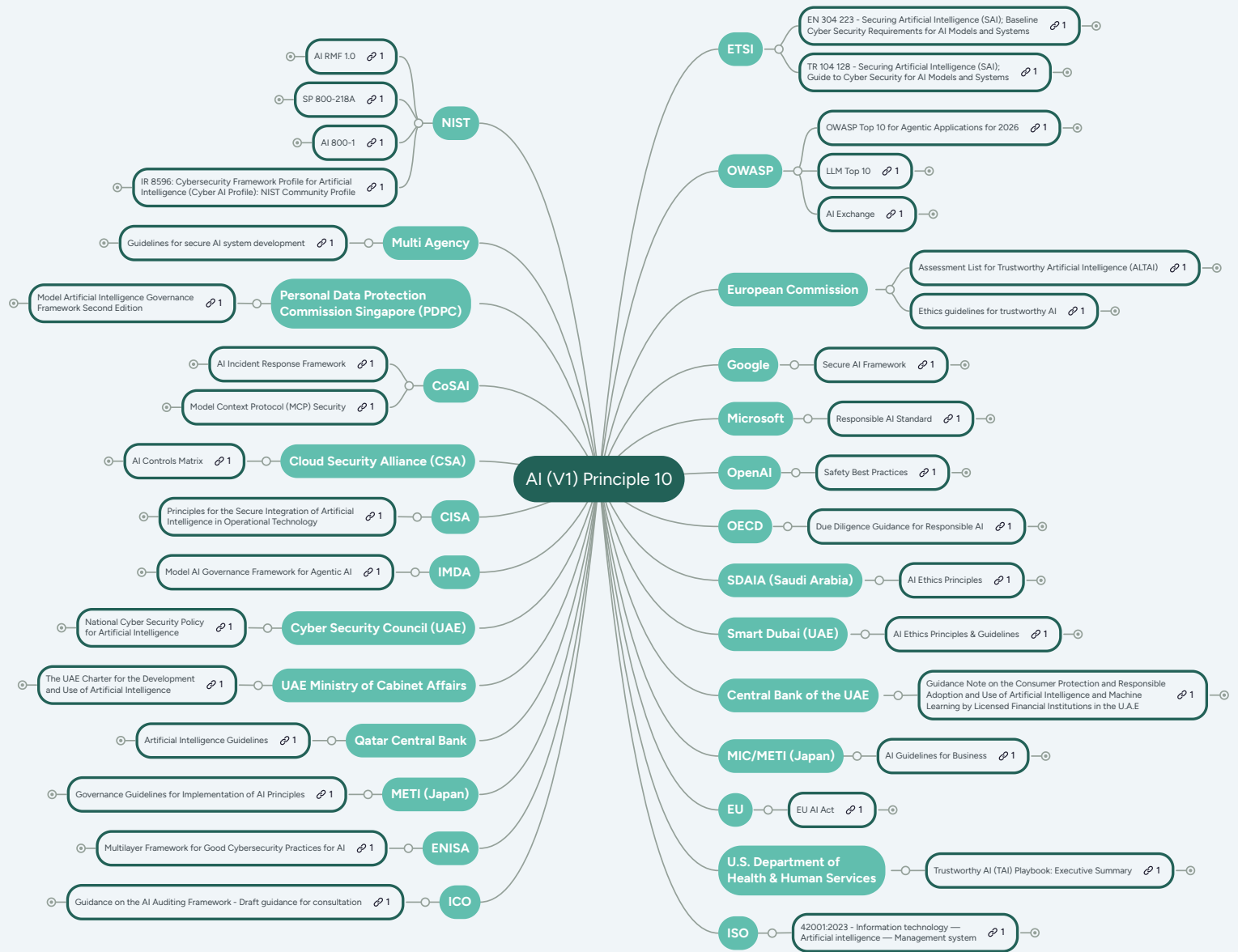




AI (V1) Principle 10

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 5.3.1-1

System Operators shall convey to End-users in an accessible way where and how their data will be used, accessed and stored (for example, if it is used for model retraining, or reviewed by employees or partners). If the Developer is an external entity, they shall provide this information to System Operators.

1.1.2. Provision 5.3.1-2

System Operators shall provide End-users with accessible guidance to support their use, management, integration, and configuration of AI systems. If the Developer is an external entity, they shall provide all necessary information to help System Operators.

1.1.3. Provision 5.3.1-2.1

System Operators shall include guidance on the appropriate use of the model or system, which includes highlighting limitations and potential failure modes.

1.1.4. Provision 5.3.1-2.2

System Operators shall proactively inform End-users of any security relevant updates and provide clear explanations in an accessible way.

1.1.5. Provision 5.3.1-3

Developers and System Operators should support End-users and Affected Entities during and following a cyber security incident to contain and mitigate the impacts of an incident. The process for undertaking this should be documented and agreed in contracts with End-users.

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 5.3.1-1

"System Operators shall convey to End-users in an accessible way where and how their data will be used, accessed, and stored (for example, if it is used for model retraining, or reviewed by employees or partners). If the Developer is an external entity, they shall provide this information to System Operators." (ETSI TS 104 223 [i.1])

Related threats/risks:

Insufficient communication about data usage, access, or storage practices can lead to misunderstandings and mistrust among End-users. This lack of transparency increases the risk of misuse or unauthorized access to data, as users might not fully understand or consent to how their data is handled, potentially resulting in regulatory and reputational damage.

Example Measures/Controls:

Establish Transparent Data Usage Communication: Provide a transparent overview of data usage policies, purpose of processing, specifying whether data will be used for model retraining, third-party access, or employee review. Each processing activity needs to be logged, and a compatibility assessment needs to be made for each new purpose. Ensure end users understand all potential uses of their data and how it contributes to AI model performance or security and that documentation is in an accessible format

1.2.2. Provision 5.3.1-2

"System Operators shall provide End-users with accessible guidance to support their use, management, integration, and configuration of AI systems. If the Developer is an external entity, they shall provide all necessary information to help System Operators." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without clear guidance, End-users might mismanage the software, leading to data leaks, vulnerabilities, or system misuse, exposing AI systems to security risks.

Example Measures/Controls:

Provide Comprehensive User Guides and Tutorials: Develop and distribute detailed user guides and tutorials that cover secure configuration, integration steps, and recommended usage practices, ensuring users understand safe operation protocols. Provide accessible notification options, such as screen reader-compatible text alerts, and customisable notifications for users with sensory impairments.

1.2.3. Provision 5.3.1-2.1

"System Operators shall include guidance on the appropriate use of the model or system, which includes highlighting limitations and potential failure modes." (ETSI TS 104 223 [i.1])

Related threats/risks:

Failing to highlight limitations and potential failure modes can lead to End-users relying on the AI system for unsupported or inappropriate tasks, increasing the risk of operational errors, data misuse, or unintended consequences.

Example Measures/Controls:

Highlight Model Limitations and Failure Modes: Clearly outline the model's appropriate uses, limitations, and potential failure modes to inform end-users of scenarios in which the model might produce inaccurate or unreliable outputs.

1.2.4. Provision 5.3.1-2.2

"System Operators shall proactively inform End-users of any security relevant updates and provide clear explanations in an accessible way." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without proactive communication about security-relevant updates, End-users can fail to understand changes in system behaviour or associated risks. This lack of awareness can lead to improper use or failure to take necessary precautions, increasing the system's exposure to potential exploitation or security breaches.

Example Measures/Controls:

Notify Users of Security Updates: Proactively inform End-users about security update, detailing the purpose and impact of each update to promote user compliance. Ensure all users, can be informed by using accessible formats.

1.2.5. Provision 5.3.1-3

"Developers and System Operators should support affected End-users and Affected Entities during and following a cyber security incident to contain and mitigate the impacts of an incident. The process for undertaking this should be documented and agreed in contracts with End-users." (ETSI TS 104 223 [i.1])

Related threats/risks:

Failure to support affected End-users and Affected Entities during an incident can result in prolonged recovery times, mismanagement of containment efforts, and increased reputational damage due to inadequate communication or guidance.

Example Measures/Controls:

Establish a Documented Incident Support and Communication Process: Develop and document a support process for responding to incidents, covering steps for containment, impact assessment, and recovery, and specify the roles and responsibilities of Developers and System Operators. This should include specialist skills and AI expertise that might be require. Provide support through accessible formats, ensuring usability for all affected stakeholders.

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. ASI09: Human-Agent Trust Exploitation - 4

Allow reporting of suspicious interactions: In user-interactive systems, provide plain-language risk summary (not model-generated rationales) and a clear option for users to flag suspicious or manipulative agent behavior, triggering automated review or a temporary lockdown of agent capabilities.

2.1.2. ASI09: Human-Agent Trust Exploitation - 8

Human-factors and UI safeguards: Visually differentiate high-risk recommendations using cues such as red borders, banners, or confirmation prompts, and periodically remind users of manipulation patterns and agent limitations. Where appropriate, avoid persuasive or emotionally manipulative language in safety-critical flows. Maintain appropriate training and assessment of personnel to ensure familiarity and consistency of perception of human-factors and UI.

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. LLM02: Sensitive Information Disclosure - 7

Provide guidance on avoiding the input of sensitive information. Offer training on best practices for interacting with LLMs securely.

2.2.2. LLM02: Sensitive Information Disclosure - 8

Maintain clear policies about data retention, usage, and deletion. Allow users to opt out of having their data included in training processes.

2.2.3. LLM02: Sensitive Information Disclosure - 9

Limit the ability for users to override or access the system's initial settings, reducing the risk of exposure to internal configurations.

2.2.4. LLM09: Misinformation - 3

Encourage users to cross-check LLM outputs with trusted external sources to ensure the accuracy of the information. Implement human oversight and fact-checking processes, especially for critical or sensitive information. Ensure that human reviewers are properly trained to avoid overreliance on AI-generated content.

2.2.5. LLM09: Misinformation - 5

Identify the risks and possible harms associated with LLM-generated content, then clearly communicate these risks and limitations to users, including the potential for misinformation.

2.2.6. LLM09: Misinformation - 7

Design APIs and user interfaces that encourage responsible use of LLMs, such as integrating content filters, clearly labeling AI-generated content and informing users on limitations of reliability and accuracy. Be specific about the intended field of use limitations.

2.2.7. LLM09: Misinformation - 8

Provide comprehensive training for users on the limitations of LLMs, the importance of independent verification of generated content, and the need for critical thinking. In specific contexts, offer domain-specific training to ensure users can effectively evaluate LLM outputs within their field of expertise.

2.3. AI Exchange

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

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

2.3.1. 1.3. Controls to limit the effects of unwanted behaviour - AI TRANSPARENCY

AI transparency: Informing users on the AI system's properties to enable them to adjust how they rely on it, what data they are willing to send to it, and what additional mitigations to apply. These AI system properties can include:

Rough working of the model

The training approach

Type of data used and the source

Expected accuracy and robustness of the AI system's output

Any residual (security) risks

Note that transparency here is about providing abstract information regarding the AI system and is therefore something else than explainability of model decisions. The simplest form of transparency is to inform users that an AI model is being involved. This is for example required by the EU AI Act for chatbots.

See the DISCRETE control for the balance between being transparent and being discrete about the model.

Example: Informing users that when they choose an agent to perform a task, that the agent could be manipulated if it reads untrusted data and what consequences that could have (residual security risk) - followed by a recommendation to configure the permissions of the agent to the minimal set for the task.

2.3.2. 1.3. Controls to limit the effects of unwanted behaviour - EXPLAINABILITY

Explainability: Explaining how individual model decisions are made, a field referred to as Explainable AI (XAI), can aid in gaining user trust in the model. In some cases, this can also prevent overreliance, for example, when the user observes the simplicity of the 'reasoning' or even errors in that process. See this Stanford article on explainability and overreliance. Explanations of how a model works can also aid security assessors to evaluate AI security risks of a model.

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 #4 Transparency

A crucial component of achieving Trustworthy AI is transparency which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system.

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. 1.1.2 Human Agency

Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals. AI systems can sometimes be deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy. The overall principle of user autonomy must be central to the system's functionality. Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them.

3.2.2. 1.4.3 Communication

AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations.

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. User Data Management

Store, process, and use all user data (e.g. prompts and logs) from AI applications in compliance with user consent.

4.1.2. User Transparency and Controls

Inform users of relevant AI risks with disclosures, and provide transparency and control experiences for use of their data in AI applications.

4.1.3. User Policies and Education

Publish easy to understand AI security and privacy policies and education for users.

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. A3.7

If an intended use is not supported by evidence, or if evidence comes to light that refutes that the system is fit for purpose for the intended use at any point in the system's use:

- 1) remove the intended use from customer-facing materials and make current customers aware of the issue, take action to close the identified gap, or discontinue the system,
- 2) revise documentation related to the intended use, and
- 3) publish the revised documentation to customers.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

5.1.2. A3.8

Communicate with care about system benefits; follow any applicable guidance from your attorney.

5.1.3. T1.2

Design the system, including, when possible, the system UX, features, reporting functions, and educational materials, so that stakeholders identified in requirement T1.1 can:

- 1) understand the system's intended uses,
- 2) interpret relevant system behavior effectively (i.e., in a way that supports informed decision making), and
- 3) remain aware of the possible tendency of over-relying on outputs produced by the system ("automation bias").

For the two categories of stakeholders identified in requirement T1.1, document:

- 1) how the system design will support their understanding of the system's intended uses, and
- 2) how the system aids their ability to interpret relevant system responses, and
- 3) how the system design discourages automation bias.

5.1.4. T3.2

Design the system, including system UX, features, reporting functions, educational materials, and outputs so that stakeholders identified in T3.1 will be informed of the type of AI system they are interacting with or exposed to. Ensure that any image, audio, or video outputs that are intended to be used outside the system are labelled as being produced by AI.

5.1.5. T3.3

Define and document the method to be used to evaluate whether each stakeholder identified in T3.1 is informed of the type of AI system they are interacting with or exposed to.

5.1.6. F1.9

Publish information for customers about:

- 1) identified demographic groups for which performance may not meet any target minimum performance level,
- 2) any remaining performance disparities between identified demographic groups that may exceed the target maximum, and
- 3) any justifiable factors that account for these performance levels and differences.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

5.1.7. RS2.4

Provide training and documentation for system owners, developers, customer support and other stakeholders responsible for managing the system to support their remediation and mitigation of predictable failures identified in requirement RS2.1. Document the training and documentation provided.

5.1.8. RS3.7

If evidence comes to light that refutes the system is fit for purpose for an intended use at any point in the system's use:

- 1) remove the intended use from customer-facing materials and make current customers aware of the issue, take action to close the identified gap, or discontinue the system,
- 2) revise documentation related to the intended use, and
- 3) publish the revised documentation to customers.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

6. OpenAI

6.1. Safety Best Practices

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

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

6.1.1. Allow users to report issues

Users should generally have an easily-available method for reporting improper functionality or other concerns about application behavior (listed email address, ticket submission method, etc). This method should be monitored by a human and responded to as appropriate.

7. OECD

7.1. Due Diligence Guidance for Responsible AI

Link: https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/oecd-due-diligence-guidance-for-responsible-ai_7831bb49/41671712-en.pdf

Link: https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/oecd-due-diligence-guidance-for-responsible-ai_7831bb49/41671712-en.pdf

7.1.1. Step 5 - Communicate actions to address impacts

Communicate externally relevant information on due diligence policies, processes, activities conducted to identify and address actual or potential adverse impacts, including the findings and outcomes of those activities. Communication could take a variety of forms depending on the target audience (e.g., stakeholder consultations and public communication through the enterprise's annual, sustainability or corporate responsibility reports or other appropriate forms of disclosure required by legislation or voluntary initiatives).

8. SDAIA (Saudi Arabia)

8.1. 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>

8.1.1. Principle 6 – Transparency & Explainability - Plan and Design - 1

When designing a transparent and trusted AI system, it is vital to ensure that stakeholders affected by AI systems are fully aware and informed of how outcomes are processed. They should further be given access to and an explanation of the rationale for decisions made by the AI technology in an understandable and contextual manner. Decisions should be traceable. AI system owners must define the level of transparency for different stakeholders on the technology based on data privacy, sensitivity, and authorization of the stakeholders.

8.1.2. Principle 6 – Transparency & Explainability - Plan and Design - 2

The AI system should be designed to include an information section in the platform to give an overview of the AI model decisions as part of the overall transparency application of the technology. Information sharing as a sub-principle should be adhered to with end-users and stakeholders of the AI system upon request or open to the public, depending on the nature of the AI system and target market. The model should establish a process mechanism to log and address issues and complaints that arise to be able to resolve them in a transparent and explainable manner.

8.1.3. Principle 6 – Transparency & Explainability - Build and Validate - 2

Transparent and explainable algorithms ensure that stakeholders affected by AI systems, both individuals and communities, are fully informed when an outcome is processed by the AI system by providing the opportunity to request explanatory information from the AI system owner. This enables the identification of the AI decision and its respective analysis which facilitates its auditability as well as its explainability.

8.1.4. Principle 6 – Transparency & Explainability - Deploy and Monitor - 1

Upon deployment of the AI system, performance metrics relating the AI system's output, accuracy and alignment to priorities and objectives, as well as its measured impact on individuals and communities should be documented, available and accessible to stakeholders of the AI technology.

8.1.5. Principle 6 – Transparency & Explainability - Deploy and Monitor - 2

Information on any system failures, data breaches, system breakdowns, etc. should be logged and stakeholders should be informed about these instances keeping the performance and execution of the AI system transparent. Periodic UI and UX testing should be conducted to avoid the risk of confusion, confirmation of biases, or cognitive fatigue of the AI system.

9. Smart Dubai (UAE)

9.1. 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>

9.1.1. 1.2.3.2

In the case that critical decisions are of civic interest, public release of the results of the audit should be considered as a means of ensuring public processes remain accountable to those affected by them.

9.1.2. 1.2.4.3

AI operator organisations should make affected AI subjects aware of these procedures, and should design them in a convenient and user-friendly way.

9.1.3. 1.2.5.1

When informing an AI subject about significant choices they will make, AI systems should not unreasonably restrict the available options or otherwise attempt to influence their value judgements without the explicit consent of the AI subject in question.

9.1.4. 1.2.7.2

AI developer organisations should consider notifying customers and AI operator organisations of the use cases for which the system has been designed, and those for which it is not suitable.

9.1.5. 1.3.2.1

AI operator organisations should inform AI subjects when a significant decision affecting them has been made by an AI system.

9.1.6. 1.3.2.2

If an AI system can convincingly impersonate a human being, it should do so only after notifying the AI subject that it is an AI system.

9.1.7. 1.4.1.1

AI operator organisations could consider informing the affected AI subjects in understandable, non-technical language of:

- the data that is ingested by the system;
- the types of algorithms employed;
- the categories into which people can be placed, and;
- the most important features driving the outcomes of decisions

9.1.8. 1.4.1.3

AI operator organisations should consider providing affected AI subjects with a means to request explanations for specific significant decisions, to the extent possible given the state of present research and the choice of model.

9.1.9. 1.4.2.1

AI operator organisations should consider providing a means by which people affected by a significant decision informed by AI can access the reasoning behind that decision.

10. Central Bank of the UAE

10.1. Guidance Note on the Consumer Protection and Responsible Adoption and Use of Artificial Intelligence and Machine Learning by Licensed Financial Institutions in the U.A.E

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>

10.1.1. 4. Transparency and Explain ability - a

LFIs should be transparent with customers and relevant stakeholders about the use of AI, particularly in respect of high-impact decisions, and if they are communicating or interacting with an AI application. LFIs should be clear as to how AI systems operate and make decisions and be able to disclose the same.

10.1.2. 4. Transparency and Explain ability - d

LFIs should provide customers with meaningful information regarding the logic of AI decisions and make available mechanisms for customers to seek clarification or redress.

11. MIC/METI (Japan)

11.1. 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

11.1.1. Human-Centric - 4

In addition to ensuring fairness, to prevent information poverty and digital poverty and allow more people to enjoy the benefits of AI, pay attention to make it easy for socially vulnerable people to use AI. Adopt universal design, ensure accessibility, and provide relevant stakeholders with education and support.

11.1.2. Human-Centric - 5

Offer rational information about the functions and peripheral technologies of the AI system or service, and allow users to use functions that timely and appropriately offer the information for judging choices. For example, default settings, provision of understandable options, provision of feedbacks, alerts in an emergency, and handling of errors.

11.1.3. Transparency - 2 (a)

Based on the relations with AI and the nature and purpose of AI, provide and explain information summarizing the items listed below according to the knowledge and ability of each stakeholder.

- AI systems and services in general
- Fact that AI is used and its scope
- Methods for data collection and annotation
- Methods for training and evaluation
- Information on the underlying AI models
- Capabilities and limitations of the AI system or service, and proper/improper use by AI business users
- Relevant laws applicable in the country/region where those provided with the AI system or service or AI business users are located

11.1.4. Transparency - 2 (b)

Encourage a variety of stakeholders to engage actively through dialogues and collect various opinions on social impacts and safety.

11.1.5. Transparency - 2 (c)

In addition, show actual advantages of providing or using the AI system or service and risks to relevant stakeholders.

11.1.6. Transparency - 3 (a)

Provision of information to stakeholders described above "(2) Providing relevant stakeholders with information" doesn't assume disclosure of algorithms or source code, but it assumes providing them to the extent that satisfies social rationality based on the characteristics and uses of the technologies to be adopted while respecting privacy and trade secrets.

11.1.7. Transparency - 4

Share necessary explanations for those to be explained with actors who explain to analyze requirements of such explanation to gain relevant stakeholders' understanding and sense of safety to provide proof of AI operations.

AI provider: Inform the AI developer about things that are required to be explained.

AI business user: Inform the AI developer and AI provider about things that are required to be explained.

11.1.8. Accountability - 5 (b)

As necessary, set opportunities for accepting comments from stakeholders on incorrect AI output and the like, and conduct objective monitoring of the output.

11.1.9. Accountability - 5 (c)

Set policies to handle cases that might affect the interests of stakeholders. Execute those policies reliably and report the progress regularly to the stakeholders as necessary.

11.1.10. Education/literacy - 3

To improve the safety of the whole AI system or AI service, provide stakeholders with education and literacy advancement as necessary.

11.1.11. Innovation - 3

Provide necessary information to the extent that does not hinder the innovation of the information provider.

12. EU

12.1. 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

12.1.1. 13.2 Transparency and Provision of Information to Deployers

High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers.

12.1.2. 13.3 Transparency and Provision of Information to Deployers

The instructions for use shall contain at least the following information:

- (a) the identity and the contact details of the provider and, where applicable, of its authorised representative;
- (b) the characteristics, capabilities and limitations of performance of the high-risk AI system, including:
 - (i) its intended purpose;
 - (ii) the level of accuracy, including its metrics, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;
 - (iii) any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights referred to in Article 9(2);
 - (iv) where applicable, the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output;
 - (v) when appropriate, its performance regarding specific persons or groups of persons on which the system is intended to be used;
 - (vi) when appropriate, specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk AI system;
 - (vii) where applicable, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately;
- (c) the changes to the high-risk AI system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment, if any;
- (d) the human oversight measures referred to in Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployers;
- (e) the computational and hardware resources needed, the expected lifetime of the high-risk AI system and any necessary maintenance and care measures, including their frequency, to ensure the proper functioning of that AI system, including as regards software updates;
- (f) where relevant, a description of the mechanisms included within the high-risk AI system that allows deployers to properly collect, store and interpret the logs in accordance with Article 12.

12.1.3. 15.3 Accuracy, Robustness and Cybersecurity

The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.

12.1.4. 26.7 Obligations of deployers of high-risk AI systems

Before putting into service or using a high-risk AI system at the workplace, deployers who are employers shall inform workers' representatives and the affected workers that they will be subject to the use of the high-risk AI system. This information shall be provided, where applicable, in accordance with the rules and procedures laid down in Union and national law and practice on information of workers and their representatives.

12.1.5. 26.8 Obligations of deployers of high-risk AI systems

Deployers of high-risk AI systems that are public authorities, or Union institutions, bodies, offices or agencies shall comply with the registration obligations referred to in Article 49. When such deployers find that the high-risk AI system that they envisage using has not been registered in the EU database referred to in Article 71, they shall not use that system and shall inform the provider or the distributor.

12.1.6. 26.11 Obligations of deployers of high-risk AI systems

Without prejudice to Article 50 of this Regulation, deployers of high-risk AI systems referred to in Annex III that make decisions or assist in making decisions related to natural persons shall inform the natural persons that they are subject to the use of the high-risk AI system. For high-risk AI systems used for law enforcement purposes Article 13 of Directive (EU) 2016/680 shall apply.

12.1.7. 50.2 Transparency obligations for providers and deployers of certain AI systems

Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate or prosecute criminal offences.

12.1.8. 50.4 Transparency obligations for providers and deployers of certain AI systems

Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated. This obligation shall not apply where the use is authorised by law to detect, prevent, investigate or prosecute criminal offence. Where the content forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme, the transparency obligations set out in this paragraph are limited to disclosure of the existence of such generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work.

Deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated. This obligation shall not apply where the use is authorised by law to detect, prevent, investigate or prosecute criminal offences or where the AI-generated content has undergone a process of human review or editorial control and where a natural or legal person holds editorial responsibility for the publication of the content.

12.1.9. 50.5 Transparency obligations for providers and deployers of certain AI systems

The information referred to in paragraphs 1 to 4 shall be provided to the natural persons concerned in a clear and distinguishable manner at the latest at the time of the first interaction or exposure. The information shall conform to the applicable accessibility requirements.

12.1.10. 53.1 Obligations for providers of general purpose AI models

Providers of general-purpose AI models shall:

- (a) draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation, which shall contain, at a minimum, the information set out in Annex XI for the purpose of providing it, upon request, to the AI Office and the national competent authorities;
- (b) draw up, keep up-to-date and make available information and documentation to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems. Without prejudice to the need to observe and protect intellectual property rights and confidential business information or trade secrets in accordance with Union and national law, the information and documentation shall:
 - (i) enable providers of AI systems to have a good understanding of the capabilities and limitations of the general-purpose AI model and to comply with their obligations pursuant to this Regulation; and
 - (ii) contain, at a minimum, the elements set out in Annex XII;
- (c) put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;
- (d) draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.

12.1.11. 53.2 Obligations for providers of general purpose AI models

The obligations set out in paragraph 1, points (a) and (b), shall not apply to providers of AI models that are released under a free and open-source licence that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. This exception shall not apply to general-purpose AI models with systemic risks.

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

13.1. 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>

13.1.1. Transparent / Explainable

All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection

14. ISO

14.1. 42001:2023 - Information technology — Artificial intelligence — Management system

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

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

14.1.1. 7.3

Awareness

14.1.2. 7.4

Communication

15. ICO

15.1. Guidance on the AI Auditing Framework - Draft guidance for consultation

Link: <https://ico.org.uk/media2/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf>

Link: <https://ico.org.uk/media2/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf>

15.1.1. What steps should we take to manage the risks of privacy attacks on AI models? - 3

As part of your procurement policy there should be sufficient information sharing between each party to perform your respective assessments as necessary. In some cases, ML model providers and clients will be joint controllers and therefore need to perform a joint risk assessment.

16. ENISA

16.1. Multilayer Framework for Good Cybersecurity Practices for AI

Link:

<https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf>

Link:

<https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf>

16.1.1. From the lab to the market 7

How do you inform the national stakeholders about the relevant legal instruments and standards available? (e.g. regulatory sandboxes)

16.1.2. Networking 2

Are there national initiatives that focus on collaboration about threat intelligence (AI threats, vulnerabilities and security controls) to the users/community?

16.1.3. Networking 5

Have you developed appropriate collaboration with the national AI stakeholders for information sharing?

17. METI (Japan)

17.1. Governance Guidelines for Implementation of AI Principles

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

17.1.1. Action Target 3-1-2

In case that a certain degree of gaps may potentially occur with AI systems of AI system operators that provide services to AI system users, they should, under the leadership of top management, provide sufficient information about the gaps and measures to address the gaps, as well as make a contact point easily accessible.

17.1.2. Action Target 3-3

Under the leadership of top management and with due consideration for trade secrets, companies that develop and operate AI systems and those that provide data should, except where everything from the preparation of data sets for training and other purposes to AI system development and operation is performed entirely within their own department, clarify and actively share, in accordance with the Principle of Fair Competition, AI system operational issues that the company or department is unable to fully address on their own and the information necessary to address these issues. In doing so, in order to facilitate the exchange of information clarified above, the AI system developer, AI system operator, and data provider are encouraged to agree on scope of information disclosure in advance and consider measures to protect trade secrets, for example, by entering a non-disclosure agreement.

18. Qatar Central Bank

18.1. 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

18.1.1. 14.2.2

The User must provide the results of its own use testing and its specific data sources or other inputs and Human Oversight plan.

18.1.2. 15.2

The Entity must make available to QCB the full range of data an AI Provider is contractually obligated to provide to the Entity.

18.1.3. 19.5

In the event of a serious incident, the Entity should report the matter to QCB immediately after the Provider has established a causal link between the AI System and the serious incident or the reasonable likelihood of such a link.

18.1.4. 20.1

An Entity must notify Customers when they are interacting with an AI System.

18.1.5. 20.2

An Entity must be transparent with Customers about their use of AI through their conduct and through accurate, understandable, and accessible plain language disclosure.

18.1.6. 20.4

An Entity must provide information to Customers how to use the AI System and ensure Customers always have access to the instructions.

18.1.7. 20.5

An Entity must provide clear explanations of the types of data, types of variables and factors that influence the decision-making process used by AI Systems upon Customers' request.

18.1.8. 20.6

An Entity must disclose the manner in which an AI decision may affect an individual Customer, and whether the decision is reversible.

18.1.9. 20.8

An Entity will provide a visible and accessible feedback channel for questions or comments.

18.1.10. 21.3

Any subsequent complaint by an aggrieved Customer must be handled through standard customer complaint processes.

19. UAE Ministry of Cabinet Affairs

19.1. 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.>

19.1.1. 5. Transparency

The UAE seeks to create a clear understanding of AI and how systems operate and make decisions, which helps build trust, enhance responsibility, and accountability in the use of these technologies.

20. Cyber Security Council (UAE)

20.1. 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

20.1.1. 3.6.2 Incident Reporting and Management for AI/ML - 2

This process should include clear guidelines on identifying and classifying incidents, timely reporting mechanisms, designated roles and responsibilities for incident response, and procedures for post-incident analysis and learning.

20.1.2. 3.6.2 Incident Reporting and Management for AI/ML - 3

The entity should ensure AI/ML security incidents are reported in a timely manner to all relevant stakeholders, including internal teams, third-party service providers, and regulatory bodies, in line with applicable laws and regulations, and contractual agreements.

21. IMDA

21.1. 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>

21.1.1. 2.2.1 End Users

Organisations may deploy agents to users within or outside their organisation. In doing so, organisations should ensure that users are provided sufficient information to hold the organisation accountable, as well as any information relating to the user's own responsibilities. More information can be found in Enabling end-user responsibility below.

21.1.2. 2.4 Enable end-user responsibility

Ultimately, end users are the ones who use and rely on agents, and human accountability also extends to these users. Organisations should provide sufficient information to end users to promote trust and enable responsible use.

Organisations should consider:

- **Transparency:** Users should be informed of the agents' capabilities (e.g. scope of agent's access to user's data, actions the agent can take) and the contact points whom users can escalate to if the agent malfunctions.
- **Education:** Users should be educated on proper use and oversight of agents (e.g. training should be provided on an agent's range of actions, common failure modes like hallucinations, usage policies for data), as well as the potential loss of trade craft i.e. as agents take over more functions, basic operational knowledge could be eroded. Hence sufficient training (especially in areas where agents are prevalent) should be provided to ensure that humans retain core skills.

21.1.3. 2.4.1 Different users, different needs

Organisations should cater to different users with different information needs, to enable such users to use AI responsibly. Broadly, there are two main archetypes of end-users – those who interact with agents, and those who integrate agents into their work processes or oversee them.

21.1.4. 2.4.2 Users who interact with agents

Such users usually interact with agents that act on behalf of the organisation, e.g. customer service or sales agents. These agents tend to be external facing, although they can also be deployed within the organisation e.g. a human resource agent that interacts with other users in the organisation.

For these users, focus on transparency. Organisations should share pertinent information to foster trust and facilitate proper usage of agents. Such information can include:

- **User's responsibilities:** Clearly define the user's responsibilities, such as asking the user to double-check all information provided by the agent.
- **Interaction:** Declare upfront that the users are interacting with agents.
- **Agents' range of actions and decisions:** Inform the users on the range of actions and decisions that the agent is authorised to perform and make.
- **Data:** Be clear on how user data is collected, stored, and used by the agents, in accordance with the organisation's data privacy policies. Where necessary, obtain explicit consent from users before collecting or using their data for the agents.
- **Human accountability and escalation:** Provide users with the respective human contact points who are responsible for the agents, whom the users can alert if the agents malfunction or if they are dissatisfied with a decision.

22. CISA

22.1. 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>

22.1.1. 1.3.3 Educate Personnel on AI - Leveraging explainable AI

Leveraging explainable AI by having operators request that AI outputs include clear and transparent documentation of decision-making processes; this enables humans to better understand and validate outputs.

22.1.2. 4.1.6 - Explore new AI explainability and transparency tools.

Explainable AI (XAI) and transparent AI are evolving fields of research that seek to make AI systems more understandable. Explainability focuses on making the reasoning behind individual AI decisions understandable to users, while transparency emphasizes making the overall AI system's development and operation open and accessible. Essentially, explainability clarifies why an AI made a specific decision, while transparency focuses on how the AI system works as a whole. Critical infrastructure owners and operators should, where possible, explore interpretable models or tools that make AI decisions more understandable to humans.

23. Cloud Security Alliance (CSA)

23.1. AI Controls Matrix

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

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

23.1.1. BCR-07

Establish and maintain communication channels with all relevant stakeholders in the course of business continuity and resilience procedures.

23.1.2. DSP-18

The providers should implement and describe to customers the procedure to manage and respond to requests for disclosure of Personal Data by Law Enforcement Authorities according to applicable laws and regulations.

23.1.3. GRC-14

Evaluate, document, and communicate the degree of explainability of the AI Services, including possible limitations and exceptions.

23.1.4. SEF-03

Establish, document, approve, communicate, apply, evaluate and maintain a security incident response plan, which includes but is not limited to: a communication strategy for notifying relevant internal departments, impacted AICs, and other business critical relationships (such as supply-chain) that may be impacted.

23.1.5. SEF-07

Define and implement, processes, procedures and technical measures for security breach notifications. Report material security breaches and assumed security breaches including any relevant supply chain breaches, as per applicable SLAs, laws and regulations.

23.1.6. STA-04

Provide SSRM Guidance to the Customer detailing information about the SSRM applicability throughout the supply chain.

24. CoSAI

24.1. 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>

24.1.1. 3.3.2. Detection and Analysis Phase - Initial Triage - Priority Assignment

- Assign priority based on impact
- Implement notification procedures
- Mobilize specialized resources

24.1.2. 3.3.3. Containment, Eradication, and Recovery Phase - Recovery Procedures - User Communication

- Communication templates
- Transparency guidelines
- Communication protocols
- User feedback mechanisms

24.2. 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>

24.2.1. 3.2.8 Secure Tool and UX Design

Tool and UX design represent a critical security control point in Model Context Protocol (MCP) deployments. While much attention is paid to model safety and prompt injection defenses, the tools that agents invoke are often the actual execution surface where security boundaries are crossed and sensitive operations are performed. Poor tool design can undermine even the most robust authentication and authorization controls by creating overly permissive capabilities or delegating security-critical decisions to the LLM itself.

Each tool should have a single, clearly defined purpose with explicit boundaries on what it can and cannot do. When possible, create use-case driven or purpose-built tools, avoiding excessively powerful tools, e.g., execute a prepared statement versus executing any SQL statement. Tool implementations should not rely on the LLM to perform security-critical operations, validate inputs, or enforce constraints.

Safe and secure execution should not rely solely on the human user, who may not understand the security implications of frequent security prompts and can easily become fatigued. Security-relevant messages and elicitations should be clear, indicating the implications of the request, and unambiguous what is being requested.

25. Personal Data Protection Commission Singapore (PDPC)

25.1. Model Artificial Intelligence Governance Framework Second Edition

Link: <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/smodelaigovframework2.pdf>

Link: <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/smodelaigovframework2.pdf>

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

Reviewing communications channels and interactions with stakeholders to provide disclosure and effective feedback channels.

25.1.1.2. Reproducibility - c)

Making available replication files (i.e. files that replicate each step of the AI model's developmental process) to facilitate the process of testing and reproducing behaviours;

25.1.1.3. Interacting with consumers - a)

Making sure that consumers are aware that the products or services that they are considering are AI-enabled. Such information could be provided as part of a general product description.

25.1.1.4. Interacting with consumers - b)

Providing information so that consumers know how the AI-enabled features are expected to behave during normal use. The information could be provided in more detailed descriptions or specifications of product features. This, however, may not be necessary for every feature that is AI-enabled. Organisations are encouraged to identify those features where providing additional information in this manner will enhance consumer trust. Similarly, if AI is used in decision-making, information may be provided so that consumers understand how decisions made with the assistance of AI may affect them. This can likewise be provided through descriptions of how the service will be provided.

25.1.1.5. Interacting with consumers - c)

For AI-enabled features that consumers interact with regularly, providing information so that they understand why the AI-enabled feature is behaving in a certain way, and providing preference settings to allow consumers some influence over future behaviour where possible. As doing so requires more engineering effort (such as in the providing additional user interfaces to user history), and the level of information provided may be somewhat more detailed and personalised than feature descriptions, organisations will have to decide which of their product features will benefit from provision of this level of detail.

25.1.1.6. Interacting with consumers - d)

For AI-augmented decisions that affect consumers, consider providing additional information so that they understand why the decisions were made; and for certain categories of such decisions, providing an appropriate channel to contest such decisions. The level of information that is provided will necessarily be detailed but this may not be necessary except for those scenarios where a customer is affected by the decision.

26. Multi Agency

26.1. Guidelines for secure AI system development

Link: <https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf>

Link: <https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf>

26.1.1. Make it easy for users to do the right things

You recognise that each new setting or configuration option is to be assessed in conjunction with the business benefit it derives, and any security risks it introduces. Ideally, the most secure setting will be integrated into the system as the only option. When configuration is necessary, the default option should be broadly secure against common threats (that is, secure by default). You apply controls to prevent the use or deployment of your system in malicious ways.

You provide users with guidance on the appropriate use of your model or system, which includes highlighting limitations and potential failure modes. You state clearly to users which aspects of security they are responsible for, and are transparent about where (and how) their data might be used, accessed or stored (for example, if it is used for model retraining, or reviewed by employees or partners).

27. NIST

27.1. 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>

27.1.1. MAP 5.2

Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.

27.1.2. MEASURE 3.3

Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.

27.1.3. MEASURE 4.1

Measurement approaches for identifying AI risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.

27.1.4. MANAGE 4.3

Incidents and errors are communicated to relevant AI actors, including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.

27.2. 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>

27.2.1. PO.1.3

Communicate requirements to all third parties who will provide commercial software components to the organization for use by the organization's own software.

Include AI model development security in the requirements being communicated for third-party software components.

27.3. AI 800-1

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf>

Link: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf>

27.3.1. Practice 6.1: Monitor for evidence of misuse - 4

Request that distribution channels monitor for misuse and share information regarding this monitoring.

27.3.2. Practice 6.3: Establish misuse reporting mechanisms - 1

Adopt policies that protect and reward individuals who report model issues related to misuse risk.

27.3.3. Practice 6.3: Establish misuse reporting mechanisms - 3

Communicate the identified issues or misuse instances clearly with employees and contractors, as appropriate.

27.3.4. Practice 7.1: Publish transparency reports - 4

Share steps that downstream developers and deployers of AI systems that integrate the foundation model should take to manage misuse risk.

27.3.5. Practice 7.1: Publish transparency reports - 5

Share relevant details about the internal organizational processes used to create risk assessments and to make deployment and development decisions.

27.3.6. Practice 7.1: Publish transparency reports - 6

Make the transparency reports publicly available, update them on a regular basis (e.g., with each new major version of the model), and include key information related to misuse risk.

27.3.7. Practice 7.2: Disclose information about risk management practices - 1

Share information covering the practices used to achieve the objectives listed in this document, including at least as much detail as described in the documentation sections for each practice.

27.3.8. Practice 7.2: Disclose information about risk management practices - 2

Share information about threat profiles with other entities across the AI supply chain and the public to develop a joint knowledge base and save resources.

27.3.9. Practice 7.3: Report misuse incidents - 3

Share verified reports of misuse with relevant third parties, such as AI incident databases and other model developers.

27.4. 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>

27.4.1. ID.RA-08

Processes for receiving, analyzing, and responding to vulnerability disclosures are established

27.4.2. DE.AE-08

Incidents are declared when adverse events meet the defined incident criteria

27.4.3. RS.CO-02

Internal and external stakeholders are notified of incidents

27.4.4. RS.CO-03

Information is shared with designated internal and external stakeholders

27.4.5. RC.CO-03

Recovery activities and progress in restoring operational capabilities are communicated to designated internal and external stakeholders

27.4.6. RC.CO-04

Public updates on incident recovery are shared using approved methods and messaging