.
- Integrating adversarial input detection into existing security operations tooling. For example, forwarding flagged inputs to SIEM platforms, correlating detection with authentication and network logs, enabling SOC teams to triage AI-related security events.

B003: Manage public release of technical details

Control shoulds

- Documenting limitations on technical information release. For example, limiting public disclosure of model architectures, algorithms, training data details, system configurations, and performance metrics, requiring approval before sharing technical specifications or implementation details.
- Controlling organizational information to balance transparency with security. For example, limiting disclosure of AI team details, development timelines, and other information that could reveal technical capabilities, reviewing public communications for sensitive information.

Control mays

- Establishing approval processes. For example, requiring designated review for public content referencing AI capabilities in e.g. publications, presentations, and marketing materials, and documenting approved disclosures with business justification.

B004: Prevent AI endpoint scraping

Control shoulds

- Implementing systems distinguishing between high-volume legitimate usage and adversarial behavior. For example, using behavioral analytics and user profiling to calibrate detection thresholds and prevent false positives against trusted users.
- Implementing rate limiting and query restrictions. For example, establishing per-user quotas to prevent model extraction, blocking excessive query patterns, implementing progressive restrictions for suspicious behavior, or using economic disincentives for high-volume usage.
- Conducting simulated external attack testing of AI endpoints. For example, performing automated attack simulations, testing endpoint protection effectiveness against high-volume and distributed attacks, and documenting methodologies appropriate to organizational threat profile.
- Maintaining endpoint security through remediation. For example, tracking identified vulnerabilities, implementing protective measures based on testing outcomes, and regularly updating endpoint defenses and detection thresholds.

B005: Implement real-time input filtering

Control shoulds

- Integrating automated moderation tools to filter inputs before they reach the foundation model. For example, integrating third-party moderation APIs, implementing custom filtering rules, configuring blocking or warning actions for flagged content, and establishing confidence thresholds based on risk category and severity.

Control mays

- Documenting the moderation logic and rationale. For example, explaining chosen moderation tools, threshold justifications, and decision criteria for different risk categories.
- Providing feedback to users when inputs are blocked.

- Logging flagged prompts for analysis and refinement of filters, while ensuring compliance with privacy obligations.
- Periodically evaluating filter performance and adjusting thresholds accordingly. For example, accuracy, latency, false positives/negatives.

B006: Prevent unauthorized AI agent actions

Control shoulds

- Implementing technical restrictions that limit agent capabilities to authorized scope. For example, restricting agent access to approved backend services, APIs and MCP servers, enforcing network segmentation or API gateway rules, or implementing service-level authorization preventing access to sensitive systems.
- Deploying monitoring and alerting for agent actions that exceed security boundaries. For example, logging all agent service interactions, alerting on access attempts to unauthorized systems or APIs, or anomaly detection flagging unusual connection patterns.

Control mays

- Implementing additional safeguards to contain runtime risk. For example, enabling sandboxed execution environments with configurable filesystem, network, and credential restrictions for agent-executed code and first-party MCP servers, monitoring MCP tool definitions for unauthorized changes after initial approval, providing pre-execution authorization hooks that verify runtime tool calls against defined policy before execution proceeds, or scanning agent configuration artifacts such as hooks, skills and rules for prompt injection or unauthorized behavior.

B007: Enforce user access privileges to AI systems

Control shoulds

- Implementing system-level access controls tailored to AI systems. For example, using role-based or attribute-based access to restrict access to model configuration, training datasets, tool-calling capabilities, or prompt logs, based on job function and system sensitivity.
- Restricting administrative and configuration privileges to authorized personnel. For example, limiting ability to alter system behavior, tools, or models.
- Conducting access reviews and updates at least quarterly. For example, validating access assignments, updating based on policy or role changes, documenting access changes with AI-specific context (e.g. model access justification, changes to agent capability boundaries, or access to sensitive prompt/response history).

B008: Protect AI system deployment environment

Control shoulds

- Enforcing caller authentication across API endpoints and agentic interfaces. For example, applying scoped API tokens or signed requests for model API access; enforcing OAuth 2.0 or OIDC token validation with appropriate scoping for MCP server connections; implementing mutual authentication for agent-to-agent interfaces.
- Securing data in transit across model API endpoints and agentic interfaces. For example, enforcing TLS for all model API endpoint traffic, MCP server connections, and agent-to-agent communication channels; implementing credential rotation policies for long-lived service connections.

Control mays

- Enforcing data integrity across agentic interfaces. For example, implementing cryptographic message signing for agent-to-agent communication; applying schema validation and input sanitization to MCP tool call inputs and outputs.
- Securing model hosting environments. For example, using up-to-date and minimal container images, scanning for known vulnerabilities in dependencies and base images, and applying infrastructure-level isolation techniques based on risk level (e.g. container namespaces, VM separation, or dedicated GPU access).
- Verifying model integrity before and during deployment. For example, using cryptographic checksums or signed artifacts to detect tampering, scanning model files for malicious payloads.

B009: Limit output over-exposure

Control shoulds

- Reducing or limiting the number of results shown in outputs to relevant only to balance security and utility. For example, character limits, limits on inference time.

Control mays

- Providing user-facing notices or documentation about output limitations.
- Limiting the fidelity of model outputs in certain use cases. For example, applying output rounding, threshold bands, or obfuscation techniques to reduce the risk of model inversion.

B010: Promote secure patterns in generated code

Control shoulds

- Implementing safeguards to promote secure patterns for common web vulnerability classes in generated code. For example, defaulting to parameterized queries or safe ORM patterns for database access, applying output encoding and contextual escaping to prevent cross-site scripting, or enforcing HTTPS/TLS in generated network and API code.
- Implementing safeguards to promote secure patterns for authentication and authorization in generated code. For example, defaulting to vetted authentication libraries or patterns on auth-related prompts, applying access control checks and least-privilege patterns by default, or avoiding generation of custom cryptographic or authentication primitives.
- Implementing safeguards to prevent dependency risks in generated code. For example, avoiding wildcard version specifications (e.g., * or latest) in generated manifests, verifying that suggested packages exist in their respective registries before including them in generated code, or applying guidance that avoids known-abandoned or typosquatted packages.

Control mays

- Implementing safeguards to promote secure defaults for session management in generated code. For example, applying HttpOnly, Secure, and SameSite flags to session cookies by default, or defaulting to CSRF protections on state-changing endpoints.
- Implementing safeguards to promote defensive programming patterns in generated code. For example, defaulting to input validation on user-supplied data, applying safe error handling that avoids exposing stack traces or internal details, or defaulting to rate limiting on generated public-facing endpoints.
- Implementing safeguards to prevent generated logging code from capturing secrets or personal information. For example, defaulting logging patterns to exclude sensitive fields, or applying guidance that avoids logging full request bodies or auth headers.

C001: Define AI risk taxonomy

Control shoulds

- Defining risk categories with severity levels and examples based on industry and deployment context. For example, classifying harmful outputs such as distressed outputs, angry responses, high-risk advice, offensive content, bias, and deception, identifying other high-risk use cases such as safety-critical instructions, legal recommendations, financial advice.
- Aligning risk taxonomy with external frameworks and standards.
- Establishing severity grading appropriate to organizational context and risk tolerance. For example, implementing consistent scoring methodology across risk categories, defining thresholds for flagging and human review.
- Maintaining taxonomy currency with documented change management. For example, updating based on emerging threats or incidents.

C002: Conduct pre-deployment testing

Control shoulds

- Conducting pre-deployment testing with documented results and identified issues. For example, structured hallucination testing, adversarial prompting, safety unit tests, and scenario-based walkthroughs.
- Completing risk assessments of identified issues before system deployment. For example, potential impact analysis, mitigation strategies, and residual risk evaluation.
- Obtaining approval sign-offs from designated accountable. For example, documented rationale for approval decisions and maintained records for review purposes.

Control mays

- Integrating AI system testing into established software development lifecycle (SDLC) gates. For example, including threat modelling and risk evaluation during design phases, requiring risk evaluation and sign-off at staging or pre-production milestones, aligning with CI/CD or MLOps pipelines, and documenting test artefacts in shared repositories."
- Implementing pre-deployment vulnerability scanning of AI artifacts and dependencies. For example, scanning AI models and ML libraries for security vulnerabilities, validating runtime behavior for unsafe operations, and analyzing outputs for harmful content before deployment.

C003: Prevent harmful outputs

Control shoulds

- Implementing content filtering for harmful output types. For example, detecting and blocking distressed responses, angry language, offensive content, biased statements, and deceptive information.
- Implementing guardrails for advice generation. For example, restricting high-risk recommendations in sensitive domains, requiring disclaimers for guidance.

Control mays

- Implementing bias detection and mitigation controls. For example, monitoring for discriminatory patterns, implementing fairness checks in outputs.
- Evaluating harm mitigation controls using performance metrics.

C004: Prevent out-of-scope outputs

Control shoulds

- Detecting and blocking out-of-scope requests. For example, detecting conversations outside intended use cases, blocking prohibited topics, providing redirection messages when users hit boundaries, and escalating or restricting access for repeated violations.
- Tracking out-of-scope violations and updating boundaries. For example, logging boundary violations, adjusting restrictions based on misuse patterns.

Control mays

- Providing user guidance on system capabilities and limitations. For example, communicating what the AI system can and cannot do, intended use cases, and topics or requests outside the system's scope.

C005: Prevent agent-specific high risk outputs

Control shoulds

- Implementing detection and blocking mechanisms aligned with organizational risk taxonomy. For example, deploying filtering based on defined risk categories and severity thresholds.
- Implementing response actions for detected risks. For example, blocking high-severity outputs, flagging medium-risk content for review, logging violations for monitoring and analysis.

Control mays

- Establishing escalation procedures for flagged high-risk content. For example, defining when human review is required and establishing approval workflows for edge cases.
- Implementing automated real-time interventions. For example, blocking or modifying outputs based on severity.

C006: Prevent output vulnerabilities

Control shoulds

- Establishing output sanitization and validation procedures before presenting content to users. For example, encoding or stripping potentially malicious content, validating structured outputs against safe schemas, blocking unsafe URLs, and enforcing secure rendering modes.
- Implementing content handling and security labelling based on trust level. For example, marking untrusted or third-party content, distinguishing external data from system-generated content, and applying differentiated security controls based on content source.

Control mays

- Detecting advanced output-based attack patterns. For example, identifying prompt injection attempts, model subversion techniques, payloads targeting downstream systems, or obfuscated exploits designed to bypass filters.

C007: Flag high risk outputs for human review

Control shoulds

- Defining high-risk output criteria drawing on risk taxonomy.
- Implementing automated detection mechanisms for high-risk outputs. For example, using content filtering, risk scoring, or classification models to identify outputs requiring review or flagging.

Control mays

- Establishing human review workflows for flagged high-risk outputs. For example, assigning reviewers, defining escalation procedures for complex cases, managing review queues with response time tracking, documenting review decisions, and reviewing workflow effectiveness regularly.

C008: Monitor AI risk categories

Control shoulds

- Establishing ongoing monitoring of AI outputs across risk categories. For example, conducting regular evaluations prioritized by risk severity, sampling outputs for review, and tracking system behavior patterns.

Control mays

- Maintaining documentation. For example, recording identified scenarios with clear examples, updating risk taxonomy based on monitoring findings and incidents.
- Integrating AI output monitoring with existing security tools. For example, forwarding alerts and flagged outputs to SIEM platforms, applying standard logging formats (e.g. JSON, syslog) to support automated threat detection workflows.

C009: Enable real-time feedback and intervention

Control shoulds

- Enabling user intervention capabilities. For example, providing mechanisms for users to pause, stop, or redirect system behavior, implementing feedback collection tools for users to report issues or concerns, ensuring technical controls persist across devices and interaction contexts.
- Ensuring accessibility of feedback and intervention mechanisms. For example, adhering to WCAG 2.1 standards for color contrast, screen reader compatibility, keyboard navigation, and clear messaging for users with disabilities.

Control mays

- Reviewing user feedback and intervention logs at regular intervals, analyzing findings using structured methodologies (e.g., categorizing by risk domain, frequency, and severity), and integrating corrective actions into product backlogs or compliance workflows, with records maintained for traceability.

C010: Third-party testing for harmful outputs

Control shoulds

- Appointing qualified third-party assessors. Including selecting assessors with relevant technical capabilities for identified risk areas, maintaining records of assessor qualifications and independence.
- Conducting regular testing. Including performing assessments of harmful outputs at least every quarter, defining testing scope and methodologies based on risk classifications and industry benchmarks like ToxiGen, coordinating with internal security and testing teams.
- Maintaining documentation. Including testing scope, results, and remediation actions taken, tracking follow-up activities and resolution timelines.

C011: Third-party testing for out-of-scope outputs

Control shoulds

- Appointing qualified third-party assessors. Including selecting assessors with relevant technical capabilities for identified risk areas, maintaining records of assessor qualifications and independence.
- Conducting regular testing. Including defining testing scope and methodologies based on risk taxonomy and performing assessments of out-of-scope outputs at least every quarter.
- Maintaining documentation. Including testing scope, results, and remediation actions taken, tracking follow-up activities and resolution timelines.

C012: Third-party testing for customer-defined risk

Control shoulds

- Appointing qualified third-party assessors. Including selecting assessors with relevant technical capabilities for identified risk areas, maintaining records of assessor qualifications and independence.
- Conducting regular testing. Including defining testing scope and methodologies based on risk taxonomy and performing assessments of high-risk areas at least every quarter.
- Maintaining documentation. Including testing scope, results, and remediation actions taken, tracking follow-up activities and resolution timelines.

D001: Prevent hallucinated outputs

Control shoulds

- Implementing factual accuracy controls. For example, deploying available fact-checking mechanisms, flagging uncertain or low-confidence responses.
- Establishing information source validation. For example, requiring citations for factual claims, implementing source reliability checks.

Control mays

- Maintaining uncertainty communication. For example, displaying confidence levels, providing appropriate disclaimers for generated information.

D002: Third-party testing for hallucinations

Control shoulds

- Appointing qualified third-party assessors. Including selecting assessors with relevant technical capabilities for identified risk areas, maintaining records of assessor qualifications and independence.
- Conducting regular testing. Including defining testing scope and methodologies based on risk taxonomy and performing assessments at least every quarter.
- Maintaining documentation. Including testing scope, results, and remediation actions taken, tracking follow-up activities and resolution timelines.

D003: Restrict unsafe tool calls

Control shoulds

- Implementing tool call validation and authorization. For example, restricting tool calls to approved functions and MCP servers, validating parameters before execution.
- Enforcing rate limits and transaction caps for autonomous tool use.
- Establishing execution monitoring and logging. For example, tracking all tool calls, monitoring for unauthorized access attempts or scope violations.

Control mays

- Requiring human approval for sensitive tool operations. For example, requiring human confirmation before executing high-risk actions, multi-step tool calls, implementing approval workflows for operations beyond autonomous boundaries.
- Reviewing patterns of AI tool usage. For example, identifying anomalies, updating tool permissions, and retiring unused or high-risk functions during scheduled evaluations.

D004: Third-party testing of tool calls

Control shoulds

- Appointing qualified third-party assessors. Including selecting assessors with relevant technical capabilities for identified risk areas, maintaining records of assessor qualifications and independence.
- Conducting regular testing. Including defining testing scope and methodologies based on risk taxonomy and performing assessments of tool calls at least every quarter.
- Maintaining documentation. Including testing scope, results, and remediation actions taken, tracking follow-up activities and resolution timelines.

E001: AI failure plan for security breaches

Control shoulds

- Assigning a breach response lead from existing staff. For example, IT manager, security officer, or designated executive with authority to engage external counsel and specialists as needed.
- Defining breach notification procedures. For example, customer communications, regulatory reporting requirements, and vendor notifications based on applicable privacy laws.
- Implementing security remediation measures. For example, system freeze capabilities, vulnerability fixes, access control updates, and coordination with external security consultants when internal expertise is insufficient.
- Establishing evidence collection requirements with guidance on preserving evidence for potential legal review. For example, system logs, user activity records, and basic documentation.

E002: AI failure plan for harmful outputs

Control shoulds

- Implementing customer communication protocols. For example, disclosure procedures, explanation of corrective actions, and follow-up commitments with executive approval for significant incidents.
- Establishing immediate mitigation steps with designated staff responsibilities. For example, system freeze capabilities, output suppression, customer notification, and system adjustments.

Control mays

- Defining harmful output categories with reference to risk taxonomy. For example, discriminatory content, offensive material, inappropriate recommendations, ideally with concrete examples.
- Coordinating external support engagement. For example, legal counsel consultation, PR support, and insurance claim procedures.

E003: AI failure plan for hallucinations

Control shoulds

- Implementing customer communication protocols. For example, disclosure procedures, explanation of corrective actions, and follow-up commitments with executive approval for significant incidents.
- Establishing immediate mitigation steps with designated staff responsibilities. For example, system freeze capabilities, model adjustments, output validation improvements, customer notification, and enhanced monitoring.

Control mays

- Defining hallucination incident types.
- Coordinating potential external support. For example, legal consultation for significant claims, financial review when needed, and insurance coverage activation.

E004: Assign accountability

Control shoulds

- Defining AI system changes requiring approval including model selection, material changes to the meta prompt, adding / removing guardrails, changes to end-user workflow, other changes that drive material. For example, +/-10% performance on evals.
- Assigning an accountable lead as approver for each of these changes. Can follow a RACI structure to formalize roles of those consulted and informed.

Control mays

- Implementing code signing and verification processes for AI models, libraries, and deployment artefacts to ensure only digitally signed components are approved for production use.

E005: Document data storage security

Control shoulds

- Documenting data storage security. For example, assessments around cloud vs. on-premises processing.

E006: Conduct vendor due diligence

Control shoulds

- Defining assessment criteria for foundational or upstream AI models. For example, data handling and ownership practices, PII controls, security measures, compliance status, open-source.
- Conducting documented assessments. For example, scoring results, verification activities such as certifications reviewed and references contacted, and approval decisions.
- Maintaining assessment records with sufficient detail for audit purposes and retaining due diligence evidence before vendor approval.

E007: [Retired] Document system change approvals

Control shoulds

- Documenting formal review and approval decisions for changes defined in E004: Assign accountability.

E008: Review internal processes

Control shoulds

- Reviewing decision processes every quarter including AI system changes, foundational model selection, security assessment.
- Maintaining a centralized repository of decision records and internal review of these record. For example, supporting evidence reviewed, remediation plans.
- Documenting and tracking remediation of any risks identified.

Control mays

- Collecting and implementing external feedback on AI systems. For example, system risks, new threat patterns, new mitigation strategies.

E009: Monitor third-party access

Control shoulds

- Configuring logging for third-party interactions. For example, capturing API connections, user access sessions, data exchanges, and service integrations.
- Capturing access metadata. For example, user identification, authentication timestamps, accessed resources, session duration, origin IP addresses, and resource usage patterns.

Control mays

- Generating alerts on anomalous third-party access patterns against defined detection rules. For example, alerting on unexpected call volume or out-of-scope endpoints for service connections, repeated failed logins for human access, or credential use outside approved scope - with each alert routed to a responsible owner for disposition.

E010: Establish AI acceptable use policy

Control shoulds

- Defining prohibited AI usage for end-users. For example, jailbreak attempts, malicious prompt injection, unauthorized data extraction, generation of harmful content, and misuse of customer data.
- Implementing detection and monitoring tools. For example, prompt analysis, output filtering, usage pattern anomalies, and suspicious access attempts.
- Implementing user feedback when policy is breached. For example, showing alerts or error messages when inputs violate acceptable use.

Control mays

- Real-time monitoring, blocking, or alerting capabilities.
- Maintaining logging and tracking systems. For example, incident creation, violation tracking with case assignment and resolution documentation.
- Conducting regular effectiveness reviews. For example, quarterly analysis of violation trends, tool performance assessment, policy updates based on emerging threats, and user training adjustments.

E011: Record processing locations

Control shoulds

- Maintaining AI infrastructure location documentation. For example, geographic locations of foundation model processing locations and inference endpoint regions, documenting third-party AI service provider data handling locations.
- Reviewing and updating documentation regularly.

Control mays

- Implementing transfer compliance procedures. For example, assessing data transfer requirements for AI training data and inference processing, maintaining approved transfer mechanisms for foundation model providers and AI infrastructure, mitigating transfer risk for cross-border AI model training.

E012: Document regulatory compliance

Control shoulds

- Identifying relevant regulations. For example, data protection laws. For example, GDPR, CCPA, sector-specific requirements, emerging AI standards. For example, EU AI Act.
- Documenting compliance procedures and strategies appropriate for company size and operations.
- Reviewing the repository every 6 months and when additional requirements may be triggered. For example, regulations change or business operations expand into new jurisdictions.

E013: Implement quality management system

Control shoulds

- Defining quality objectives, metrics, and risk management approach for AI systems. For example, establishing performance targets, safety thresholds, risk assessment methodologies, and measurement processes appropriate to system risk level.
- Establishing change management, approval processes, and documentation standards. For example, defining review and approval requirements for AI system changes, assigning accountability for quality decisions, documenting design and development procedures.
- Implementing defect tracking, continuous improvement, and post-market monitoring. For example, maintaining issue tracking systems, conducting root cause analysis, documenting corrective actions, establishing post-market monitoring processes.

Control mays

- Establishing data management and record-keeping systems. For example, documenting data governance procedures, maintaining technical documentation, implementing record retention policies for model training data and system outputs.
- Documenting communication procedures with regulatory authorities and stakeholders. For example, establishing protocols for regulatory reporting, stakeholder notifications for incidents, and procedures for authority interactions.

E014: [Retired] Share transparency reports

Control mays

- This requirement was merged into E017 at the Q1, 2026 standard update. See aiuc-1.com/changelog for more information.

E015: Log AI system activity

Control shoulds

- Capturing system activity details to support incident investigation and behavior explanation. For example, logging inputs, processing steps, outputs, and metadata for AI systems.
- Implementing log storage with appropriate retention periods, access controls, and data sanitation to support auditing and incident response.

Control mays

- Capturing full execution chains of agentic workflows to support investigation of agent-specific incidents. For example, logging agent provenance metadata, tool call parameters and results, sub-agent delegations and their outcomes, approval/authorization events (e.g., human-in-the-loop approvals), and reasoning traces where available.
- Implementing technical controls to ensure logs are tamper-evident and independently verifiable. For example, ensuring that captured records cannot be modified or deleted after creation, ensuring sequence integrity so that gaps, omissions, and reordering are detectable during incident investigation or audit.

E016: Implement AI disclosure mechanisms

Control shoulds

- Implementing AI disclosure for text-based interactions. For example, displaying clear notices when users interact with AI chatbots, virtual assistants, or automated messaging systems.
- Implementing AI disclosure for voice-based interactions. For example, providing audio notifications at the beginning of voice calls or interactions.
- Labelling AI-generated media and documents in a machine-readable and detectable format. For example, marking AI-generated images, videos, audio, or documents with metadata, watermarks, or labels indicating artificial generation.
- Disclosing when autonomous AI agents or systems are performing actions. For example, notifying users when AI systems are making decisions, processing requests, or executing tasks without human oversight.
- Establishing reactive disclosure capabilities when users ask if they are interacting with AI.

E017: Document system transparency policy

Control shoulds

- Creating transparency documentation for major AI systems. For example, documenting system characteristics, data provenance, and model behavior for systems meeting documentation criteria.

Control mays

- Defining policies for sharing transparency documentation with external stakeholders. For example, establishing when reports are shared, specifying recipient categories, determining what information is disclosed to each stakeholder type.
- Documenting sharing procedures including approval workflows, version control, and distribution tracking. For example, establishing approval requirements before external sharing, maintaining version control of shared documents, tracking which stakeholders received which versions.
- Documenting platform-level and deployer-level security responsibilities for AI systems. For example, delineating which security obligations are managed by the platform versus the deploying organization.

F001: Prevent AI cyber misuse

Control shoulds

- Results of testing from foundation model developer on offensive cyber capabilities and mitigations.

Control mays

- Implementing malicious use detection and blocking. For example, deploying available content filtering to detect requests for malicious code generation, attack planning, and vulnerability exploitation guidance, configuring automated blocking of cyber attack assistance requests, maintaining databases of prohibited use patterns.

F002: Prevent catastrophic misuse

Control shoulds

- Results of testing from foundation model developer on CBRN capabilities and mitigations.

Control mays

- Establishing catastrophic misuse monitoring. For example, monitoring AI system interactions for patterns indicating weapons development or mass harm intent, implementing real-time alerting for detected catastrophic misuse attempts, documenting suspicious queries and system responses.