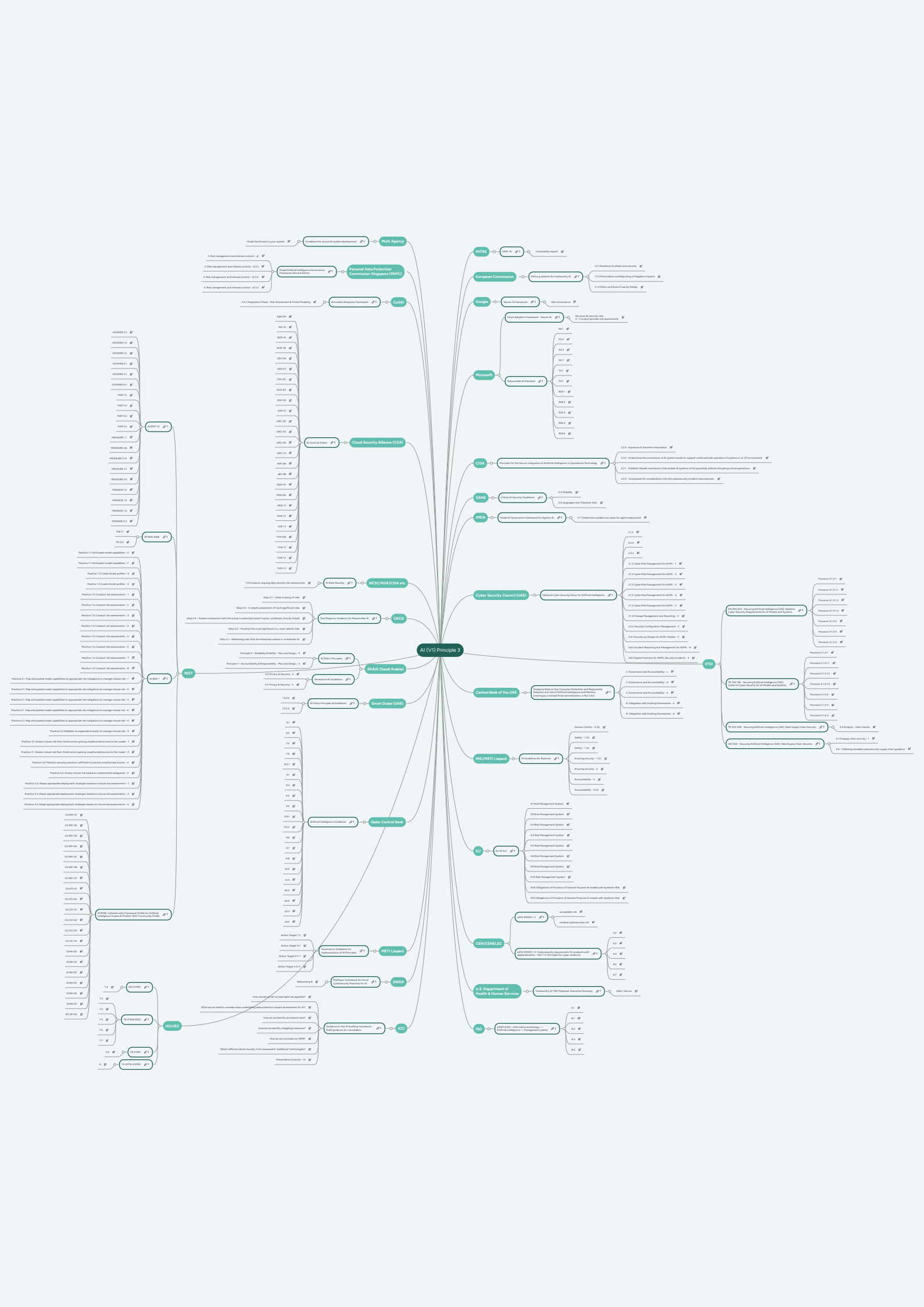




AI (V1) Principle 3

1. MITRE

1.1. SAFE-AI

Link: https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

Link: https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

1.1.1. Vulnerability exploit

Code development and testing practices for AI-enabled systems do not always conform to traditional software development practices. Consequently, assessing AI system vulnerabilities may raise unexpected challenges (e.g., in some cases it may be difficult to even identify what to test). These challenges are of course a prime opportunity for attackers to exploit gaps in the vulnerability assessment and initiate attacks. Avoiding these undesirable outcomes requires stringent approaches to vulnerability assessment and monitoring. All known potential threats, vulnerabilities, and attack vectors associated with an AI-enabled system must be identified early during the design phase (e.g., by using ATLAS) and the risks must be managed. It is critical to define metrics and procedures for detecting, tracking, and measuring known risks, errors, incidents, or negative impacts. Metrics should also account for known AI design and implementation failure modes associated with properties like brittleness. The deployed AI-enabled system must be continuously tested for errors or vulnerabilities

2. European Commission

2.1. 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>

2.1.1. 1.2.1 Resilience to attack and security

Systems and data can also become corrupted by malicious intention or by exposure to unexpected situations. Insufficient security processes can also result in erroneous decisions or even physical harm. For AI systems to be considered secure, possible unintended applications of the AI system (e.g. dual-use applications) and potential abuse of the system by malicious actors should be taken into account, and steps should be taken to prevent and mitigate these.

2.1.2. 1.7.2 Minimisation and Reporting of Negative Impacts

Both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system. The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimise negative impact. These assessments must be proportionate to the risk that the AI systems pose.

2.1.3. 2.1.2 Ethics and Rule of Law by Design

Methods to ensure values-by-design provide precise and explicit links between the abstract principles which the system is required to respect and the specific implementation decisions. The idea that compliance with norms can be implemented into the design of the AI system is key to this method. Companies are responsible for identifying the impact of their AI systems from the very start, as well as the norms their AI system ought to comply with to avert negative impacts. Different "by-design" concepts are already widely used, e.g. privacy-by-design and security-by-design. As indicated above, to earn trust AI needs to be secure in its processes, data and outcomes, and should be designed to be robust to adversarial data and attacks. It should implement a mechanism for fail-safe shutdown and enable resumed operation after a forced shut-down (such as an attack).

3. ETSI

3.1. 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

3.1.1. Provision 5.1.3-1

Developers and System Operators shall analyse threats and manage security risks to their systems. Threat modelling should include regular reviews and updates and address AI-specific attacks, such as data poisoning, model inversion, and membership inference.

3.1.2. Provision 5.1.3-1.1

The threat modelling and risk management process shall be conducted to address any security risks that arise when a new setting or configuration option is implemented or updated at any stage of the AI lifecycle.

3.1.3. Provision 5.1.3-1.2

Developers shall manage the security risks associated with AI models that provide superfluous functionalities, where increased functionality leads to increased risk. For example, where a multi-modal model is being used but only single modality is used for system function.

3.1.4. Provision 5.1.3-1.3

System Operators shall apply controls to risks identified through the analysis based on a range of considerations, including the cost of implementation in line with their corporate risk tolerance.

3.1.5. Provision 5.1.3-2

Where AI security threats are identified that cannot be resolved by Developers, this shall be communicated to System Operators so they can threat model their systems. System Operators shall communicate this information to End-users, so they are made aware of these threats. This communication should include detailed descriptions of the risks, potential impacts, and recommended actions to address or monitor these threats.

3.1.6. Provision 5.1.3-3

Where an external entity has responsibility for AI security risks identified within an organizations infrastructure, System Operators should attain assurance that these parties are able to address such risks

3.1.7. Provision 5.1.3-4

Developers and System Operators should continuously monitor and review their system infrastructure according to risk appetite. It is important to recognize that a higher level of risk will remain in AI systems despite the application of controls to mitigate against them.

3.2. 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

3.2.1. Provision 5.1.3-1

"Developers and System Operators shall analyse threats and manage security risks to their systems. Threat modelling should include regular reviews and updates and address AI-specific attacks, such as data poisoning, model inversion, and membership inference." (ETSI TS 104 223 [i.1])

Related threats/risks:

AI systems face unique threats, such as data poisoning, model inversion, and membership inference attacks, which traditional threat models cannot account for. New threats will emerge that will need to be incorporated in threat modelling and risk management.

Example Measures/Controls:

Perform Threat Modelling including AI threats: Apply threat modelling that captures potential impacts on stakeholders including both AI and traditional cyberattacks. Document each identified threat in detail, outlining the likelihood and severity of potential impacts to the AI model and the broader system and list mitigations using standardized OWASP or MITRE controls. Both OWASP AI Exchange [i.10] or MITRE ATLAS [i.8] provide threat taxonomies and related mitigations which both types of attacks and they can be used in threat modelling. If the AI system processes or was built on personal data ICO's guidance on AI and security is useful to consult for regulatory compliance.

3.2.2. Provision 5.1.3-1.1

"The threat modelling and risk management process shall be conducted to address any security risks that arise when a new setting or configuration option is implemented or updated at any stage of the AI lifecycle." (ETSI TS 104 223 [i.1])

Related threats/risks:

Failure to conduct threat modelling and risk management when implementing or updating settings or configurations during the AI lifecycle can lead to unmitigated security vulnerabilities, such as configuration errors or unanticipated attack vectors, increasing the risk of exploitation and system compromise.

Example Measures/Controls:

Conduct Threat Modelling for Configuration Changes: Perform threat modelling whenever settings or configurations are implemented or updated to identify and mitigate security risks throughout the AI lifecycle.

3.2.3. Provision 5.1.3-1.2

"Developers shall manage the security risks associated with AI models that provide superfluous functionalities, where increased functionality leads to increased risk. For example, where a multi-modal model is being used but only single modality is used for system function." (ETSI TS 104 223 [i.1])

Related threats/risks:

Allowing AI models to retain superfluous functionalities that are not required for the system's purpose can introduce unnecessary security risks, such as expanded attack surfaces, increased vulnerability to exploitation, and potential misuse of unused features, compromising the overall security of the system.

Example Measures/Controls 1:

Restrict Superfluous Functionalities: Limit AI model functionalities to those essential for the system's purpose to reduce the attack surface and minimize security risks associated with unused features.

Example Measures/Controls 2:

Integrate Threat Modelling with AI Governance: Require completed threat models for governance approval at critical stages of the AI lifecycle, ensuring documented risk understanding and mitigation before deployment providing support and guidance and ensuring cross-discipline input (ethics, privacy, legal, etc.) to threat modelling.

3.2.4. Provision 5.1.3-1.3

"System Operators shall apply controls to risks identified through the analysis based on a range of considerations, including the cost of implementation in line with their corporate risk tolerance." (ETSI TS 104 223 [i.1])

Related threats/risks:

When risk tolerance is not clearly defined in the context of AI-specific risks such as data poisoning or model misuse, this could result into inadequately prioritized controls leading to breaches, operational disruptions, or unethical decision-making.

Example Measures/Controls:

Develop a Prioritization Framework for AI Risk Controls: Use an AI-specific risk-scoring system to prioritize mitigations and controls based on the impact of threats, likelihood of occurrence, and in alignment with organizational risk tolerance. This should account for regulatory risk, including data protection, and cover AI-specific vulnerabilities, such as adversarial manipulation, model drift, and bias.

3.2.5. Provision 5.1.3-2

"Where AI security threats are identified that cannot be resolved by Developers, this shall be communicated to System Operators so they can threat model their systems. System Operators shall communicate this information to End-users, so they are made aware of these threats. This communication should include detailed descriptions of the risks, potential impacts, and recommended actions to address or monitor these threats." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without clear communication on unresolved risks, System Operators and End-users can lack awareness, limiting their ability to apply safeguards effectively.

Example Measures/Controls:

Document and Communicate Identified Unresolved Risks: Ensure clear documentation and timely communication of any unresolved threats to all relevant stakeholders.

3.2.6. Provision 5.1.3-3

"Where an external entity has responsibility for AI security risks identified within an organizations infrastructure, System Operators should attain assurance that these parties are able to address such risks." (ETSI TS 104 223 [i.1])

Related threats/risks:

Reliance on third parties without adequate verification could expose the AI system to unmanaged vulnerabilities.

Example Measures/Controls:

Conduct AI-Specific Security Assessments for Third Parties: Ensure third-party components and vendors undergo security assessments that specifically address AI-related risks and adherence with ETSI TS 104 223 [i.1].

3.2.7. Provision 5.1.3-4

"Developers and System Operators should continuously monitor and review their system infrastructure according to risk appetite. It is important to recognize that a higher level of risk will remain in AI systems despite the application of controls to mitigate against them." (ETSI TS 104 223 [i.1])

Related threats/risks:

Residual risk can be exploited by malicious actors, especially as evolving threats introduce new vulnerabilities or amplify existing ones, leading to potential breaches, disruptions, or compromised AI integrity.

Example Measures/Controls:

Establish Continuous AI Risk Monitoring Controls: Implement a regular review processes of AI developments to determine whether emerging vulnerabilities, improved mitigation techniques, or advancements in AI models necessitates updates to the risk assessment controls.

3.3. TR 104 048 - Securing Artificial Intelligence (SAI); Data Supply Chain Security

Link: https://www.etsi.org/deliver/etsi_tr/104000_104099/104048/01.01.01_60/tr_104048v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_tr/104000_104099/104048/01.01.01_60/tr_104048v010101p.pdf

3.3.1. 6.5 Analysis - Hash checks

Existing cryptographic mechanisms can be used for protecting the integrity of data in an efficient way. For verification of data integrity there is a trade-off between efficiency and security, which should be balanced according to the risk level of the application.

3.4. SAI 002 - Securing Artificial Intelligence (SAI); Data Supply Chain Security

Link: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/002/01.01.01_60/gr_SAI002v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/002/01.01.01_60/gr_SAI002v010101p.pdf

3.4.1. 6.1.3 Supply chain security - 1

Understanding the risks associated with the supply chain, particularly for high-value components such as datasets. This includes understanding the security posture of the suppliers.

3.4.2. 6.5 - Following standard cybersecurity supply chain guidance

Data, models and the roles and risks associated with them can be understood and assessed in the same way as any other component of a system.

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. Risk Governance

Inventory, measure, and monitor residual risk to AI in your organization.

5. Microsoft

5.1. Cloud Adoption Framework - Secure AI

Link: <https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/ai/secure>

Link: <https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/ai/secure>

5.1.1. Discover AI security risks 4 - Conduct periodic risk assessments

New threats emerge as AI models, usage patterns, and threat actors evolve over time. Regular assessments ensure your security posture adapts to changing risk landscapes. Run recurring assessments to identify vulnerabilities in models, data pipelines, and deployment environments, and use assessment findings to guide your risk mitigation priorities.

5.2. 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>

5.2.1. A2.1

Review defined Restricted Uses to determine whether the system meets the definition of any Restricted Use. If it does, document this in the Impact Assessment, and follow the requirements for the Restricted Use.

5.2.2. A2.2

Answer prompts in the Impact Assessment template to determine whether the system meets the definition of a Sensitive Use. If it does, report it to the Office of Responsible AI, and follow any additional requirements resulting from a Sensitive Uses review.

5.2.3. A2.3

Review your systems at least annually against the definitions for Sensitive Uses and Restricted Uses. If there are systems that meet the criteria for Sensitive Uses, report them to the Office of Responsible AI. If there are systems that meet the criteria for Restricted Uses, notify the Office of Responsible AI.

5.2.4. A5.7

If there are Responsible Release Criteria for metrics or rubrics that have not been met, consult with the reviewers named in the Impact Assessment, and in the case of Sensitive Uses, with the Office of Responsible AI, to develop a plan detailing how the gap will be managed until it can be closed. Document that plan.

5.2.5. T3.1

Identify stakeholders who will use or be exposed to the system, in accordance with the Impact Assessment requirements. Document these stakeholders using the Impact Assessment template.

5.2.6. F3.7

Publish information for customers about these risks involving identified demographic groups. When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

5.2.7. RS2.1

Define predictable failures, including false positive and false negative results for the system as a whole and how they would impact stakeholders for each intended use. Use the Impact Assessment template to document any adverse impacts of these failures on stakeholders.

5.2.8. RS2.2

For each case of a predictable failure likely to have an adverse impact on a stakeholder, document the failure management approach:

- 1) When possible, design and build the system to avoid this failure. Describe the design solution. Estimate the time range for resolving predictable failures for each designed solution or indicate that the failure will be prevented by design.
- 2) When a failure cannot be prevented by design, build a fallback option that may be used when this failure occurs. Describe the fallback option and document the estimated time required to invoke and use the fallback option.
- 3) Provide training and documentation for stakeholders accountable for system oversight that supports their resolution of the failure. Describe the documentation and training.

5.2.9. RS3.3

When new uses, critical operational factors, or changes in the supported range of an operational factor are identified, determine whether any new use or operational factor can be supported with the existing system, will be supported but require additional work, or will not be supported.

- When new uses or operational factors identified are to be supported, evaluate the updated system in accordance with requirement RS1.6, add the new intended use to the Impact Assessment, and publish updated communication in accordance with requirement RS1.9.
- When these new uses or operational factor range changes cannot or will not be accommodated to ensure reliable and safe performance of the system update customer documentation described in RS1.9 to include the new use as an unsupported use.

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

5.2.10. RS3.4

When a system is to be used for a Sensitive Use that imposes qualification or quality control requirements beyond the intended uses and/or operational factor ranges, conduct an evaluation specific to this use. If the required Responsible Release Criteria cannot be met, the Office of Responsible AI will review the results and decide how to proceed. Document any changes to the Responsible Release Criteria and document the results of evaluation.

5.2.11. RS3.6

If there are targets in Ongoing Evaluation Checkpoints that are no longer satisfied, consult with named reviewers, and in the case of Sensitive Uses, with the Office of Responsible AI, to develop and implement a plan to close any gaps. Document the process, its results, and conclusions.

6. CISA

6.1. 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>

6.1.1. 2.2.3 - Exposure of Sensitive Information

Minimize risk by not sharing sensitive data with AI models, especially if the AI models are hosted in an environment hosted or controlled by external parties, such as public cloud infrastructure.

6.1.2. 4.1.2 - Understand the correctness of AI system results to support continued safe operation of systems in an OT environment.

It is vital for critical infrastructure owners and operators to understand the states where an AI system can fail to produce accurate and reliable results. This understanding includes expectations for false positives and false negatives in the system's performance, and how the false positives compare to the base rate of true positives.

6.1.3. 4.2.1 - Establish failsafe mechanisms that enable AI systems to fail gracefully without disrupting critical operations.

Incorporate new AI system failure states, including how to bypass or replace an AI system, into existing functional safety and incident response processes. Integrating an AI system into existing OT networks inevitably generates new failure states for the overall critical infrastructure system. Therefore, operators responsible for revising the existing functional safety and incident response processes should incorporate these new failure states as they are critical to ensuring safe operation of these systems.

6.1.4. 4.2.3 - Incorporate AI considerations into the cybersecurity incident response plan.

Despite organizations' best efforts at mitigation, risk cannot be reduced to zero; incidents are inevitable. To account for this, critical infrastructure owners and operators should update their incident response plans and functional safety procedures with steps for responding to malicious activity directed against an AI system and AI system failure.

7. SANS

7.1. Critical AI Security Guidelines

Link: <https://sansorg.egnyte.com/dl/bvkYQxrW8QMj>

Link: <https://sansorg.egnyte.com/dl/bvkYQxrW8QMj>

7.1.1. 4.5 Modality

Although powerful, multimodal implementations can increase the attack surface. Common sense and research suggest safety and alignment can prove inconsistent across different modalities. As an example, a text-only prompt that might have been considered unsafe could be allowed if the text were instead submitted as an image.

7.1.2. 4.6 Languages and Character Sets

It has been shown that multilingual and multicharacter models can introduce new vulnerabilities and expand the attack surface. Multilingual jailbreak challenges have been observed when utilizing prompts in a language other than that used in the primary training data.¹² Research has shown this can result in jailbreaking or providing instructions to deliberately attack vulnerable LLMs. The same is true for character sets, which have been shown to increase hallucinations and comprehension errors.¹³ Additional research highlights that when instructions involve Unicode characters outside the standard Latin or variants of other languages, a reduction in guardrail efficiency is observed.

8. IMDA

8.1. 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>

8.1.1. 2.1.1 Determine suitable use cases for agent deployment

Risk identification and assessment is the first step when considering if an agentic use case is suitable for development or deployment. Risk is a function of likelihood (probability of the risk manifesting) and impact (severity of impact if the risk manifests).

9. Cyber Security Council (UAE)

9.1. 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

9.1.1. 2.1.2

The entity shall identify, assess, and mitigate unique cyber security risks associated with AI/ML systems in a systematic and ongoing manner.

9.1.2. 2.2.4

The entity shall implement a systematic process to identify, assess, and remediate vulnerabilities in AI/ML systems to enhance their resilience against cyber security threats.

9.1.3. 2.3.3

The entity shall ensure the security of the AI/ML inference process by protecting against potential threats and maintaining the integrity and confidentiality of inference data.

9.1.4. 3.1.2 Cyber Risk Management for AI/ML - 1

The entity should have a formal cyber security risk management process that specifically addresses the unique cyber security risks associated with AI/ML.

9.1.5. 3.1.2 Cyber Risk Management for AI/ML - 2

The AI/ML cyber security risk management process should be integrated into the entity's broader risk management program, ensuring that cyber security risks associated with AI/ML systems are classified, prioritized, and considered as part of the organization's overall risk profile.

9.1.6. 3.1.2 Cyber Risk Management for AI/ML - 3

The entity should conduct cyber security risk assessments at regular intervals and at key stages in the AI/ML lifecycle, such as during model design, data collection, training, and deployment.

9.1.7. 3.1.2 Cyber Risk Management for AI/ML - 4

The entity should assess and mitigate cyber security risks related to cloud-based AI solutions, particularly those that utilize client data under the End User License Agreement (EULA) for unsupervised learning in the background.

9.1.8. 3.1.2 Cyber Risk Management for AI/ML - 5

The entity should regularly review and update the AI/ML cyber security risk management process to address new threats, vulnerabilities, and risk mitigation strategies.

9.1.9. 3.1.4 Change Management and Reporting - 2

This process should consider the potential cyber security impacts of changes, including changes that could affect the confidentiality, integrity, or availability of AI/ML systems or the data they process.

9.1.10. 3.2.2 Security Configuration Management - 2

The entity should periodically review and update AI/ML secure configuration baselines to ensure continued effectiveness against evolving threats and changes in system architecture or technology.

9.1.11. 3.3.1 Security by Design for AI/ML Models - 2

The entity should perform threat modeling during the design and development of AI/ML models to identify potential cyber security threats and design appropriate mitigating controls.

9.1.12. 3.6.2 Incident Reporting and Management for AI/ML - 6

The entity should conduct periodic reviews and updates to the automated response rules based on lessons learned from exercises and actual incidents to ensure continuous improvement and alignment with emerging threats.

9.1.13. 3.6.3 Digital Forensics for AI/ML Security Incidents - 3

The entity should incorporate the forensic insights into the post-incident review process, cyber security strategy and risk management, for enhancing the overall resilience of its AI/ML systems.

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. 2. Governance and Accountability - c

Regular reporting should be required by and provided to Senior Management and Boards of LFIs, covering performance and risk.

10.1.2. 2. Governance and Accountability - d

Governance structures should facilitate informed decision-making, enable the identification and mitigation of risks and ensure that AI and ML systems and applications are aligned with the institution's risk appetite and legal obligations. AI-related risks should be incorporated into the institution's governance framework in a cohesive and consolidated manner, including with specific adaptable roles and responsibilities for the Audit and Risk Committee, Risk Management, Internal Audit, and IT.

10.1.3. 2. Governance and Accountability - e

Boards and senior management should ensure that risk committees and control functions (e.g., compliance, internal audit and risk management) understand AI-driven processes and can challenge outcomes where appropriate, with the adoption, deployment and use of AI being an integral part of the Risk Management framework.

10.1.4. 8. Integration with Existing Frameworks - a

Whilst the focus of this guidance is Consumer Protection, risk management with respect to AI and ML activities should be embedded within the institution's enterprise-wide risk management framework, including conduct risk, credit risk, operational risk and cybersecurity risk. AI risk assessments should not operate in isolation but should inform and be informed by the institution's overall risk appetite and controls.

10.1.5. 8. Integration with Existing Frameworks - d

LFIs should create processes to rate the risk of each AI system/application/technology they deploy or use, to enable appropriate risk assessment, monitoring and management of the AI whilst deployed in use and developing, which may be influenced by data quality and sensitivity, capability of the AI, controls in place, impact of the AI and dependence on the AI and/or third parties in the use of the AI.

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 - 2 (b)

When developing, providing, or using an AI system or service, pay attention and take necessary countermeasures against the risk of heavy dependence on AI, such as automated biases.

11.1.2. Safety - 1 (d)

Conduct appropriate risk analyses to take countermeasures against risks (avoidance, mitigation, transference, or acceptance).

11.1.3. Safety - 1 (e)

If there are potential hazards to the lives, bodies, properties, and minds of humans and the environment, organize measures to be taken in advance and offer related information to stakeholders. Clearly specify measures that should be taken by relevant stakeholders and the terms of use.

11.1.4. Ensuring security - 1 (c)

Bearing in mind that relevant stakeholders might make unexpected judgments by mixing detailed information into inference target data, recognize that vulnerabilities cannot be completely eliminated from AI systems and services.

11.1.5. Ensuring security - 2

New methods for attacking AI systems and services from the outside are increasing on a daily basis. In order to address those risks, check the matters to be noted.

11.1.6. Accountability - 2

Provide and explain information on how AI business actors conform to common guiding principles regularly to stakeholders, including suppliers, according to their knowledge and competence. This information summarizes, for example, the following items:

General:

- Whether any risk is found that prevents the common guiding principles from being implemented, to what extent it prevents the implementation of those guiding principles
- Implementation progress of the common guiding principles

Human-centric:

- How disinformation is considered, and how diversity, inclusion, user support, and sustainability are ensured

Safety:

- Known risks relating to AI systems and services, countermeasures against them, and how to ensure safety against them

Fairness:

- Possibility that technological elements forming AI models will introduce bias. Those elements may include training data, AI model training process, prompts expected to be entered by AI business users or non-business users, and reference information and collaborating external services used by AI models for inference.

Privacy protection:

- Risks of infringements of privacy of AI business actors or stakeholders entailed by the AI system or service, countermeasures against those risks, and actions expected to be taken when the privacy breach actually occurred.

Ensuring security:

- Conformity to standards required to facilitate collaboration between AI systems and services or with other systems if such collaboration occurs
- Any risks that may occur when the AI system or service collaborates with other AI systems and services via the Internet, and measures to be taken against the risks

11.1.7. Accountability - 5 (a)

As necessary, establish and publicly report policies, including those created by each AI business actor on AI governance or privacy in relation to risk management or safety assurance associated with the use of AI systems and services. Those policies involve social responsibilities, including sharing visions with and giving out and providing information to society and general citizens.

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

12.1.1.1. 9.1 Risk Management System

A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems.

12.1.1.2. 9.2 Risk Management System

The risk management system shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating. It shall comprise the following steps:

- (a) the identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose;
- (b) the estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse;
- (c) the evaluation of other risks possibly arising, based on the analysis of data gathered from the post-market monitoring system referred to in Article 72;
- (d) the adoption of appropriate and targeted risk management measures designed to address the risks identified pursuant to point (a).

12.1.1.3. 9.3 Risk Management System

The risks referred to in this Article shall concern only those which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information.

12.1.1.4. 9.4 Risk Management System

The risk management measures referred to in paragraph 2, point (d), shall give due consideration to the effects and possible interaction resulting from the combined application of the requirements set out in this Section, with a view to minimising risks more effectively while achieving an appropriate balance in implementing the measures to fulfil those requirements.

12.1.1.5. 9.5 Risk Management System

The risk management measures referred to in paragraph 2, point (d), shall be such that the relevant residual risk associated with each hazard, as well as the overall residual risk of the high-risk AI systems is judged to be acceptable.

In identifying the most appropriate risk management measures, the following shall be ensured:

- (a) elimination or reduction of risks identified and evaluated pursuant to paragraph 2 in as far as technically feasible through adequate design and development of the high-risk AI system;
- (b) where appropriate, implementation of adequate mitigation and control measures addressing risks that cannot be eliminated;
- (c) provision of information required pursuant to Article 13 and, where appropriate, training to deployers.

With a view to eliminating or reducing risks related to the use of the high-risk AI system, due consideration shall be given to the technical knowledge, experience, education, the training to be expected by the deployer, and the presumable context in which the system is intended to be used.

12.1.1.6. 9.6 Risk Management System

High-risk AI systems shall be tested for the purpose of identifying the most appropriate and targeted risk management measures. Testing shall ensure that high-risk AI systems perform consistently for their intended purpose and that they are in compliance with the requirements set out in this Section.

12.1.1.7. 9.9 Risk Management System

When implementing the risk management system as provided for in paragraphs 1 to 7, providers shall give consideration to whether in view of its intended purpose the high-risk AI system is likely to have an adverse impact on persons under the age of 18 and, as appropriate, other vulnerable groups.

12.1.1.8. 9.10 Risk Management System

For providers of high-risk AI systems that are subject to requirements regarding internal risk management processes under other relevant provisions of Union law, the aspects provided in paragraphs 1 to 9 may be part of, or combined with, the risk management procedures established pursuant to that law.

12.1.1.9. 55.8 Obligations of Providers of General-Purpose AI models with Systemic Risk

In addition to the obligations listed in Articles 53 and 54, providers of general-purpose AI models with systemic risk shall:

- (a) perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks;
- (b) assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk;
- (c) keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them;
- (d) ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model.

12.1.10. 55.9 Obligations of Providers of General-Purpose AI models with Systemic Risk

Providers of general-purpose AI models with systemic risk may rely on codes of practice within the meaning of Article 56 to demonstrate compliance with the obligations set out in paragraph 1 of this Article, until a harmonised standard is published. Compliance with European harmonised standards grants providers the presumption of conformity to the extent that those standards cover those obligations. Providers of general-purpose AI models with systemic risks who do not adhere to an approved code of practice or do not comply with a European harmonised standard shall demonstrate alternative adequate means of compliance for assessment by the Commission.

13. CEN/CENELEC

13.1. prEN 40000-1-1

Link: <https://genorma.com/en/standards/pren-40000-1-1>

Link: <https://genorma.com/en/standards/pren-40000-1-1>

13.1.1. acceptable risk

the level of risk deemed acceptable for intended and reasonably foreseeable use, considering state of the art and safety.

13.1.2. residual cybersecurity risk

risk remaining after treatment.

13.2. prEN 40000-1-2: Cybersecurity requirements for products with digital elements - Part 1-2: Principles for cyber resilience

Link: <https://genorma.com/en/standards/pren-40000-1-2>

Link: <https://genorma.com/en/standards/pren-40000-1-2>

13.2.1. 5.2

Risk Based Approach to Cybersecurity

13.2.2. 6.3

Risk acceptance criteria and risk management methodology

13.2.3. 6.4

Risk assessment

13.2.4. 6.6

Risk Communication

13.2.5. 6.7

Risk monitoring

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

14.1. 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>

14.1.1. Safe / Secure

AI systems should be protected from risks (including Cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity

15. ISO

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

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

15.1.1. 6.1

Actions to address risks and opportunities

15.1.2. 8.1

Operational planning and control

15.1.3. 8.2

AI risk assessment

15.1.4. 8.3

AI risk treatment

15.1.5. 8.4

AI system impact assessment

16. ICO

16.1. 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>

16.1.1. How should we set a meaningful risk appetite?

To manage the risks to individuals that arise from the processing of personal data in your AI systems, it is important that you develop a mature understanding and articulation of fundamental rights, risks, and how to balance these and other interests. Ultimately, it is necessary for you to:

- assess the risks to individuals' rights that your use of AI poses;
- determine how you need to address these; and
- establish the impact this has on your use of AI.

You should ensure your approach fits both your organisation and the circumstances of your processing. Where appropriate, you should also use risk assessment frameworks.

16.1.2. What do we need to consider when undertaking data protection impact assessments for AI?

DPIAs are a key part of data protection law's focus on accountability and data protection by design.

You should not see DPIAs as a mere box ticking compliance exercise. They can effectively act as roadmaps for you to identify and control the risks to rights and freedoms that use of AI can pose. They are also a perfect opportunity for you to consider and demonstrate your accountability for the decisions you make in the design or procurement of AI systems.

16.1.3. How do we identify and assess risks?

The DPIA process will help you to objectively identify the relevant risks. You should assign a score or level to each risk, measured against the likelihood and the severity of the impact on individuals.

The use of personal data in the development and deployment of AI systems may not just pose risks to individuals' information rights. When considering sources of risk, a DPIA should consider the potential impact of material and non-material damage or harm on individuals.

For instance, machine learning systems may reproduce discrimination from

historic patterns in data, which could fall foul of equalities legislation. Similarly, AI systems that stop content being published based on the analysis of the creator's personal data could impact their freedom of expression. In such contexts, you should consider the relevant legal frameworks beyond data protection.

16.1.4. How do we identify mitigating measures?

Against each identified risk, you should consider options to reduce the level of assessed risk further. Examples of this could be data minimisation or providing opportunities for individuals to opt out of the processing.

You should ask your DPO for advice when considering ways to reduce or avoid risk, and you should record in your DPIA whether your chosen measure reduces or eliminates the risk in question.

16.1.5. How do we conclude our DPIA?

You should record:

- what additional measures you plan to take;
- whether each risk has been eliminated, reduced or accepted;
- the overall level of 'residual risk' after taking additional measures
- the opinion of your DPO, if you have one; and
- whether you need to consult the ICO.

16.1.6. What's different about security in AI compared to 'traditional' technologies?

Some of the unique characteristics of AI mean compliance with data protection law's security requirements can be more challenging than with other, more established technologies, both from a technological and human perspective.

From a technological perspective, AI systems introduce new kinds of complexity not found in more traditional IT systems that you may be used to using. Depending on the circumstances, your use of AI systems is also likely to rely heavily on third party code and/or relationships with suppliers. Also, your existing systems need to be integrated with several other new and IT components, which are also intricately connected.

This complexity may make it more difficult to identify and manage some security risks, and may increase others, such as the risk of outages.

16.1.7. Preventative Controls - 13

Document a DPIA, including thorough assessment of the security risks and the mitigants / controls to reduce the likelihood and impact of an attack.

17. ENISA

17.1. 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>

17.1.1. Networking 8

Do you impose dynamic risk assessment to be conducted by the AI stakeholders?

18. METI (Japan)

18.1. 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

18.1.1. Action Target 1-3

Companies that develop and operate AI systems should, under the leadership of top management, evaluate and re-evaluate in a timely manner their AI proficiency based on the extent of the company's experience in developing and operating AI systems, the number of employees, including engineers, involved in the development and operation of AI systems and their degree of experience, and the degree of AI literacy of these employees with respect to AI technology and ethics, except in situations where a company assesses negative impacts of their AI system are minor based on analyses of Action Targets 1-1 and 1-2 in light of the company's business domain and scale, etc. If the negative impacts are assessed to be minor and no evaluation of AI proficiency is carried out, companies should be prepared to explain their rationale to their stakeholders.

18.1.2. Action Target 3-1

Companies that develop and operate AI systems should, under the leadership of top management, identify a gap between AI governance goals and current state in the AI systems that they are developing and operating, and if any negative impacts are found upon evaluating the impacts resulting from the gap, determine whether or not the negative impacts would be acceptable, taking into account their severity, scope, and frequency of occurrence. They should incorporate processes that prompt a reexamination of how the AI systems should be developed and operated in an appropriate stage such as during AI system design, development, before they are used, and after their usage begins, to address cases where the negative impacts are found not to be acceptable. Those in operations positions should make these processes concrete. In addition, those who are not directly involved in the development and operation of AI systems should be included in the gap analysis between AI governance goals and current state. It should be noted that it would not be appropriate to stop the development or provision of AI simply because a gap was found. As such, gap analysis is merely a step for evaluating negative impacts and simply serve as a starting point for improvement.

18.1.3. Action Target 3-1-1

Companies that develop and operate AI systems should, under the leadership of top management, check whether a standard gap analysis process in their industry is available and incorporate it into their own process if such a process is available.

18.1.4. Action Target 3-4-2

Companies that develop and operate AI systems should, under the leadership of top management, consider defining response guidelines and plans so that upon occurrence of an AI incident or dispute, they can promptly give an explanation to AI system users, identify the extent of the impact and damage, clarify legal responsibilities, consider relief measures and measures to prevent the spread of damage and recurrence, or take other relevant actions. Further, they should consider conducting rehearsal exercise relevant to such guidelines and plans, as appropriate.

19. Qatar Central Bank

19.1. 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

19.1.1. 6.1

An Entity must create a defined AI strategy based on the Entity's needs and risk appetite. It must also be consistent with the Entity's relevant strategies and internal policies and processes.

19.1.2. 6.2

An Entity must conduct a periodic review of its AI strategy at a time consistent with other strategic reviews.

19.1.3. 7.2

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

- Approving the level of AI exposures to be tolerated in the overall risk framework.
- Deciding whether an Entity's existing governance structures are fit for purpose.
- Assigning clear lines of accountability and responsibility.
- Ensuring adequate human resourcing of all AI functions.

19.1.4. 7.6

An Entity must manage risks associated with the use of AI within the enterprise risk management structure.

19.1.5. 8.2.1

Management of AI-related risks through regular audits covering regulatory compliance, governance, customer interactions, risk management process, systems, and control evaluation, and applying required mitigation controls.

19.1.6. 9.1

An Entity must evaluate risks associated with deployment of AI within the organization.

19.1.7. 9.3

An Entity must determine AI risk levels by conducting risk and criticality assessments of the process and functions that the AI System implementation is a part of or connected to.

19.1.8. 9.4

An Entity must manage the process so that the AI risk assessment stays in line with overall process and function risk assessment.

19.1.9. 9.5

An Entity must inform its AI risk score by using additional factors.

19.1.10. 9.5.1

An AI System that allows no direct Human Oversight will probably be judged a higher-than-normal risk, while an AI System that works at delivering "AI-assisted" outcomes, where the final determination is always made by a human with relevant expertise, will likely be judged a lower risk category.

19.1.11. 9.5.2

An Entity assessing the use of an AI System developed by a third-party vendor must consider the access to information allowed and the reputation of the vendor as important risk assessment variables.

19.1.12. 9.6

The Entity will determine whether a specific AI System is "High Risk" using the definition in Section (2), and the risk evaluation and rating per clauses (9.1) to (9.5).

19.1.13. 9.7

Notwithstanding the determination made in clause (9.6), an Entity must classify an AI System as high-risk if or when there is a level of potential harm to natural persons from the system, namely with respect to:

- Interactions determining consumer access to financial services offerings.
- Internal organizational decisions that affect employees in a material manner.
- Processing sensitive personal information.

19.1.14. 9.8

An Entity, if it is the Provider, must have or develop the risk management system that specifically is designed to handle AI related risk with capacity to handle the profile of each AI System.

19.1.15. 10.4

High-risk systems must be highlighted.

19.1.16. 12.3

An Entity should conduct a risk assessment of the outsourcing service provider, including the location of the data when it is processed, stored, and transmitted, and any relevant vendor contracted by the third party.

19.1.17. 20.3

An Entity should ensure that Customers are informed of products and/or services that utilize AI, the associated risks, and limitations of the technology:

- Prior to providing the service initially (for non-high-risk systems like chatbots).
- Each update of an AI System will be treated as a new version requiring Customer notification.
- Use of AI can be disclosed in the general product description or terms of use in line with clause (17.2).
- Each time Customers (or employees) interact with the service for High-Risk AI like credit, insurance, or employee issues.

19.1.18. 20.9

An Entity should collect Customer consent to acceptance of the risks associated with the use of AI prior to providing the service. For non-high-risk AI Systems this could be a one-time event as above or in general terms of use.

19.1.19. 23.3

An Entity must support its exemption request with a clear and documented business case or rationale.

19.1.20. 23.5

An Entity must duly record in the Entity's exemptions Register any approved exemptions and assign an expiration date - the date by which the exemption will be mitigated or resolved by the Entity seeking the exemption.

20. Smart Dubai (UAE)

20.1. 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>

20.1.1. 1.2.2.2

AI operator organisations should identify the likely impact of incorrect automated decisions on AI subjects and, in the case where incorrect decisions are likely to cause significant cost or inconvenience, consider mitigating measures.

20.1.2. 1.2.2.3

AI operator organisations should consider internal risk assessments or ethics frameworks as a means to facilitate the identification of risks and mitigating measures

21. SDAIA (Saudi Arabia)

21.1. 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>

21.1.1. Principle 5 – Reliability & Safety - Plan and Design - 2

Planning to set out a robust and reliable AI system that works with different sets of inputs and situations is essential to prevent unintended harm and mitigate risks of system failures when positioned against unknown and unforeseen events.

21.1.2. Principle 7 – Accountability & Responsibility - Plan and Design - 2

Organizations can put in place additional instruments such as impact assessments, risk mitigation frameworks, audit and due diligence mechanisms, redress, and disaster recovery plans.

21.2. Generative AI Guidelines

Link: <https://sdaia.gov.sa/en/SDAIA/about/Files/GenerativeAIPublicEN.pdf>

Link: <https://sdaia.gov.sa/en/SDAIA/about/Files/GenerativeAIPublicEN.pdf>

21.2.1. 4.5 Privacy & Security - 3

Assess the risks resulting from the use of the GenAI tool according to the AI ethics principles; little or no risk, limited risk, high risk, unacceptable risk.

21.2.2. 4.5 Privacy & Security - 5

The privacy impact assessment and risk management assessment should be continuously revisited to ensure that societal and ethical considerations are regularly evaluated.

22. ISO/IEC

22.1. DIS 27090

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27090:dis:ed-1:v1:en>

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27090:dis:ed-1:v1:en>

22.1.1. 7.4

Model and Mitigation Deterioration Over Time

22.2. TR 27563:2023

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:27563:ed-1:v1:en>

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:27563:ed-1:v1:en>

22.2.1. 7.3

Identify security and privacy concerns

22.2.2. 7.4

Identify security and privacy risks

22.2.3. 7.5

Identify security and privacy controls

22.2.4. 7.6

Identify security and privacy assurance concerns

22.2.5. 7.7

Identify security and privacy plan requirements

22.3. TR 27091

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27091:dis:ed-1:v1:en>

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27091:dis:ed-1:v1:en>

22.3.1. 6.3

Other privacy risks to AI systems

22.4. TS 42119-2:2025

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:ts:42119:-2:ed-1:v1:en>

Link: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:ts:42119:-2:ed-1:v1:en>

22.4.1. 6

Identifying risks in AI systems

23. OECD

23.1. 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

23.1.1. Step 2.1 – Initial scoping of risks

Carry out a scoping exercise to identify where risks may be present and where they may be most significant.

Multiple frameworks exist at the international, regional and national level that describe risks related to the development and use of AI systems and recommend actions companies should take to address those risks. One objective of this guidance is to support enterprises' implementation of other risk management frameworks. While the list of frameworks can inform a range of potentially relevant risks for an enterprise's due diligence efforts, it is non-exhaustive and many of the risks overlap and may be linked to each other. Likewise, not all frameworks are relevant for every enterprise. Each enterprise is expected to identify its priority risk areas based on its individual circumstances.

When prioritising the order in which risks are to be addressed (see Step 2.4), enterprises should take into account that certain risks are closely linked to or may enable others.

As with many new technologies, public and private malign actors may find ways to exploit AI systems. The significant dual-use potential of AI systems and ability to repurpose AI systems can lead to harmful uses even when their design was intended to be innocuous.

23.1.2. Step 2.2 – In-depth assessment of most significant risks

Starting with the most significant areas of risk identified, carry out iterative and increasingly in-depth assessments of prioritised risks related to (1) the enterprise's own activities and (2) the enterprise's business relationships (e.g., suppliers, customers and users).

23.1.3. Step 2.3 – Assess involvement with the actual or potential impact (cause, contribute, directly linked)

Assess the enterprise's involvement with the actual or potential adverse impacts identified. Specifically assess whether the organisation or relevant business relationship caused (or would cause) the adverse impact; or contributed (or would contribute) to the adverse impact; or whether the adverse impact is (or would be) directly linked to its operations, products or services by a business relationship. An enterprise's relationship to adverse impact is not static. It may change, for example as situations evolve and depending upon the degree to which due diligence and steps taken to address identified risks and adverse impacts decrease the risk of the impacts occurring.

23.1.4. Step 2.4 – Prioritise the most significant (i.e., most salient) risks

Drawing from the information obtained on actual and potential adverse impacts, prioritise the most significant (i.e., most salient) risks and adverse impacts for action, based on severity and likelihood. Prioritisation will be relevant where it is not possible to address all potential and actual adverse impacts immediately. Once the most significant adverse impacts are identified and dealt with, the enterprise should move on to address less significant foreseeable impacts.

Where the risk of adverse impacts is most significant will be specific to the enterprise. Thus, it is important for enterprises to demonstrate a credible prioritisation process.

Engaging with relevant stakeholders, including workers, workers' representatives and trade unions, on how to prioritise and publicly communicating the rationale behind how prioritisation decisions are made can be useful for establishing trust in the enterprise's due diligence approach. In some cases, prioritisation may also be informed by domestic legal obligations.

23.1.5. Step 3.1 – Addressing risks that the enterprise causes or contributes to

Cease activities that are causing or contributing to adverse impacts based on the enterprise's assessment of its involvement with the impact.

Develop and implement plans to prevent and mitigate potential (future) adverse impacts.

24. NCSC/NSA/CISA etc

24.1. AI Data Security

Link: https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

Link: https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

24.1.1. 1.10 Conduct ongoing data security risk assessments

Conduct ongoing risk assessments using industry-standard frameworks, such as the NIST SP 800-3r2, Risk Management Framework (RMF) [4] [21], and the NIST AI 100-1, Artificial Intelligence RMF [3]. These assessments should evaluate the AI data security landscape, identify risks, and prioritize actions to minimize security incidents. Continuously improve data security measures to keep pace with evolving threats and vulnerabilities, learn from security incidents, stay up to date with emerging technologies, and maintain a robust security posture.

25. Cloud Security Alliance (CSA)

25.1. AI Controls Matrix

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

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

25.1.1. A&A-06

Establish, document, approve, communicate, apply, evaluate and maintain a risk-based corrective action plan to remediate audit findings, regularly review and report remediation status to relevant stakeholders.

25.1.2. AIS-01

Establish, document, approve, communicate, apply, evaluate and maintain policies and procedures for application security. Review and update the policies and procedures at least annually or after significant system changes.

25.1.3. BCR-01

Establish, document, approve, communicate, apply, evaluate and maintain business continuity management and operational resilience policies and procedures. Review and update the policies and procedures at least annually, or when significant changes occur that could impact risk exposure.

25.1.4. BCR-02

Determine the impact of business disruptions and risks to establish criteria for developing business continuity and operational resilience strategies and capabilities. Review and update the risk assessment and impact analysis at least annually or upon significant changes.

25.1.5. CEK-06

Manage and adopt changes to cryptography-, encryption-, and key management-related systems (including policies and procedures) that fully account for downstream effects of proposed changes, including residual risk, cost, and benefits analysis.

25.1.6. CEK-07

Establish and maintain an encryption and key management risk program that includes provisions for risk assessment, risk treatment, risk context, monitoring, and feedback.

25.1.7. CEK-20

Define, implement and evaluate processes, procedures and technical measures to assess the risk to operational continuity versus the risk of the keying material and the information it protects being exposed if control of the keying material is lost, which include provisions for legal and regulatory requirements.

25.1.8. DCS-05

Classify and document the physical, and logical assets (e.g., applications) based on the organizational business risk. Review and update the assets' classification at least annually or upon significant changes.

25.1.9. DSP-09

Conduct a Data Protection Impact Assessment (DPIA) to evaluate the origin, nature, particularity and severity of the risks upon the processing of personal data, according to any applicable laws, regulations and industry best practices.

25.1.10. DSP-21

Define, implement and evaluate processes, procedures and technical measures to prevent data poisoning in AI models and continuously detect such.

25.1.11. GRC-02

Establish and maintain a formal, documented, and leadership-sponsored AI Risk Management (AIRM) program that includes policies and procedures for identification, evaluation, ownership, treatment, and acceptance of risks.

25.1.12. GRC-03

Review all relevant organizational policies and associated procedures at least annually or when a substantial change occurs within the organization.

25.1.13. GRC-09

Define, document and enforce policies and procedures on the acceptable use of AI services offered by the organization. Ensure effectiveness by continuous risk assessments, reviews and human oversight.

25.1.14. GRC-10

Establish, document, and communicate to all relevant stakeholders an AI Impact Assessment process and its criteria to regularly evaluate the ethical, societal, operational, legal, and security impacts of the AI system throughout its lifecycle.

25.1.15. IAM-08

Review and revalidate user access for least privilege and separation of duties with a frequency that is commensurated with organizational risk tolerance and at least annually, or upon significant changes.

25.1.16. I&S-08

Identify and document high-risk environments.

25.1.17. MDS-01

Define, implement, and evaluate policies, procedures, and technical measures that ensure the security of the Training Pipeline. Regularly review and update policies, procedures and technical measures to address new security threats and best practices.

25.1.18. MDS-06

Define, implement, and evaluate processes and technical measures to assess adversarial threats specific to each AI model.

25.1.19. MDS-11

Perform a risk-based evaluation of the model and model serving infrastructure for model failure. Define and implement measures to mitigate model and model serving infrastructure failures, and regularly evaluate throughout the AI system's lifecycle.

25.1.20. MDS-12

Establish a process to evaluate risk associated with open models. Periodically review these risk factors, and implement a process to monitor and mitigate any determined vulnerabilities.

25.1.21. STA-15

Define and implement a process for conducting security assessments periodically for all organizations within the supply chain.

25.1.22. TVM-08

Use a risk-based model for effective prioritization of vulnerability remediation using an industry recognized framework.

25.1.23. TVM-11

Define and implement processes, procedures and technical measures to apply guardrails to the AI system. Continuously evaluate guardrails for changes in regulatory requirements and risk scenarios.

25.1.24. TVM-12

Define implement and evaluate threat analysis process and procedures to identify, assess and review the threat landscape for Cloud and AI systems. Build threat models according to industry best practices to inform the risk mitigation strategy.

25.1.25. TVM-13

Use a risk-based method for the prioritization and mitigation of threats, leveraging an industry-recognized framework to guide threat decision-making and protection measures.

26. CoSAI

26.1. 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>

26.1.1. 3.3.1. Preparation Phase - Risk Assessment & Threat Modeling

- Identify critical assets in each architecture pattern
- Map ATLAS threat vectors to components
- Assess likelihood and impact
- Prioritize security controls
- Document risk thresholds

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. GOVERN 1.3

Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.

27.1.2. GOVERN 1.4

The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.

27.1.3. GOVERN 1.5

Ongoing monitoring and periodic review of the risk management process and its outcomes are planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review.

27.1.4. GOVERN 2.1

Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.

27.1.5. GOVERN 3.1

Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds).

27.1.6. GOVERN 4.2

Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.

27.1.7. MAP 1.5

Organizational risk tolerances are determined and documented.

27.1.8. MAP 3.2

Potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness – as connected to organizational risk tolerance – are examined and documented.

27.1.9. MAP 4.2

Internal risk controls for components of the AI system, including third-party AI technologies, are identified and documented.

27.1.10. MAP 5.1

Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.

27.1.11. MEASURE 1.1

Approaches and metrics for measurement of AI risks enumerated during the MAP function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.

27.1.12. MEASURE 2.8

Risks associated with transparency and accountability – as identified in the MAP function – are examined and documented.

27.1.13. MEASURE 2.10

Privacy risk of the AI system – as identified in the MAP function – is examined and documented.

27.1.14. MEASURE 3.1

Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.

27.1.15. MEASURE 3.2

Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.

27.1.16. MANAGE 1.2

Treatment of documented AI risks is prioritized based on impact, likelihood, and available resources or methods.

27.1.17. MANAGE 1.3

Responses to the AI risks deemed high priority, as identified by the MAP function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.

27.1.18. MANAGE 1.4

Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented.

27.1.19. MANAGE 2.3

Procedures are followed to respond to and recover from a previously unknown risk when it is identified.

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. PW.1.1

Use forms of risk modeling – such as threat modeling, attack modeling, or attack surface mapping – to help assess the security risk for the software.

Incorporate relevant AI model-specific vulnerability and threat types in risk modeling. Examples of these vulnerability and threat types include poisoning of training data, malicious code or other unwanted content in inputs and outputs, denial-of-service conditions arising from adversarial prompts, supply chain attacks, unauthorized information disclosure, theft of AI model weights, and misconfiguration of data pipelines.

27.2.2. RV.2.2

Plan and implement risk responses for vulnerabilities.

Risk responses for AI models should consider the time and expenses that may be associated with rebuilding them.

Establish and implement criteria and processes for when to stop using an AI model and when to roll back to a previous version and its 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 1.1: Anticipate model capabilities - 5

Assess the degree of uncertainty in these estimates based on the difference between the proxy model and the planned model, how those differences are expected to affect their capabilities, and the reliability and completeness of the evaluations available for proxy models.

27.3.2. Practice 1.1: Anticipate model capabilities - 7

If capability forecasts are uncertain and include the possibility that the model may require increased risk mitigations, consider increasing the frequency of capability measurements during the development process (Objective 4) and expanding the organization's planned risk mitigation measures (Objective 5).

27.3.3. Practice 1.2 Create threat profiles - 3

Use real data, case studies, or expert opinions to inform threat profiles and help identify gaps.

27.3.4. Practice 1.2 Create threat profiles - 4

Develop a plan for identifying and adding threat profiles to this list as future research reveals new potential or ongoing misuse risks (Objective 2).

27.3.5. Practice 1.3: Conduct risk assessments - 1

Consider both quantitative and qualitative assessments of likelihood and impact, if possible.

27.3.6. Practice 1.3: Conduct risk assessments - 2

Account for alternative tools already available to threat actors, such as existing models or other digital tools, and assess the model's marginal risk of misuse relative to that baseline.

27.3.7. Practice 1.3: Conduct risk assessments - 3

Consider the new risk the model may introduce, such as the model's ability to help an actor increase the scale, prevalence, or frequency, decrease the cost, or improve the effectiveness or efficiency of their malicious activity.

27.3.8. Practice 1.3: Conduct risk assessments - 4

Consider malicious actors' potential motivations, willingness, and capacity to enact harm, as well as the number of malicious actors that may exist.

27.3.9. Practice 1.3: Conduct risk assessments - 5

Account for existing mitigations and barriers, such as limitations on access to physical resources needed to enact harm, and the level of societal preparedness to defend against that harm.

27.3.10. Practice 1.3: Conduct risk assessments - 6

Use information about the real-world impact and use of proxy models to inform the assessment; update these assessments as new information about real-world impact materializes.

27.3.11. Practice 1.3: Conduct risk assessments - 7

Update estimates if changes in the broader ecosystem either increase or decrease vulnerability to potential harms.

27.3.12. Practice 1.3: Conduct risk assessments - 8

Recognize areas of uncertainty and sensitivity and account for these areas when communicating and using impact assessments.

27.3.13. Practice 2.1. Map anticipated model capabilities to appropriate risk mitigations to manage misuse risk - 1

Consider the costs and benefits of planned risk mitigations, using both qualitative and quantitative comparisons.

27.3.14. Practice 2.1. Map anticipated model capabilities to appropriate risk mitigations to manage misuse risk - 2

Consider the risk in the context of organizational risk tolerances, which may reflect legal and regulatory obligations.

27.3.15. Practice 2.1. Map anticipated model capabilities to appropriate risk mitigations to manage misuse risk - 3

Identify and articulate evidence to justify the adequacy of planned risk mitigations.

27.3.16. Practice 2.1. Map anticipated model capabilities to appropriate risk mitigations to manage misuse risk - 4

Refine planned methods to mitigate misuse risk periodically based on adjustments to identified threat profiles, changes in factors that affect risk tolerance (such as increases in expected benefits), and information about real-world performance.

27.3.17. Practice 2.1. Map anticipated model capabilities to appropriate risk mitigations to manage misuse risk - 5

As measurement and real-world monitoring is performed, continue to re-assess the safeguards necessary to manage misuse risk for each particular capability.

27.3.18. Practice 2.2 Establish an organizational plan to manage misuse risk - 4

Plan to monitor evidence of real-world misuse and potential risk. Develop processes for adjusting these organizational plans, as well as deployment or development approaches, as necessary.

27.3.19. Practice 3.1: Assess misuse risk from threat actors gaining unauthorized access to the model - 1

Consider possible threat actors' motivations and level of sophistication related to gaining unauthorized access to the model.

27.3.20. Practice 3.1: Assess misuse risk from threat actors gaining unauthorized access to the model - 3

Consider the threat posed by insiders, such as an individual involved in developing or deploying the model who may behave maliciously or collaborate with an external attacker.

27.3.21. Practice 3.2: Maintain security practices sufficient to prevent unauthorized access - 4

Re-assess the risk of unauthorized access as security practices are implemented.

27.3.22. Practice 5.2: Assess misuse risk based on implemented safeguards - 2

Estimate misuse risk based on appropriate measurements conducted under Objective 4 and comparisons to proxy models conducted in Practice 1.1.

27.3.23. Practice 5.3: Adopt appropriate deployment strategies based on misuse risk assessments - 1

Assess residual risk by incorporating the overall assessed misuse risk from the selected deployment strategy and mitigations from implemented safeguards.

27.3.24. Practice 5.3: Adopt appropriate deployment strategies based on misuse risk assessments - 3

Consider whether additional safeguards are feasible to implement prior to deployment, whether additional time could be used to carry out a more reliable estimate of risk, or whether a more limited deployment may be more appropriate given the level of assessed risk.

27.3.25. Practice 5.3: Adopt appropriate deployment strategies based on misuse risk assessments - 4

Consider leaving a buffer between the estimated level of risk—given the implemented safeguards and the deployment strategy—and the associated anticipated real-world risk (hereafter referred to as a 'margin of safety'). This margin of safety could incorporate how threat actors may continue to acquire new knowledge about how to misuse or augment the model after it is deployed²⁶ and how to circumvent safeguards. Consider a larger margin of safety to manage risks that are more severe or less certain.

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. GV.RM-01

Risk management objectives are established and agreed to by organizational stakeholders

27.4.2. GV.RM-02

Risk appetite and risk tolerance statements are established, communicated, and maintained

27.4.3. GV.RM-03

Cybersecurity risk management activities and outcomes are included in enterprise risk management processes

27.4.4. GV.RM-04

Strategic direction that describes appropriate risk response options is established and communicated

27.4.5. GV.RM-05

Lines of communication across the organization are established for cybersecurity risks, including risks from suppliers and other third parties

27.4.6. GV.RM-06

A standardized method for calculating, documenting, categorizing, and prioritizing cybersecurity risks is established and communicated

27.4.7. GV.RM-07

Strategic opportunities (i.e., positive risks) are characterized and are included in organizational cybersecurity risk discussions

27.4.8. GV.PO-01

Policy for managing cybersecurity risks is established based on organizational context, cybersecurity strategy, and priorities and is communicated and enforced

27.4.9. GV.PO-02

Policy for managing cybersecurity risks is reviewed, updated, communicated, and enforced to reflect changes in requirements, threats, technology, and organizational mission

27.4.10. GV.OV-01

Cybersecurity risk management strategy outcomes are reviewed to inform and adjust strategy and direction

27.4.11. GV.OV-02

The cybersecurity risk management strategy is reviewed and adjusted to ensure coverage of organizational requirements and risks

27.4.12. GV.OV-03

Organizational cybersecurity risk management performance is evaluated and reviewed for adjustments needed

27.4.13. GV.SC-01

A cybersecurity supply chain risk management program, strategy, objectives, policies, and processes are established and agreed to by organizational stakeholders

27.4.14. ID.RA-02

Cyber threat intelligence is received from information sharing forums and sources

27.4.15. ID.RA-03

Internal and external threats to the organization are identified and recorded

27.4.16. ID.RA-04

Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded

27.4.17. ID.RA-05

Threats, vulnerabilities, likelihoods, and impacts are used to understand inherent risk and inform risk response prioritization

27.4.18. ID.RA-06

Risk responses are chosen, prioritized, planned, tracked, and communicated

27.4.19. ID.RA-07

Changes and exceptions are managed, assessed for risk impact, recorded, and tracked

27.4.20. RC.RP-04

Critical mission functions and cybersecurity risk management are considered to establish post-incident operational norms

28. Personal Data Protection Commission Singapore (PDPC)

28.1. 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>

28.1.1. 2. Risk management and internal controls - a)

Organisations can consider implementing a sound system of risk management and internal controls that specifically addresses the risks involved in the deployment of the selected AI model.

28.1.2. 2. Risk management and internal controls - b) (i)

Such measures include:

Using reasonable efforts to ensure that the datasets used for AI model training are adequate for the intended purpose, and to assess and manage the risks of inaccuracy or bias, as well as reviewing exceptions identified during model training. Virtually, no dataset is completely unbiased. Organisations should strive to understand the ways in which datasets may be biased and address this in their safety measures and deployment strategies.

28.1.3. 2. Risk management and internal controls - b) (iv)

Reviewing the internal governance structure and measures when there are significant changes to organisational structure or key personnel involved.

28.1.4. 2. Risk management and internal controls - b) (v)

Periodically reviewing the internal governance structure and measures to ensure their continued relevance and effectiveness.

29. Multi Agency

29.1. 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>

29.1.1. Model the threats to your system

As part of your risk management process, you apply a holistic process to assess the threats to your system, which includes understanding the potential impacts to the system, users, organisations, and wider society if an AI component is compromised or behaves unexpectedly. This process involves assessing the impact of AI-specific threats and documenting your decision making.

You recognise that the sensitivity and types of data used in your system may influence its value as a target to an attacker. Your assessment should consider that some threats may grow as AI systems increasingly become viewed as high value targets, and as AI itself enables new, automated attack vectors.