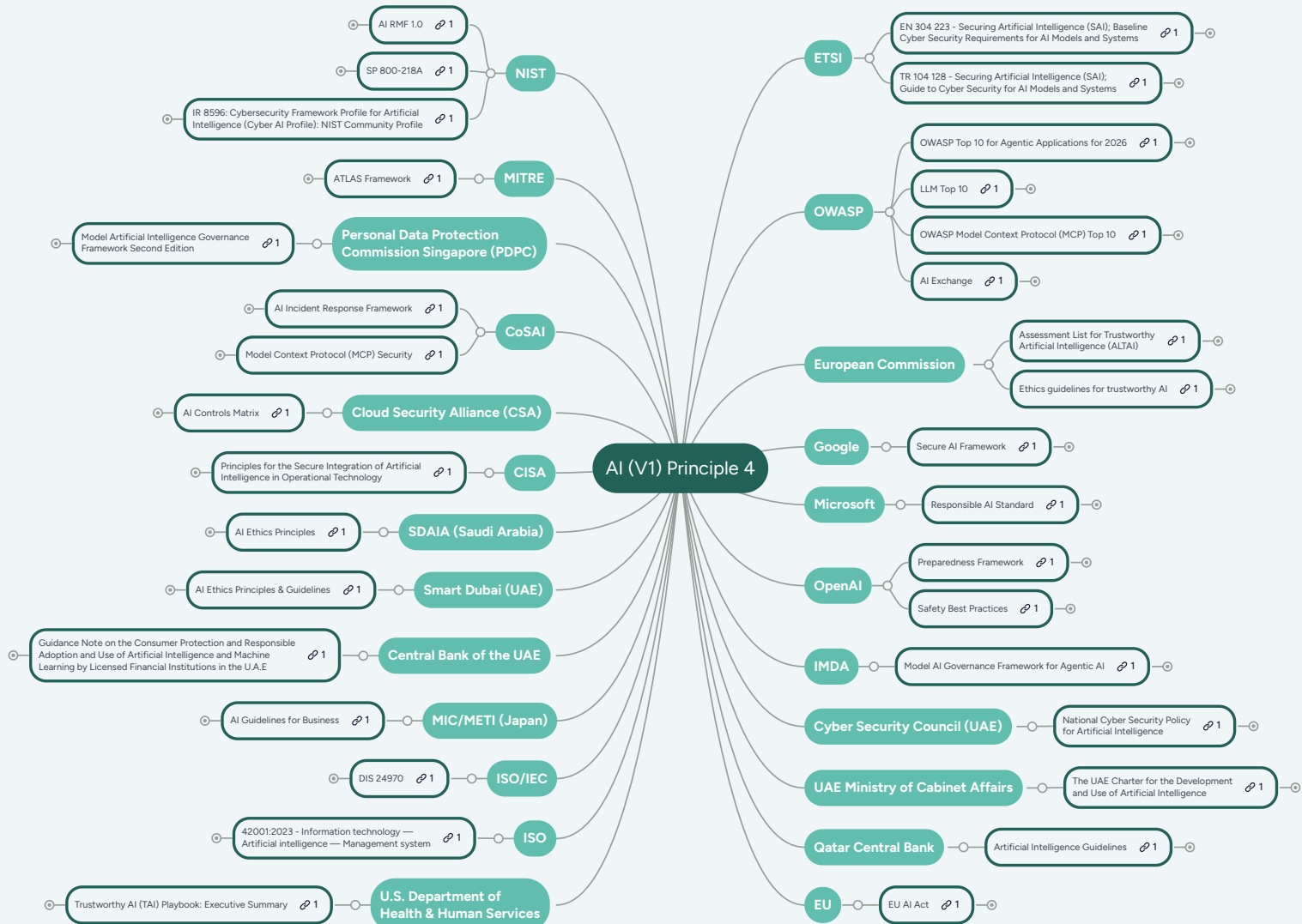




AI (V1) Principle 4

1. ETSI

1.1. EN 304 223 - Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems

Link: https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

Link: https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

1.1.1. Provision 5.1.4-1

When designing an AI system, Developers and/or System Operators should incorporate and maintain capabilities to enable human oversight.

1.1.2. Provision 5.1.4-2

Developers should design systems to make it easy for humans to assess outputs that they are responsible for in said system (such as by ensuring that models outputs are explainable or interpretable).

1.1.3. Provision 5.1.4-3

Where human oversight is a risk control, Developers and/or System Operators shall design, develop, verify and maintain technical measures to reduce the risk through such oversight.

1.1.4. Provision 5.1.4-4

Developers should verify that the security controls specified by the Data Custodian have been built into the system.

1.1.5. Provision 5.1.4-5

Developers and System Operators should make End-users aware of prohibited use cases of the AI system.

1.2. TR 104 128 - Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems

Link: https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

1.2.1. Provision 5.1.4-1

"When designing an AI system, Developers and/or System Operators should incorporate and maintain capabilities to enable human oversight." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without built-in human oversight, AI systems will generate incorrect outputs or decisions that are difficult to interpret, verify, or override, increasing risks of data protection compliance, unintended consequences, misuse, or harmful impacts.

Example Measures/Controls 1:

Implement Mechanisms for Human Oversight: Control: Implement features that allow human operators to easily interpret, verify, and act on AI outputs, including manual release and overrides. Ensure that the design meet obligations around automated decisions and encourages meaningful human decision-making rather than passive acceptance of AI recommendations.

Example Measures/Controls 2:

Measure and Validate Accuracy of Human Oversight Decisions: Regularly test and measure the accuracy of human oversight decisions, validating that operators can correctly interpret and act on AI outputs and identifying areas for improvement. Assess not just individual performance but how the system supports human understanding and engagement to foster effective sociotechnical communication between the operator and AI.

1.2.2. Provision 5.1.4-2

"Developers should design systems to make it easy for humans to assess outputs that they are responsible for in said system (such as by ensuring that models outputs are explainable or interpretable)." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without clarity and ease of use, users can not perform oversight effectively leading to failures and harm.

Example Measures/Controls:

Develop User-Friendly Human Responsibility UI: Implement UIs that display outputs, decision-making rationales, and logs clearly to make it easy for human operators to assess outputs and understand their accountability. Ensure systems are designed to encourage rigorous assessment by humans and not condition them to simply click an approve button.

1.2.3. Provision 5.1.4-3

"Where human oversight is a risk control, Developers and/or System Operators shall design, develop, verify, and maintain technical measures to reduce the risk through such oversight." (ETSI TS 104 223 [i.1])

Related threats/risks:

Ineffective oversight technical measures can compromise the risk reduction effort by overburdening or failing to adequately support human reviewers.

Example Measures/Controls:

Implement Validation and Enforcement of Oversight controls: Design and implement technical measures that provide guardrails to assist human reviewers in understanding, interpreting, and acting on AI outputs.

1.2.4. Provision 5.1.4-4

"Developers should verify that the security controls specified by the Data Custodian have been built into the system." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without validation of Data Custodian controls, the system can lack necessary data protection and governance measures, potentially leading to security vulnerabilities or regulatory non-compliance.

Example Measures/Controls:

Conduct Validation of Custodian: Verify that all controls specified by the Data Custodian have been implemented correctly, with testing to validate effectiveness and alignment with data protection requirements and guidance.

1.2.5. Provision 5.1.4-5

"Developers and System Operators should make End-users aware of prohibited use cases of the AI system." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without clear communication on prohibited uses, end-users can unintentionally misuse the AI system, leading to legal, ethical, or operational risks.

Example Measures/Controls 1:

Document and Train Users on Prohibited Use Cases: Clearly define and document prohibited use cases for the AI system, ensuring end-users understand limitations and restrictions. Use threat modelling to identify and inform users of all known harmful states and unmitigated risks.

Example Measures/Controls 2:

Monitor for Prohibited Use Cases: Implement controls to actively monitor, detect, and prevent prohibited use cases.

2. OWASP

2.1. OWASP Top 10 for Agentic Applications for 2026

Link: <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>

Link: <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>

2.1.1. ASI02: Tool Misuse and Exploitation - 2

Action-Level Authentication and Approval. Require explicit authentication for each tool invocation and human confirmation for high-impact or destructive actions (delete, transfer, publish). Display a pre-execution plan or dry-run diff before final approval; where possible, present a dry-run or diff preview to the user before high-impact actions are approved.

2.1.2. ASI03: Identity and Privilege Abuse - 4

Apply Human-in-the-Loop for Privilege Escalation: Require human approval for high-privilege or irreversible actions to provide a safety net that would stop Memory-Based Escalation, Cross-Agent Trust Exploitation, and Maintenance Window attacks.

2.1.3. ASI04: Agentic Supply Chain Vulnerabilities - 8

Supply chain kill switch: Implement emergency revocation mechanisms that can instantly disable specific tools, prompts, or agent connections across all deployments when a compromise is detected, preventing further cascading damage.

2.1.4. ASI05: Unexpected Code Execution (RCE) - 6

Access control and approvals: Require human approval for elevated runs; keep an allowlist for auto-execution under version control; enforce role and action-based controls.

2.1.5. ASI06: Memory & Context Poisoning - 7

Resilience and verification: Perform adversarial test, use snapshots/rollback and version control, and require human review for high-risk actions. Where you operate shared vector or memory stores, use per-tenant namespaces and trust scores for entries, decaying or expiring unverified memory over time and supporting rollback/quarantine for suspected poisoning.

2.1.6. ASI08: Cascading Failures - 5

Output validation and human gates: Checkpoints, governance agents, or human review for high risk before agent outputs are propagated downstream.

2.1.7. ASI09: Human-Agent Trust Exploitation - 1

Explicit confirmations: Require multi-step approval or "human in the loop" before accessing extra sensitive data or performing risky actions.

2.1.8. ASI09: Human-Agent Trust Exploitation - 5

Adaptive Trust Calibration: Continuously adjust the level of agent autonomy and required human oversight based on contextual risk scoring. Implement confidence weighted cues (e.g., "low-certainty" or "unverified source") that visually prompt users to question high-impact actions, reducing automation bias and blind approval. Develop and continuously maintain appropriate training of human personnel involved in the evolving human oversight of autonomous agentic systems.

2.1.9. ASI10: Rogue Agents - 4

Containment & Response: Implement rapid mechanisms like kill-switches and credential revocation to instantly disable rogue agents. Quarantine suspicious agents in sandboxed environments for forensic review.

2.2. LLM Top 10

Link: <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>

Link: <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>

2.2.1. LLM01: Prompt Injection - 5

Implement human-in-the-loop controls for privileged operations to prevent unauthorized actions.

2.2.2. LLM06: Excessive Agency - 6

Utilise human-in-the-loop control to require a human to approve high-impact actions before they are taken. This may be implemented in a downstream system (outside the scope of the LLM application) or within the LLM extension itself. For example, an LLM-based app that creates and posts social media content on behalf of a user should include a user approval routine within the extension that implements the 'post' operation.

2.3. OWASP Model Context Protocol (MCP) Top 10

Link: <https://owasp.org/www-project-mcp-top-10/>

Link: <https://owasp.org/www-project-mcp-top-10/>

2.3.1. MCP02:2025 - Privilege Escalation via Scope Creep - 6

Runtime Controls & Guardrails Implement runtime policy enforcement (PDP/PIP) to block disallowed commands or tool calls. Apply action whitelists, safe execution sandboxes, and require multi-step confirmation for high-impact operations.

2.3.2. MCP02:2025 - Privilege Escalation via Scope Creep - 8

Separation of Duties & Approval Flows Separate the authority to grant permissions from the authority to deploy code or change production settings. Require human-in-the-loop approvals for non-routine privilege grants.

2.3.3. MCP03:2025 - Tool Poisoning - 6

Runtime Enforcement & Guardrails

- Don't allow agents to interpret schema changes as immediate action drivers without revalidation.
- Require a "schema attestation" that binds the schema hash to a specific agent identity and session.
- Implement runtime sanity checks: if an operation's semantic impact exceeds a threshold (e.g., destructive verbs, data volume), pause execution and require human approval.

2.3.4. MCP03:2025 - Tool Poisoning - 11

Patch CI/CD and registry processes to require signed commits and multi-party approvals where missing.

2.3.5. MCP05:2025 – Command Injection & Execution - 6

Add Human-in-the-Loop for Sensitive Actions Require approval for destructive, privileged, or system-modifying operations. Log all tool calls with full parameters and maintain immutable audit trails.

2.3.6. MCP06:2025 – Intent Flow Subversion - 5

Active Drift Detection & Human-in-the-Loop

- Monitor for “Intent Drift”—where the semantic alignment between the user’s request and the agent’s actions degrades over time.
- Automatically pause the session and require human re-authentication of the intent flow if the agent’s plan deviates from the original goal.

2.3.7. MCP10:2025 – Context Injection & Over-Sharing - 6

Human-in-the-Loop for Sensitive Context Require approval before sensitive context is: Exported Summarized Shared across agents Show a preview of context that will be reused.

2.4. AI Exchange

Link: https://owaspai.org/docs/ai_security_overview/

Link: https://owaspai.org/docs/ai_security_overview/

2.4.1. 1.3. Controls to limit the effects of unwanted behaviour - OVERSIGHT

Oversight of model behaviour by humans or automated mechanisms (e.g., using rules), where human oversight provides not only more intelligent validation through common sense and domain knowledge, but also clear accountability for deviations and outcomes.

3. European Commission

3.1. Assessment List for Trustworthy Artificial Intelligence (ALTAI)

Link: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

Link: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

3.1.1. REQUIREMENT #1 Human Agency and Oversight

AI systems should support human agency and human decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both: act as enablers for a democratic, flourishing and equitable society by supporting the user’s agency; and uphold fundamental rights, which should be underpinned by human oversight. In this section AI systems are assessed in terms of their respect for human agency and autonomy as well as human oversight.

3.1.2. REQUIREMENT #7 Accountability

The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems. This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties. When unjust or adverse impacts occur, accessible mechanisms for accountability should be in place that ensure an adequate possibility of redress.

3.2. Ethics guidelines for trustworthy AI

Link: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Link: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

3.2.1. 1.1.3 Human Oversight

Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.

3.2.2. 1.4.2 Explainability

Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

4. Google

4.1. Secure AI Framework

Link: <https://www.saif.google/secure-ai-framework>

Link: <https://www.saif.google/secure-ai-framework>

4.1.1. Agent User Control

Ensure user approval for any actions performed by agents/plugins that alter user data or act on the user's behalf.

5. Microsoft

5.1. Responsible AI Standard

Link: <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us>

Link: <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us>

5.1.1. A5.1

Identify the stakeholders who are responsible for troubleshooting, managing, operating, overseeing, and controlling the system during and after deployment. Document these stakeholders and their oversight and control responsibilities using the Impact Assessment template.

5.1.2. A5.2

Identify the system elements (including system UX, features, alerting and reporting functions, and educational materials) necessary for stakeholders identified in requirement A5.1 to effectively understand their oversight responsibilities and carry them out. Stakeholders must be able to understand:

- 1) the system's intended uses,
 - 2) how to effectively execute interactions with the system,
 - 3) how to interpret system behavior,
 - 4) when and how to override, intervene, or interrupt the system, and
 - 5) how to remain aware of the possible tendency of over-relying on outputs produced by the system ("automation bias").
- Document the system design elements that will support relevant stakeholders for each oversight and control function.

5.1.3. A5.3

When possible, design the system elements identified in A5.2. When this is not possible (for example, when Microsoft is not responsible for the system UX), provide guidance on human oversight considerations to the third party responsible for implementing the system elements identified in A5.2.

5.1.4. A5.4

Define and document the method to be used to evaluate whether each oversight or control function can be accomplished by stakeholders in realistic conditions of system use. Include the metrics or rubrics that will be used in the evaluations. When this is not possible (for example, when Microsoft is not responsible for oversight and control functions), provide guidance on evaluating oversight and control functions to the third party responsible for evaluating oversight or control functions.

5.1.5. T1.3

Define and document the method to be used to evaluate whether each stakeholder who will make decisions or be subject to decisions based on the behavior of the system can interpret the relevant system responses reasonably well. Include the metrics or rubrics that will be used in the evaluations.

6. OpenAI

6.1. Preparedness Framework

Link: <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>

Link: <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>

6.1.1. Safeguards Against a Misaligned Model - Reliable and Robust System Oversight

Effective AI and human oversight of model actions detects and prevents execution of harm or subversion of safeguards.

6.2. Safety Best Practices

Link: <https://platform.openai.com/docs/guides/safety-best-practices>

Link: <https://platform.openai.com/docs/guides/safety-best-practices>

6.2.1. Human in the loop (HITL)

Wherever possible, we recommend having a human review outputs before they are used in practice. This is especially critical in high-stakes domains, and for code generation. Humans should be aware of the limitations of the system, and have access to any information needed to verify the outputs (for example, if the application summarizes notes, a human should have easy access to the original notes to refer back).

7. IMDA

7.1. Model AI Governance Framework for Agentic AI

Link: <https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf>

Link: <https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf>

7.1.1. 2.2 Make humans meaningfully accountable 2.2.1 Clear allocation of responsibilities within and outside the organisation

By establishing chains of accountability across the agent value chain and lifecycle, while emphasising adaptive governance, so that the organisation is set up to quickly understand new developments and update their approach as the technology evolves. As deployers, organisations and humans remain accountable for the decisions and actions of agents. However, as with AI, the value chain for agentic AI involves multiple actors. Organisations should consider the allocation of responsibility both within their organisation, and vis-à-vis other organisations along the value chain. Within the organisation, organisations should allocate responsibilities for different teams across the agent lifecycle. Outside the organisation, organisations may also need to work with external parties when deploying agents e.g. model developers, agentic AI providers, or hosts of external MCP servers or tools.

7.1.2. 2.2.2 Design for meaningful human oversight - 1

Organisations should define significant checkpoints or action boundaries that require human approval, especially before sensitive actions are executed. This can include:²³

- High-stakes actions and decisions e.g. editing of sensitive data, final decisions in high-risk domains (such as healthcare or legal), actions that may trigger liability
- Irreversible actions e.g. permanently deleting data, sending communications, making payments
- Outlier or atypical behaviour e.g. when agent accesses a system or database outside of its work scope, when agent selects a delivery route that is twice as long as the median distance
- User-defined. Agents may act on behalf of users who have different risk appetites. Beyond organisation-defined boundaries, users may be given the option to define their own boundaries e.g. requiring approval for purchases above a certain amount

7.1.3. 2.2.2 Design for meaningful human oversight - 2

Apart from considering when approvals are required, organisations should also consider what form approvals should take. These considerations include:

- Keep approval requests contextual and digestible. When asking humans for approval, keep the request short and clear, instead of providing long logs or raw data that may be challenging to decipher and understand.
- Consider the form of human input required. For straightforward actions such as accessing a database, the human user can simply approve or reject. For more complex cases, such as reviewing an agent's plan before execution, it may be more productive for the human to edit the plan before giving the agent the go-ahead.

Organisations should implement measures to ensure continued effectiveness of human oversight, particularly as humans remain susceptible to alert fatigue and automation bias. These measures can include:

- Training humans to identify common failure modes e.g. inconsistent agent reasoning, agents referring to outdated policies
- Regularly auditing the effectiveness of human oversight

8. Cyber Security Council (UAE)

8.1. National Cyber Security Policy for Artificial Intelligence

Link: https://csc.gov.ae/documents/38662/0/National+Cyber+Security+Policy+for+Artificial+Intelligence_v1.1.pdf/4e02b32e-9f62-948d-4bc8-b580d596451b?t=1766994254544

Link: https://csc.gov.ae/documents/38662/0/National+Cyber+Security+Policy+for+Artificial+Intelligence_v1.1.pdf/4e02b32e-9f62-948d-4bc8-b580d596451b?t=1766994254544

8.1.1. 3.1.1 Cyber Security Policies & Procedures - 2

The AI/ML cyber security policy should define roles and responsibilities for security within the AI/ML lifecycle, providing clarity on who is responsible for implementing, monitoring, and enforcing these policies.

8.1.2. 3.2.1 Asset Management for AI/ML Systems - 4

The entity should establish clear ownership and accountability for each AI/ML asset, including responsibility for its security and compliance.

8.1.3. 3.6.2 Incident Reporting and Management for AI/ML - 5

In cases where automated response mechanisms may not be suitable, or fully effective, the entity should implement a fail-safe mechanism that allows for timely and informed human intervention.

9. UAE Ministry of Cabinet Affairs

9.1. The UAE Charter for the Development and Use of Artificial Intelligence

Link: <https://uaelegislation.gov.ae/en/policy/details/the-uae-charter-for-the-development-and-use-of-artificial-intelligence#:~:text=The%20charter%20covers%20the%20following%20priorities%20and,and%20use%20of%20AI%20in%20the%20country.>

Link: <https://uaelegislation.gov.ae/en/policy/details/the-uae-charter-for-the-development-and-use-of-artificial-intelligence#:~:text=The%20charter%20covers%20the%20following%20priorities%20and,and%20use%20of%20AI%20in%20the%20country.>

9.1.1. 6. Human Oversight

The Charter emphasizes the irreplaceable value of human judgment and human oversight over AI, aligning with ethical values and social standards to correct any errors or biases that may arise.

9.1.2. 7. Governance and Accountability

The UAE adopts a responsible and proactive stance, emphasizing the importance of governance and accountability in AI to ensure the technology is used ethically and transparently.

10. Qatar Central Bank

10.1. Artificial Intelligence Guidelines

Link: https://www.qcb.gov.qa/Services/Financial%20Technology/QCB_Artificial_Intelligence_Guideline.pdf

Link: https://www.qcb.gov.qa/Services/Financial%20Technology/QCB_Artificial_Intelligence_Guideline.pdf

10.1.1. 7.1

The Board of Directors (BOD) and senior management of an Entity remain accountable for the outcomes and decisions of the Entity's AI Systems including those systems that make decisions on behalf of the Entity.

10.1.2. 7.3

The key responsibilities of the senior management include but are not limited to:

- Must have one or more members with the knowledge to understand and manage technology risks, ideally including AI.
- Must be responsible for the assessment, understanding and monitoring of the Entity's reliance on AI.
- Must ensure responsibility for, and oversight of the various stages and activities involved in AI portfolio deployment is allocated to the appropriate personnel and/ or departments.
- Must provide information to the BOD that is clear, consistent, robust, timely, well-targeted and contain an appropriate level of technical detail to allow the BOD to provide effective oversight and challenge management.

10.1.3. 7.4

An Entity should establish either a function overseeing AI or delegating its responsibility to an existing function within an Entity.

10.1.4. 7.5

The function responsible for overseeing AI must utilize or create appropriate committees to assess AI use cases prior to implementation.

10.1.5. 8.2

An Entity must allocate key roles and responsibilities associated with managing the Entity's AI portfolio.

10.1.6. 8.2.3

Clearly allocate roles & responsibilities between model owners, developers and approvers.

10.1.7. 13.1

Any AI Systems must have a Human Oversight protocol.

10.1.8. 13.2

High risk AI Systems must be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI System is in use.

10.1.9. 13.3

The User must assign Human Oversight to a Supervisor who has the necessary competence, training, and authority to operate or oversee the relevant AI System.

10.1.10. 13.4

An Entity must ensure that the Supervisor is given tools and authority as appropriate and proportionate to the circumstances to:

- Understand the capacities and limitations of the High-Risk AI System and are able to duly monitor its operation.
- Correctly interpret the High-Risk AI System's output, considering for example the interpretation tools and methods available.
- To decide, in any situation, not to use the High-Risk AI System or otherwise disregard, override or reverse the output of the High-Risk AI System.
- Have the authority and means to intervene in the operation of the High-Risk AI System or interrupt the system through a "stop" button or a similar procedure.

10.1.11. 13.6.5

The Supervisor(s) must have the capacity to close the system down in the event outputs from or data around the system seems aberrant.

10.1.12. 13.7.1

An entity that plans to provide or use an AI System that requires Human Oversight in a monitoring role, with the ability to take over control must ensure there will be appropriate human-machine interface tools the ability to allow a Supervisor to take over control.

10.1.13. 13.7.3

An Entity may ensure the design and development of AI Systems in such a way that it allows a Supervisor to oversee the AI Algorithm's functioning and allow decision making in a timely manner.

10.1.14. 13.7.4

An Entity must ensure that appropriate Human Oversight measures should be in place before use.

10.1.15. 13.7.6

An Entity must ensure the AI System is responsive to the Supervisor.

10.1.16. 21.1

An Entity must put in place a mechanism for Customers to raise inquiries about AI decisions and request reviews of decisions made by AI Systems with no human intervention.

10.1.17. 23.4

An Entity must ensure that the appropriate individuals or levels of authority within the Entity consistently approve the exemption.

10.1.18. 23.6

An Entity must remain accountable for any risks stemming from approved exemptions.

11. EU

11.1. EU AI Act

Link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

Link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

11.1.1. 13.1 Transparency and Provision of Information to Deployers

High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3.

11.1.2. 14.1 Human Oversight

High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.

11.1.3. 14.2 Human Oversight

Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular where such risks persist despite the application of other requirements set out in this Section.

11.1.4. 14.4 Human Oversight

For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate:

- (a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;
- (b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
- (c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;
- (d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;
- (e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.

11.1.5. 26.2 Obligations of deployers of high-risk AI systems

Deployers shall assign human oversight to natural persons who have the necessary competence, training and authority, as well as the necessary support.

11.1.6. 26.3 Obligations of deployers of high-risk AI systems

The obligations set out in paragraphs 1 and 2, are without prejudice to other deployer obligations under Union or national law and to the deployer's freedom to organise its own resources and activities for the purpose of implementing the human oversight measures indicated by the provider.

12. U.S. Department of Health & Human Services

12.1. Trustworthy AI (TAI) Playbook: Executive Summary

Link: <https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook-executive-summary.pdf>

Link: <https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook-executive-summary.pdf>

12.1.1. Responsible / Accountable

Policies should outline governance and who is held responsible for all aspects of the AI solution (e.g., initiation, development, outputs, decommissioning)

13. ISO

13.1. 42001:2023 - Information technology — Artificial intelligence — Management system

Link: <https://www.iso.org/standard/42001>

Link: <https://www.iso.org/standard/42001>

13.1.1. 5.1

Leadership and commitment

13.1.2. 5.3

Roles, responsibilities and authorities

14. ISO/IEC

14.1. DIS 24970

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:24970:dis:ed-1:v1:en>

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:24970:dis:ed-1:v1:en>

14.1.1. 7.4

Triggers from human oversight

15. MIC/METI (Japan)

15.1. AI Guidelines for Business

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf

15.1.1. Human-Centric - 2 (d)

Carefully handle AI outputs, especially when they can be relevant to procedures that might significantly affect the society, such as an election and decision-making in a community.

15.1.2. Safety - 1 (c)

Ensure controllability that allows humans to control AI as necessary including objective monitoring and handling, in accordance with the characteristics and purposes of the relevant AI, in light of the severity and possibility of rights violations that can result from AI use or unintended AI behaviors.

15.1.3. Fairness - 2 (a)

To prevent AI from generating unfair results, consider implementing timely human interventions, rather than letting AI make the decisions alone.

15.1.4. Accountability - 3

Appoint someone as the person responsible for executing its accountability in each AI business actor.

15.1.5. Accountability - 4

As for responsibilities shared among actors, clarify who take the responsibilities through contracts or social promises (voluntary commitments) between AI business actors including non-business users.

16. Central Bank of the UAE

Link: <https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence>

Link: <https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence>

16.1.1. 2. Governance and Accountability - b

Senior management and the Board of Directors of LFIs should be responsible and accountable for AI and ML systems and outcomes, model selection/development, deployment, accountability, appropriate human resourcing and oversight and monitoring and management on an on-going basis.

16.1.2. 6. Continuous Monitoring and Review - d

LFIs should remain responsible for outsourced AI functions and should consider: appropriate contractual rights with respect to audit and information rights from providers, be made aware of any material developments with the AI provider, appropriate termination/cease provisions, data protection, cyber security, performance guarantees, its compliance with laws/regulations and standards and any material developments with regard to the AI being outsourced/utilized.

16.1.3. 6. Continuous Monitoring and Review - e

LFIs should at all times retain the clear and immediate ability, with human intervention, to cease use of an AI model system, technology or application deployed or utilized.

16.1.4. 7. Human Oversight and Consumer Protection - a

LFIs should ensure that AI and ML systems operate under meaningful human oversight and judgement, particularly for decisions that have significant implications for consumers and in respect of the ongoing selection of, determination as to third party providers of, deployment of and ongoing monitoring and general use of AI. Human oversight may be exercised through different models:

- (i) Human-in-the-loop – where a AI provides recommendations but a human decision maker retains full authority to approve or reject the outcome;
- (ii) Human-on-the-loop – where the AI works autonomously for routine tasks, while a human monitors outcomes and can intervene where necessary;
- (iii) Human-out-of-the-loop – where the AI operates without direct human involvement, which should only be utilised for low-risk, non-material processes with appropriate controls in place.

16.1.5. 7. Human Oversight and Consumer Protection - b

The level of human involvement should be commensurate with the identified and potential risks posed to a consumer by any AI.

16.1.6. 7. Human Oversight and Consumer Protection - c

Consumers should be able to request human review or explanation of AI generated decisions, and alternative arrangements should be available where a customer does not wish to be subject to an AI decision. LFIs should maintain clear and accessible channels for complaints and redress in line with Article 8 of the Consumer Protection Regulation. Consumers should be informed of their right to challenge decisions, correct inaccurate data inputs having impact on AI and the process to challenge data and decisions by AI. A clear complaints-handling procedure/policy should be created, provided and accessible by customers on a regular basis. Complaints should be addressed in person, efficiently, confidentially and in as short at time as is reasonable in the circumstances.

16.1.7. 8. Integration with Existing Frameworks - b

Senior management should ensure that policies and procedures for AI adoption complement, rather than duplicate, existing regulatory obligations under the Consumer Protection Regulation and other CBUAE directives. For example, consumer risk arising from AI-driven models should be treated as part of the conduct risk framework, with appropriate reporting to the board and regulators.

17. Smart Dubai (UAE)

17.1. AI Ethics Principles & Guidelines

Link: <https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf>

Link: <https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf>

17.1.1. 1.2.1.1

Accountability for loss or damages resulting from the application of AI systems should not be attributed to the system itself.

17.1.2. 1.2.1.2

AI operator organisations and AI developer organisations should consider designating individuals to be responsible for investigating and rectifying the cause of loss or damage arising from the deployment of AI systems.

17.1.3. 1.2.4.1

AI operator organisations which use AI systems to inform significant decisions should provide procedures by which affected AI subjects can challenge a specific decision concerning them.

17.1.4. 1.2.4.2

AI operator organisations should consider such procedures even for non-significant decisions.

17.1.5. 1.2.4.4

AI operator organisations should consider employing human case evaluators to review any such challenges and, when appropriate, overturn the challenged decision.

17.1.6. 1.2.4.5

AI operator organisations should consider instituting an opt-out mechanism for significant automated decisions.

18. SDAIA (Saudi Arabia)

18.1. AI Ethics Principles

Link: <https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf>

Link: <https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf>

18.1.1. Principle 5 – Reliability & Safety - Plan and Design - 5

All critical decision points in the system design should be subject to sign-off by relevant stakeholders to minimize risks and make stakeholders accountable for the decisions.

18.1.2. Principle 5 – Reliability & Safety - Build and Validate - 2

To ensure the technical robustness of an AI system rigorous testing, validation, and re-assessment as well as the integration of adequate mechanisms of oversight and controls into its development is required. System integration test sign-off should be done with relevant stakeholders to minimize risks and liability.

18.1.3. Principle 5 – Reliability & Safety - Build and Validate - 3

Automated AI systems involving scenarios where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life-and-death decisions should trigger human oversight and final determination. Furthermore, AI systems should not be used for social scoring or mass surveillance purposes.

18.1.4. Principle 7 – Accountability & Responsibility - Plan and Design - 1

This step is crucial to design or procure an AI System in an accountable and responsible manner. The ethical responsibility and liability for the outcomes of the AI system should be attributable to stakeholders who are responsible for certain actions in the AI System Lifecycle. It is essential to set a robust governance structure that defines the authorization and responsibility areas of the internal and external stakeholders without leaving any areas of uncertainty to achieve this principle. The design approach of the AI system should respect human rights, and fundamental freedoms as well as the national laws and cultural values of the kingdom.

18.1.5. Principle 7 – Accountability & Responsibility - Plan and Design - 3

It is essential to build and design a human-controlled AI system where decisions on the processes and functionality of the technology are monitored and executed, and are susceptible to intervention from authorized users. Human governance and oversight establish the necessary control and levels of autonomy through set mechanisms.

18.1.6. Principle 7 – Accountability & Responsibility - Build and Validate - 1

Model development of the AI system and algorithm should consist of the selection of features, hyperparameter tuning and performance metric selection. To achieve this, the technical stakeholders who build and validate models should be responsible for these decisions.

18.1.7. Principle 7 – Accountability & Responsibility - Build and Validate - 2

Assigning the appropriate ownership and communicating responsibilities will set the tone for accountability that would aid in steering the development of the AI system on good reasons, solid interference, and will allow the intervention of human critical judgement and expertise.

18.1.8. Principle 7 – Accountability & Responsibility - Build and Validate - 4

The appropriate stakeholders and owners of the AI technology should review and sign off the model after successful testing and validation of user acceptance testing rounds have been conducted and completed before the AI models can be productionized.

19. CISA

19.1. Principles for the Secure Integration of Artificial Intelligence in Operational Technology

Link: <https://www.cisa.gov/sites/default/files/2026-01/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508cV2.pdf>

Link: <https://www.cisa.gov/sites/default/files/2026-01/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508cV2.pdf>

19.1.1. 3.1.1 Establish Governance Mechanisms for AI in OT

Effective governance structures are essential for the safe and secure integration of AI into OT environments. This involves establishing clear policies, procedures, and accountability structures for AI decision-making processes within OT. An AI governance structure should include the key stakeholders listed below, as well as any AI vendors needed for maintaining oversight during procurement, development, design, deployment, and operations.

19.1.2. 3.1.3 - Establishing clear roles and responsibilities

Ensure everyone involved in the development, deployment, and operations and maintenance of AI systems (e.g., data owners, model developers, and end users) understands their tasks and expectations—and to avoid liability and confusion over stakeholder responsibilities in the event of safety or operational incidents

19.1.3. 4.1.1 - Human-in-the-Loop Decision-Making

Provide adequate transparency that involves operators and engineers in decision-making, especially for critical OT operations and actions. For more passive AI systems, operators and engineers can implement this by incorporating the recommendations into an existing change management process. Use caution with active AI systems directly influencing control, as problems can escalate before operators become aware of them. Where AI is actively updating control logic, use safety thresholds, alternative sensor output, or state changes that add human-in-the-loop intervention points.

20. Cloud Security Alliance (CSA)

20.1. AI Controls Matrix

Link: <https://cloudsecurityalliance.org/artifacts/ai-controls-matrix>

Link: <https://cloudsecurityalliance.org/artifacts/ai-controls-matrix>

20.1.1. GRC-06

Define and document roles and responsibilities for planning, implementing, operating, assessing, and improving governance programs.

20.1.2. GRC-12

Establish an ethics committee to review AI applications, ensuring alignment with ethical standards and organizational values.

20.1.3. GRC-15

Establish, execute, and assess processes, procedures, and technical measures to ensure human oversight and control of the AI system in compliance with regulatory requirements and organizational risk management.

20.1.4. HRS-09

Document and communicate roles and responsibilities of employees, as they relate to information assets and security.

20.1.5. STA-02

Establish, document, approve, communicate, apply, evaluate and maintain policies and procedures for the application of the Shared Security Responsibility Model (SSRM) within the organization. Review and update the policies and procedures at least annually, or upon significant changes.

20.1.6. STA-03

Apply, document, implement and manage the SSRM throughout the supply chain.

20.1.7. STA-06

Review and validate SSRM documentation.

20.1.8. STA-07

Implement, operate, and audit or assess the portions of the SSRM which the organization is responsible for.

21. CoSAI

21.1. AI Incident Response Framework

Link: <https://github.com/cosai-oasis/ws2-defenders/blob/main/incident-response/AI%20Incident%20Response.md>

Link: <https://github.com/cosai-oasis/ws2-defenders/blob/main/incident-response/AI%20Incident%20Response.md>

21.1.1. 3.3.2. Detection and Analysis Phase - Detection Mechanisms - Manual Review

- Human review procedures for flagged interactions
- Sampling of high-risk operations
- Security dashboards
- Escalation triggers

21.2. Model Context Protocol (MCP) Security

Link: <https://github.com/cosai-oasis/ws4-secure-design-agentic-systems/blob/main/model-context-protocol-security.md>

Link: <https://github.com/cosai-oasis/ws4-secure-design-agentic-systems/blob/main/model-context-protocol-security.md>

21.2.1. 3.2.9 Human-in-the-loop

There is the possibility that a large language model, legit or poisoned, decides to execute a tool in a dangerous way. MCP hosts and clients, in general, allow users to disable the confirmation prompt. There are two approaches organizations considering this risk unacceptable may implement to reduce its probability and impact:

enforce the use of MCP hosts and clients with a configuration that unprivileged users cannot change and that keeps the confirmation prompt enabled.
use elicitation on the MCP server side to request the user confirmation of actions.

22. Personal Data Protection Commission Singapore (PDPC)

22.1. Model Artificial Intelligence Governance Framework Second Edition

Link: <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf>

Link: <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf>

22.1.1. 1. Clear roles and responsibilities for the ethical deployment of AI - a)

Responsibility for and oversight of the various stages and activities involved in AI deployment should be allocated to the appropriate personnel and/or departments. If necessary and possible, consider establishing a coordinating body, having relevant expertise and proper representation from across the organisation.

22.1.2. 1. Clear roles and responsibilities for the ethical deployment of AI - c) (i)

Key roles and responsibilities that can be allocated include:
Using any existing risk management framework and applying risk control measures (see "Risk management and internal controls" below) to:

- Assess and manage the risks of deploying AI, including any potential adverse impact on the individuals (e.g. who are most vulnerable, how are they impacted, how to assess the scale of the impact, how to get feedback from those impacted, etc.).
- Decide on the appropriate level of human involvement in AI-augmented decision-making.
- Manage the AI model training and selection process

22.1.3. ALGORITHM AND MODEL - c)

Supplementary explanation tools are helpful for explaining AI models, especially models that are less interpretable (also known as "black box" systems). These tools help make the underlying rationale of an AI system's output more interpretable and intelligible to those who use the system. It is possible to use a combination of these tools to improve the explainability of an AI model's decision.

23. MITRE

23.1. ATLAS Framework

Link: <https://atlas.mitre.org/mitigations>

Link: <https://atlas.mitre.org/mitigations>

23.1.1. AML.M0029 - Human In-the-Loop for AI Agent Actions

Systems should require the user or another human stakeholder to approve AI agent actions before the agent takes them. The human approver may be technical staff or business unit SMEs depending on the use case. Separate tools, such as dedicated audit agents, may assist human approval, but final adjudication should be conducted by a human decision-maker.

The security benefits from Human In-the-Loop policies may be at odds with operational overhead costs of additional approvals. To ease this, Human In-the-Loop policies should follow the degree of consequence of the task at hand. Minor, repetitive tasks performed by agents accessing basic tools may only require minimal human oversight, while agents employed in systems with significant consequences may necessitate approval from multiple stakeholders diversified across multiple organizations.

24. NIST

24.1. AI RMF 1.0

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

24.1.1. GOVERN 2.3

Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.

24.1.2. GOVERN 3.2

Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.

24.1.3. MAP 1.2

Interdisciplinary AI actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.

24.1.4. MAP 3.5

Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the GOVERN function.

24.2. SP 800-218A

Link: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf>

Link: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf>

24.2.1. PO.2.1

Create new roles and alter responsibilities for existing roles as needed to encompass all parts of the SDLC. Periodically review and maintain the defined roles and responsibilities, updating them as needed.

Include AI model development security in SDLC-related roles and responsibilities throughout the SDLC. The roles and responsibilities should include, but are not limited to, AI model development, AI model operations, and data science.

24.2.2. PW.7.1

Determine whether code review (a person looks directly at the code to find issues) and/or code analysis (tools are used to find issues in code, either in a fully automated way or in conjunction with a person) should be used, as defined by the organization.

Code review and analysis policies or guidelines should include code for AI models and other related components.

24.3. IR 8596: Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile

Link: <https://csrc.nist.gov/pubs/ir/8596/iprd>

Link: <https://csrc.nist.gov/pubs/ir/8596/iprd>

24.3.1. GV.RR-01

Organizational leadership is responsible and accountable for cybersecurity risk and fosters a culture that is risk-aware, ethical, and continually improving

24.3.2. GV.RR-02

Roles, responsibilities, and authorities related to cybersecurity risk management are established, communicated, understood, and enforced

24.3.3. GV.RR-03

Adequate resources are allocated commensurate with the cybersecurity risk strategy, roles, responsibilities, and policies

24.3.4. GV.SC-02

Cybersecurity roles and responsibilities for suppliers, customers, and partners are established, communicated, and coordinated internally and externally