INFOCOMM MEDIA DEVELOPMENT AUTHORITY

AI VERIFY FOUNDATION

AI Verify Testing Framework

For Traditional and Generative AI

The Framework

AI Verify testing framework aims to help companies assess their AI systems against 11 internationally-recognised AI governance principles:

1. Transparency
2. Explainability
3. Repeatability / Reproducibility
4. Safety
5. Security
6. Robustness
7. Fairness
8. Data Governance
9. Accountability
10. Human Agency and Oversight
11. Inclusive Growth, Societal and Environmental Well-being

AI Verify testing framework is consistent with other international AI governance frameworks such as those from ASEAN, European Union, OECD and the US.

WHO SHOULD USE THE FRAMEWORK?

AI System Owners / Developers looking to demonstrate their implemen tation of responsible AI governance practices

Internal Compliance Teams looking to ensure responsible AI practices have been implemented

External Auditors looking to validate your clients' implementation of responsible AI practices

How to use it

Each item in the checklist consists of:

OUTCOME

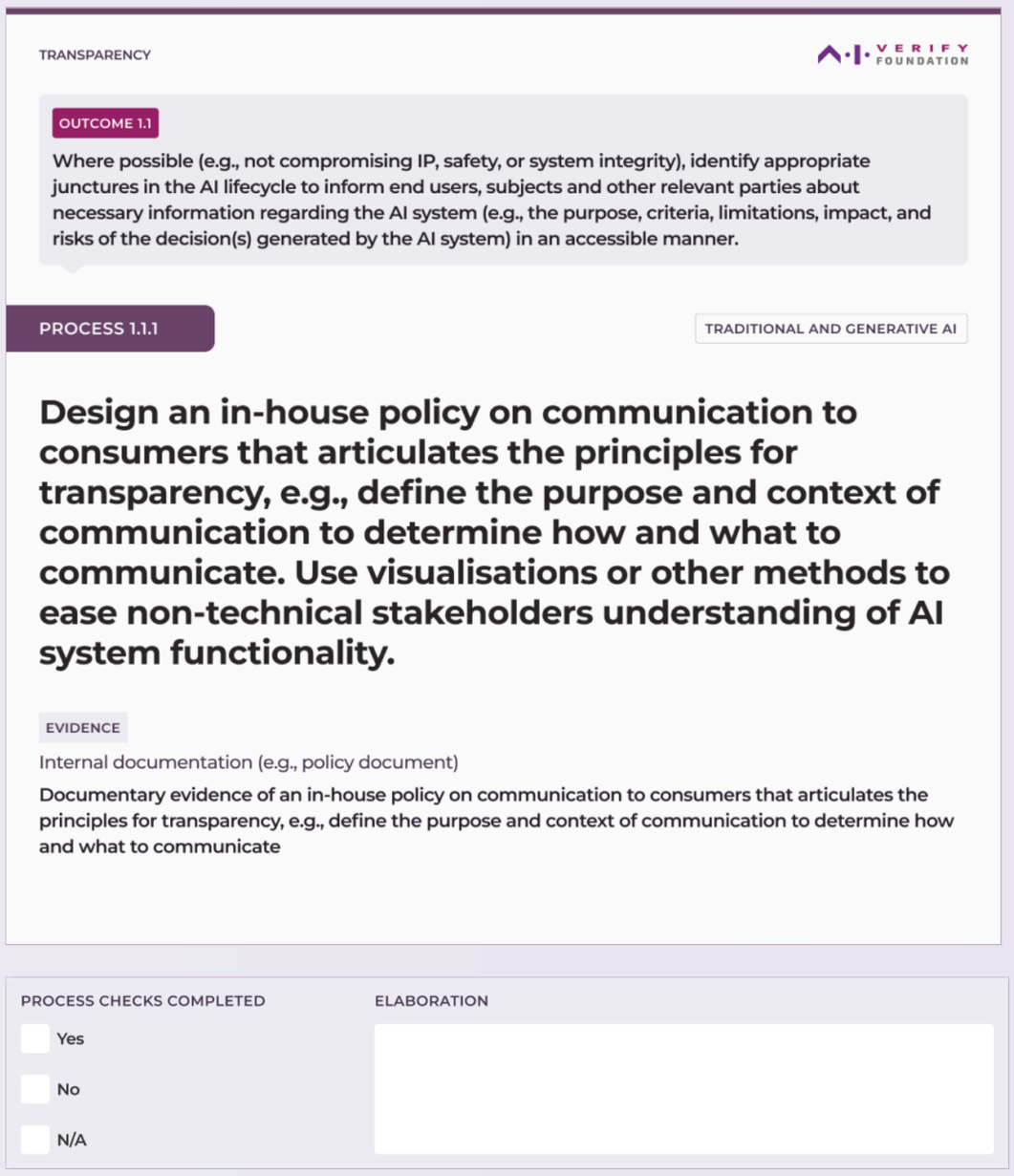
Describe the outcomes that you want to achieve for each principle.

PROCESS

Steps you need to take to achieve desired outcome.

EVIDENCE

Documentary evidence, quantitative and qualitative parameters that validate the process.



For each process, indicate if you have completed process checks and, if necessary, provide a detailed elaboration.

AI VERIFY FOUNDATION

Transparency

Ability to provide responsible disclosure to those affected by AI systems to understand the outcome.

Transparency provides visibility to the intended use and impact of the AI system. It complements existing privacy and data governance measures. Integrating transparency into the AI lifecycle helps ameliorate the problems caused by opaqueness. The desired outcomes and processes focus on ensuring communication mechanisms are in place to enable those affected by AI systems to understand how their data is collected and used, as well as the intended use and limitations of the AI system. This should be done in a manner appropriate to the use case at hand and accessible to the audience.

OUTCOME 1.1

Where possible (e.g., not compromising IP, safety, or system integrity), identify appropriate junctures in the AI lifecycle to inform end users, subjects and other relevant parties about necessary information regarding the AI system (e.g., the purpose, criteria, limitations, impact, and risks of the decision(s) generated by the AI system) in an accessible manner.

PROCESS 1.1.1

TRADITIONAL AND GENERATIVE AI

Design an in-house policy on communication to consumers that articulates the principles for transparency, e.g., define the purpose and context of communication to determine how and what to communicate. Use visualisations or other methods to ease non-technical stakeholders understanding of AI system functionality

EVIDENCE

Internal documentation (e.g., policy document)

Documentary evidence of an in-house policy on communication to consumers that articulates the principles for transparency, e.g., define the purpose and context of communication to determine how and what to communicate

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 1.1

Where possible (e.g., not compromising IP, safety, or system integrity), identify appropriate junctures in the AI lifecycle to inform end users, subjects and other relevant parties about necessary information regarding the AI system (e.g., the purpose, criteria, limitations, impact, and risks of the decision(s) generated by the AI system) in an accessible manner.

PROCESS 1.1.2

TRADITIONAL AND GENERATIVE AI

Inform relevant stakeholders that AI is used in your products and/or services

EVIDENCE

External / internal correspondence

Documentary evidence of communication to relevant stakeholders that AI is used in the organisation's products and/or services

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

PROCESS 1.1.3

GENERATIVE AI

Implement mechanisms such as labelling or disclaimers to enable users, where possible and appropriate, to know when they are interacting with an AI system, and they accurately understand content lineage and origin

EVIDENCE

External / internal correspondence

Documentary evidence of mechanisms or disclosure to relevant stakeholders to identify AI-generated content (e.g., labelling, watermark), it is not error free and there is no human behind the content with whom they might form a relationship. This can be accomplished through methods such as mental model elicitation interviews, questionnaires or interviews with test subjects after test interactions with the model/tool, or observational studies in which interactions with the model/tool are recorded and coded for behaviours that reveal assumptions or misunderstandings

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 1.1

Where possible (e.g., not compromising IP, safety, or system integrity), identify appropriate junctures in the AI lifecycle to inform end users, subjects and other relevant parties about necessary information regarding the AI system (e.g., the purpose, criteria, limitations, impact, and risks of the decision(s) generated by the AI system) in an accessible manner.

PROCESS 1.1.4

TRADITIONAL AND GENERATIVE AI

For decisions made or materially influenced by the AI system, communicate to those using or impacted by the system the: (a) factors leading to the decision, and (b) decision was partially or completely made by an automated system

EVIDENCE

External / internal correspondence

Documentary evidence of communicating to end users the factors that lead to decisions made by AI systems e.g., "You are being shown this product because you bought this item."

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

PROCESS 1.1.5

TRADITIONAL AND GENERATIVE AI

Consult end users at the earliest stages of AI system development to communicate how the technology is used and how it will be deployed

EVIDENCE

External / internal correspondence

Documentary evidence of communication with end users at early stages of AI system development concerning how the technology is used and how it will be deployed

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 1.1

Where possible (e.g., not compromising IP, safety, or system integrity), identify appropriate junctures in the AI lifecycle to inform end users, subjects and other relevant parties about necessary information regarding the AI system (e.g., the purpose, criteria, limitations, impact, and risks of the decision(s) generated by the AI system) in an accessible manner.

PROCESS 1.1.6

TRADITIONAL AND GENERATIVE AI

Publicly report relevant information (e.g., regarding accuracy, intended use cases, and limitations of the AI system) including the risk assessment, mitigation measures and training materials, to stakeholders such as end users and relevant authorities in alignment with regulations, while safeguarding intellectual property rights. Share information and reports to relevant stakeholders (e.g., AI Verify report). Information in the reports to be sufficiently clear and understandable to enable deployers and users as appropriate and relevant to interpret the model/system's output and to enable users to use it appropriately. Transparency reporting to be supported and informed by robust documentation processes such as technical documentation and instructions for use.

EVIDENCE

External / internal correspondence

Documentary evidence of communication with stakeholders concerning the AI system (e.g., model card, system card, reports), which includes (where applicable):
- accuracy;
- confidence scores;
- intended use cases;
- limitations;
- risk assessment;
- results of red-teaming;
- origin and history of training data and generated data

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 1.2

Provide information to guide end users on the proper use of the AI system in an accessible manner

PROCESS 1.2.1

TRADITIONAL AND GENERATIVE AI

Provide information such as the purpose, intended use and intended response of the AI system to end users

EVIDENCE

External / internal correspondence

Documentary evidence of communication with end users concerning the intended use and intended response of the AI system (e.g., Model Card and Data Card)

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 1.3

Provide the necessary information to end users about the use of their personal data to ensure it is processed in a fair and transparent manner

PROCESS 1.3.1

TRADITIONAL AND GENERATIVE AI

Align with existing data protection laws and regulations

EVIDENCE

Internal documentation (e.g., policy document)

Documentary evidence of internal policy requiring alignment with existing data protection laws and regulations, which include:
(in Singapore)
- PDPC's Advisory Guidelines on Key Concepts in the PDPA;
- Guide to Accountability; and
- Guide to Data Protection Impact Assessments.
(outside Singapore)
- Applicable data protection laws/regulations

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 1.4

Provide the necessary information to end users about the use of their personal data to ensure it is processed in a fair and transparent manner

PROCESS 1.4.1

TRADITIONAL AND GENERATIVE AI

Publish a privacy policy on your organization's website to share information about the use of personal data in the AI system (e.g., data practices, and decision-making processes). The general disclosure notice could include:
- Disclosure of third-party engagement
- Definition of data ownership and portability
- Depiction of the data flow and identify any leakages
- Identification of standards the company is compliant with as assurance to customers

EVIDENCE

External / internal correspondence

Documentary evidence of a privacy policy on your organization's website to share information about the use of personal data in the AI system (e.g., data practices and decision-making processes). The general disclosure notice could include:
- Disclosure of third-party engagement;
- Definition of data ownership and portability;
- Depiction of the data flow and identify any leakages; and
- Identification of standards the company is compliant with as assurance to customers

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 1.5

Communicate to end users whenever an incident has occurred

PROCESS 1.5.1

TRADITIONAL AND GENERATIVE AI

Processes in place for communication of incidents to end users

EVIDENCE

Internal documentation

Documentary evidence of communication plan

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 1.6

Provide end users and other external parties with the means to report adverse impacts of AI system to the organisation

PROCESS 1.6.1

TRADITIONAL AND GENERATIVE AI

Processes in place for reporting of adverse impacts of AI system, including protection for whistleblowers

EVIDENCE

Internal documentation

Documentary evidence of processes by which end users and other external parties are able to report adverse impacts of AI system

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 1.7

Make relevant information about the AI system to publicly available

PROCESS 1.7.1

TRADITIONAL AND GENERATIVE AI

Publish generic information on the organisation's development and use of AI systems, and its approach to AI governance through channels such as organisation's website and public regulatory filings

EVIDENCE

External correspondence

Documentary evidence of communication with general public (e.g., info on organisation's website)

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 1.7

Make relevant information about the AI system to publicly available

PROCESS 1.7.2

TRADITIONAL AND GENERATIVE AI

Make model/system cards publicly available - in an accessible form - for specific AI systems where it is warranted by public interest or is helpful to build public trust in the organisation/ its use of AI

EVIDENCE

External correspondence

Documentary evidence of communication with general public (e.g., system card published on website). System card could outline the purpose, intended use, testing approach, known limitations, risk assessment and mitigation measures. It could also cover information on the data used to train or finetune underlying model(s) and the mechanics of the AI system (e.g., which predictive and/or foundation models were used, what specific architecture was used, how the AI model is combined with rule-based systems)

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

AI VERIFY FOUNDATION

Explainability

Ability to assess the factors that led to the AI system's decision, its overall behaviour, outcomes, and implications

Explainability is about ensuring AI-driven decisions can be explained and understood by those directly using the system to enable or carry out a decision to the extent possible. The degree to which explainability is needed also depends on the aims of the explanation, including the context, the needs of stakeholders, types of understanding sought, mode of explanation, as well as the severity of the consequences of erroneous or inaccurate output on human beings. Explainability is an important component of a transparent AI system. This section focuses on system-enabled explainability. However, it may not be possible to provide an explanation for how a black box model generated a particular output or decision (and what combination of input factors contributed to that). In these circumstances, other explainability measures may be required (e.g., accountability and transparent communication). As state-of-the-art approaches to explainability become available, users should refine the process, metrics, and/or thresholds accordingly.

OUTCOME 2.1

For each model being developed, run explainability methods to help users understand the drivers of the AI model

PROCESS 2.1.1

TECHNICAL TESTING | TRADITIONAL AI

Perform analysis to determine feature contributions using technical tools

METRICS

Values obtained from technical tools

Documented testing results from use of testing tools such as AI Verify

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 2.2

Demonstrate a preference for using AI models that can better explain their output (e.g., foundation models with open weights) or that are interpretable by default (e.g., modelling with decision trees) when developing AI systems

PROCESS 2.2.1

TRADITIONAL AND GENERATIVE AI

If choosing a less explainable modelling approach, document the rationale, risk assessments, and trade-offs of the AI model
Apply explainable AI (XAI) techniques (e.g., counterfactual prompts, word clouds) as part of ongoing continuous improvement processes to mitigate risks related to unexplainable AI systems, and verify alignment with intended purpose

EVIDENCE

Internal documentation

Documentary evidence of considerations for the choice of AI model. Considerations include:
- rationale;
- risk assessment; and
- trade-offs

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

AI VERIFY FOUNDATION

Reproducibility

The ability of a system to consistently perform its required functions under stated conditions for a specific period of time, and for an independent party to produce the same results given similar inputs.

Reproducibility is a crucial requirement for achieving system resilience. With software systems, the ability to reproduce an outcome or error is key to identifying and isolating the root cause. This section focus on logging capabilities to monitor the AI system, tracking the journey of a data input through the AI lifecycle, and reviewing the input and output of the AI system.

OUTCOME 3.1

Put in place methods to record the provenance of the AI model, including the various versions, configurations, data transformations, and underlying source code

PROCESS 3.1.1

TRACEABILITY | TRADITIONAL AND GENERATIVE AI

Implement version control of source code and frameworks used to develop the model. For each version of the model, track the code version, as well as the parameters, hyperparameters, and source data used

EVIDENCE

Internal documentation of physical testing

Documentary evidence of version control of source code and frameworks used to develop the model, including considerations of how much version history is required

Each version of the model should track the following:
- code version;
- parameters;
- hyperparameters; and
- source data

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 3.2

Trace the data used by the AI system to make a certain decision(s) or recommendation(s)

PROCESS 3.2.1

TRACEABILITY | TRADITIONAL AND GENERATIVE AI

Log and capture clearly the data used to train a model version, as well as produce inference results using the model (batch scoring or API endpoint)

EVIDENCE

Internal documentation
Documentary evidence of data used.

Data (raw and synthetic data) includes:
- data used to train the AI model;
- data used to produce inference results using the AI model (batch scoring or API endpoint)

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 3.3

Where possible, trace the output (e.g., decision or recommendation) of the AI system to its original source (Note: Mechanistic interpretability is still nascent)

PROCESS 3.3.1

TRACEABILITY | TRADITIONAL AI

Link the inference results of the model (batch scoring or API endpoint) back to the underlying model and source code

EVIDENCE

Internal documentation of physical testing

Documentary evidence of linking the inference results of the model (batch scoring or API endpoint) back to the underlying model and source code

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 3.4

Put in place adequate logging practices to record the decision(s) or recommendation(s) of the AI system

PROCESS 3.4.1

TRACEABILITY | TRADITIONAL AND GENERATIVE AI

Log all inputs and inference outputs of the model, and store them for a reasonable duration so that they can be reviewed

EVIDENCE

Internal documentation

Documentary evidence of log records covering all inputs and inference outputs of the model.

Log records would cover:
- decisions(s) of AI system; and/or
- recommendation(s) of the AI system
- if a human accepted or rejected AI recommendations/decisions, especially when human-in-the-loop is required

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 3.5

Reproduce the training process for every evaluated model (except data)

PROCESS 3.5.1

REPRODUCIBILITY | TRADITIONAL AND GENERATIVE AI

Version control model artefacts by associating each artefact with the version of code, dependencies, and parameters used in training

EVIDENCE

Internal documentation

Documentary evidence of version control model artefacts.

Each artefact includes:
- version of code
- dependencies; and
- parameters used in training

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 3.6

Assess for repeatability by reviewing if the model produces a consistent output based on the same input (Note: this is not relevant when it's time to the retrain model)

PROCESS 3.6.1

REPRODUCIBILITY | TRADITIONAL AND GENERATIVE AI

Calculate multiple inferences; and check if the output falls within the accepted limits of the AI system owner

EVIDENCE

Internal documentation of physical testing

Documentary evidence of assessment conducted to review if the model produces a consistent/similar output based on the same input

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 3.7

Define the process for developing models and evaluate the process

PROCESS 3.7.1

REPRODUCIBILITY | TRADITIONAL AND GENERATIVE AI

Identify a combination of technical metrics and business metrics that AI models are designed to be assessed against

EVIDENCE

Internal documentation

Documentary evidence of metrics of AI models that are designed to be assessed against.

Metrics include:
- technical metrics; and/or
- business metrics

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 3.8

Define the process for developing models and evaluate the process

PROCESS 3.8.1

REPRODUCIBILITY | TRADITIONAL AND GENERATIVE AI

Keep track of experiments (e.g., hyperparameters and model performance) used to develop challenger models, document the rationale for developing these models, and how the final deployed model was arrived at

EVIDENCE

Internal documentation

Documentary evidence of the process in developing the AI model.

The process includes:
- hyperparameters, model performance, and other relevant aspects used to develop challenger models;
- the rationale for developing these models; and
- how the final deployed model was derived

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 3.9

Establish a strategy for reproducing the input data used in the training process for every model

PROCESS 3.9.1

REPRODUCIBILITY | TRADITIONAL AND GENERATIVE AI

Version control the input data used for training where possible. If not possible, avoid changing the raw data at the source, and keep track of the various stages or transformation steps that are part of the data pipeline for AI model development, preferably as a directed acyclic graph (DAG)

EVIDENCE

Internal documentation of physical testing

Documentary evidence of having implemented a strategy for reproducing the input data used in the training process for every model.

This strategy includes:
- data cleaning, data processing, and feature engineering
- maintaining version control of the input data used for training the AI model, where possible; or
- separating data manipulation process into extraction (data versioning) and processing; or
- avoiding changes to the raw data at the source and keeping track of the various stages or transformation steps that are part of the data pipeline for AI model development, preferably as a directed acyclic graph (DAG).

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 3.10

Establish a strategy for ensuring that assumptions still hold across subsequent model retraining process on new input data

PROCESS 3.10.1

REPRODUCIBILITY | TRADITIONAL AI

Record the statistical distribution of input features and output results so that divergence during retraining can be flagged. Monitor input parameters and evaluation metrics for anomalies across retraining runs

EVIDENCE

Internal documentation of physical testing

Documentary evidence of establishing a strategy for ensuring that assumptions still hold across subsequent model retraining process on new input data. For example:
- K-L divergence and K-S test metrics can be used to compare the statistical distributions of inputs/outputs between two training runs
- Moving average and standard deviations can be used to detect a significant change in model performance metrics

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 3.11

Reproduce consistent outputs of the AI system

PROCESS 3.11.1

REPRODUCIBILITY | TRADITIONAL AND GENERATIVE AI

Log audit trail of when and how each model was deployed, including the code used to serve the model, testing/validation data, and what version of the model artefact was used

EVIDENCE

Internal documentation

Documentary evidence of past outputs of deployed AI system, which can include:
- when and how each model was deployed;
- the code used to serve the model; and
- the version of the model artefact used

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 3.12

If using a blackbox model or third party model, assess the vendor's claim on accuracy

PROCESS 3.12.1

REPRODUCIBILITY | TRADITIONAL AND GENERATIVE AI

Curate the test set and apply the test set on the model to review performance

EVIDENCE

Internal documentation of physical testing

Documentary evidence of assessment conducted concerning vendor's claim on the accuracy, if using a blackbox or third party model

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

AI VERIFY FOUNDATION

Safety

AI should not result in harm to humans (particularly physical harm), and measures should be put in place to mitigate harm.

Safety is about ensuring AI systems do not cause any harm, especially physical harm. All systems will have some level of residual risk and must be developed with a preventative approach to risks that are not tolerable. Safety is achieved by reducing risks to a tolerable level. Usually, the higher the perceived risks of a system causing harm, the higher the demands on risk mitigation. This section adopt a risk-based approach to assess the appropriate level of tolerable risk, as well as identify and mitigate potential harm throughout the AI lifecycle. It focuses on model and content safety.

OUTCOME 4.1

Carry out regular tests to evaluate for safety and possible harms (e.g., hallucination and general toxicity)

PROCESS 4.1.1

TECHNICAL TESTING | GENERATIVE AI

Identify relevant and/or use-case appropriate benchmarks

- prepare datasets/prompts that include questions or statements that can solicit possible harms including toxic output
Document instructions given to AI red-teamers
Run benchmarking and redteaming
Testing to take place in secure environments and be performed at several checkpoints throughout the AI lifecycle in particular before deployment and placement on the market to identify risks and vulnerabilities, and to inform action to address the identified AI risks. The results of red-teaming conducted to evaluate the model's/ system's fitness for moving beyond the development stage.
Identify relevant stakeholders (e.g., AI system owner, product manager, risk team) to determine if the test scores are acceptable for the use case
Where applicable, share results with relevant stakeholders

METRICS

Values obtained from technical tools

Documented testing results from use of testing tools such as Project Moonshot

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.2

Carry out an assessment of materiality on key stakeholders

PROCESS 4.2.1

TRADITIONAL AND GENERATIVE AI

Complete and submit the Assessment of Materiality to the appropriate parties who are accountable for the AI system (e.g., AI governance committee, AI system owner, and reviewers) and highlight the risks of the proposed AI solution.
Document the justifications for decisions on materiality and the application of relevant governance and controls to demonstrate to regulators and auditors that sufficient responsibility has been taken by humans to address potential risks.
The materiality assessment could be 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. This could also be a continuous feedback process between AI system operators and stakeholders

EVIDENCE

1) Internal procedure manual 2) Internal documentation (e.g., procedure manual)

Documentary evidence of details of the assessment of materiality on key stakeholders, justifications for decisions on materiality, and the application of relevant governance/controls.

The Assessment of Materiality includes the following impact dimensions (where applicable):
- probability of harm;
- severity of harm;
- Likelihood of threat;
- Extent of human involvement;
- Complexity of AI model;
- Extensiveness of impact on stakeholders;
- Degree of Transparency; and
- Impact on trust

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.3

Assess risks, risk metrics, and risk levels of the AI system in each specific use case, including the dependency of a critical AI system's decisions on its stable and reliable behaviour

PROCESS 4.3.1

TRADITIONAL AND GENERATIVE AI

Document the intended use cases, capabilities, impact and limitations of AI models e.g., via model cards, and approaches for model release
Include relevant AI actors in the risk identification process
Process to obtain feedback for risk measurement such as red-teaming and independent evaluation
Conduct adversarial role-playing exercises red-teaming and chaos testing to identify possible failures
State the safeguard requirements (e.g., what are the risks/failures/ unacceptable outcomes to mitigate against), safeguard plans (e.g., what are the safeguards, evaluate whether the safeguards are sufficient / working as intended to address the requirements)
This documentation should be stored and retrieved together with the model artefact, as well as surfaced during a review process before the model is deployed into production

EVIDENCE

Internal documentation

Documentary evidence of risk assessment done for specific use cases.

This risk assessment includes documenting* the:
- intended use cases, capabilities, and limitations of the AI model (e.g., via model cards)
- plan to halt development or deployment if it poses unacceptable risk
- how the AI model has been adapted, improved or fine-tuned

***Note: This documentation should be stored and retrieved together with the model artefact and surfaced during a review process before the model is deployed into production**

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.3

Assess risks, risk metrics, and risk levels of the AI system in each specific use case, including the dependency of a critical AI system's decisions on its stable and reliable behaviour

PROCESS 4.3.2

GENERATIVE AI

Conduct regular evaluation of risks for models that are fine-tuned, or when adapted to new domains or when risks exceeds organisational risk tolerance
Ensure fine-tuning does not compromise safety and security controls
Develop plan to assess the safeguards after deployment, including how frequent the assessment should occur (e.g., based on time / changes in model capability/performance) and whether the plan is sufficient

EVIDENCE

Internal documentation

Documentary evidence of risk assessment done for specific use cases.

This risk assessment includes documenting* the:
- intended use cases, capabilities, and limitations of the AI model (e.g., via model cards)
- plan to halt development or deployment if it poses unacceptable risk
- how the AI model has been adapted, improved or fine-tuned

***Note: This documentation should be stored and retrieved together with the model artefact and surfaced during a review process before the model is deployed into production**

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.4

Put in place a process to continuously assess, measure and monitor outcomes and risks, including the identification of new methods and technologies for measurement of risks, as well as new risks after deployment

PROCESS 4.4.1

TRADITIONAL AND GENERATIVE AI

Assign a reviewer who is familiar with the downstream use case of an AI model to review the model post-deployment. This process should include model cards/documentation to ensure alignment between intended use cases at modelling and post-deployment. Where applicable, share or publish reports detailing the performance, feedback received, and improvements made

EVIDENCE

Internal documentation (e.g., log, register or database) / External correspondence

Documentary evidence of process for continuous risk monitoring for AI model.

Process includes:
- Assessing, measuring, and monitoring risks at modelling stage
- Assessing general risks associated with a lack of transparency and explainability
- Assessing whether the AI model will be misused (e.g., offensive cyber capabilities and Chemical, Biological, Radiological, Nuclear (CBRN) information) before deployment
- identification of new risks after the post-deployment stage
- Revaluating organisational tolerance to account for unacceptable risks

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 4.5

Assess whether the AI system might fail by considering contingency processes, input features and predicted outcomes to aid communication with stakeholders

PROCESS 4.5.1

TRADITIONAL AND GENERATIVE AI

Where feasible, use AI models that can produce confidence score together with prediction. Low confidence scores may occur when the data contains values that are outside the range of the training data, or for data regions where there were insufficient training examples to make a robust estimate. Implement mechanisms to detect if model might fail or input represents an outlier in terms of training data, e.g., return some "data outlier score" with predictions

EVIDENCE

Internal documentation of physical testing

Documentary evidence of assessment of whether the AI system might fail by considering the input features and predicted outcomes to aid communication to stakeholders

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 4.5

Assess whether the AI system might fail by considering contingency processes, input features and predicted outcomes to aid communication with stakeholders

PROCESS 4.5.2

TRADITIONAL AND GENERATIVE AI

Continuous monitoring of third party systems and fallback plan that monitors the effectiveness of risk controls and mitigation plans
Establish proper contracts with third parties, with clear assignment of liability and responsibilities
Contingency processes to handle failures or incidents (e.g., arising from third-party data or AI systems that are high risk or data redundancy such as model weights and other system artifacts) or overreliance on third party data and systems

EVIDENCE

Internal documentation / External correspondence

Documentary evidence of
- policy for monitoring of third party systems, data redundancy
- external contracts with third parties
- fallback plan / incident response plan to handle incidents and failure

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.6

Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or 'conventional')

PROCESS 4.6.1

TRADITIONAL AND GENERATIVE AI

Implement deployment strategies such as blue-green and canary deployments

EVIDENCE

Internal documentation of physical testing

Documentary evidence of:
- implementation of deployment strategies such as blue-green and canary deployments
- a plan for graceful failure or failover modes

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

PROCESS 4.6.2

TRADITIONAL AND GENERATIVE AI

Maintain backup model server in blue-green deployment mode

EVIDENCE

Internal documentation of physical testing

Documentary evidence of maintenance of the backup model server in blue-green deployment mode

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.6

Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or 'conventional')

PROCESS 4.6.3

TRADITIONAL AND GENERATIVE AI

Where feasible, use AI models that can produce a confidence score together with the prediction. Design the systems that are using the results of the AI model to handle cases where the model fails or has low confidence, falling back to backup model servers or sensible default behaviour

EVIDENCE

Internal documentation of physical testing

Documentary evidence of the use of AI models that can produce a confidence score together with the prediction, and that it can fall back to backup model servers or sensible default behaviour

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 4.6

Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or 'conventional')

PROCESS 4.6.4

TRADITIONAL AND GENERATIVE AI

Close the feedback loop by retraining models with ground truth obtained once models are in production

EVIDENCE

Internal documentation

Documentary evidence of closing the feedback loop by retraining models with ground truth obtained once models are in production

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 4.7

Identify and track residual risk that cannot be measured, mitigated and assess the organisation's tolerance for these risks

PROCESS 4.7.1

TRADITIONAL AND GENERATIVE AI

Document the assessment of the residual risk and provide reasons for the tolerance level
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. For risks difficult to assess or where metrics are not available, put in place risk tracking approaches such as developing a risk reporting matrix, communicating potential risk to affected stakeholders, monitoring risk mitigation plans and reviewing status updates regularly

EVIDENCE

Internal documentation

Documentary evidence of:
- implementation of risk tracking approaches
- assessment of risks that cannot be measured (including explanations such as technological limitations), residual risk and the reasons for the organisation's tolerance for these risks

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.8

Establish a strategy to ensure that the AI system is safe, and does not generate content that violates laws and regulations or is harmful

PROCESS 4.8.1

GENERATIVE AI

Assess existence or level of harms (e.g., bias, intellectual property infringement). Identify content that may violate laws and regulations (e.g., child sexual abuse material (CSAM), non-consensus intimate image (NCII)), inappropriate chemical, biological, radiological and nuclear (CBRN) information), offensive cyber capabilities) Implement measures to prevent AI model from generating such content (e.g., guardrails, content filters, human moderation system) and mitigating steps should it happens (e.g., integrate feedback into AI system updates). Document the mechanisms to prevent AI model from generating content that violates laws and regulations. Monitor and review output regularly for validity, safety, and aligned with socio-cultural norms, trigger alerts for intervention (e.g., use of sentiment analysis to gauge user sentiment)

EVIDENCE

Internal documentation

Documentary evidence of the organisational policies, mitigation measures if the content violates the laws and regulations

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.9

Establish policies for content provenance

PROCESS 4.9.1

GENERATIVE AI

For Use. Understand the AI system's limits, how its output may be utilised and how the AI system interacts with external networks. Identify potential content provenance harms such as deepfakes. Conduct feedback on expectations, concerns, generated content and labels on content. Use feedback to guide design of provenance data-tracking techniques. Identify methods to trace and analyse the origin and modifications of digital content (e.g., c2pa). Provenance data to include an identifier of the service or model that created the content. Integrate tools to enable real-time monitoring of each instance when content is generated, modified or shared, identify data anomalies and verify authenticity of digital content (e.g., digital signatures). Maintain records of changes to content including content by third parties. Implement content provenance management with third parties

EVIDENCE

Internal documentation / External correspondence

Documentary evidence of
- identification of harms
- tracing and analysis of digital content
- verification of authenticity of the digital content
- change record to content including timestamps, metadata and sources
- tamper-proof history of the content
- engagement and educational activities with relevant third parties

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 4.9

Establish policies for content provenance

PROCESS 4.9.2

GENERATIVE AI

For evaluation. Implement evaluation metrics by demographic groups to identify discrepancies in how content provenance mechanisms work across diverse populations. Manage statistical bias related to content provenance through techniques such as re-sampling or adversarial training. Measure effectiveness and reliability of content provenance methodologies such as watermark. Apply TEVV practices for content provenance (e.g., probing a system's synthetic data generation capabilities for potential misuse). Assess how well solutions address risks or harms

EVIDENCE

Internal documentation / External correspondence

Documentary evidence of
- considerations of demographic groups, and evaluation metrics
- technical tests conducted and results detailing the effectiveness of content provenance methodologies
- each instance when content is generated, modified or shared
- TEVV considerations

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 4.10

Scientific integrity and Test, Evaluation, Verification, and Validation (TEVV) considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation

PROCESS 4.10.1

TRADITIONAL AND GENERATIVE AI

For systems that are in experimental stage, put in place a process to document the TEVV considerations. For example, creation of measurement error models for pre-deployment metrics to demonstrate construct validity for each metric. Assess the accuracy, quality, reliability and authenticity of output by comparing to ground truth, and conducting adversarial testing to identify vulnerabilities, potential manipulation or misuse of the system

EVIDENCE

Internal documentation

Documentary evidence of TEVV considerations and policy

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 4.10

Scientific integrity and Test, Evaluation, Verification, and Validation (TEVV) considerations are identified and documented, including those related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), system trustworthiness, and construct validation

PROCESS 4.10.2

GENERATIVE AI

Document fact-checking techniques to verify accuracy of information and retention policy to keep history for TEVV. Use testing techniques to identify AI-generated content and human-generated content

EVIDENCE

Internal documentation

Documentary evidence of TEVV considerations and policy

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

AI VERIFY FOUNDATION

Security

AI security is the protection of AI systems, their data, and the associated infrastructure from unauthorised access, disclosure, modification, destruction, or disruption. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure.

Security risks related to AI systems can be common across other types of software development and deployment, e.g., concerns related to the confidentiality, integrity, and availability of the system and its training and output data; and general security of the underlying software and hardware for AI systems. Security risk management considerations and approaches are applicable in the design, development, deployment, evaluation, and use of AI systems. Security also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Organisations need to develop a risk-based approach to managing AI security. This involves identifying and assessing the risks associated with the use of AI systems and implementing appropriate security controls to mitigate those risks. Organisations should also define the roles and responsibilities of different stakeholders involved in securing AI systems, including developers, operators, and users of AI system.

OUTCOME 5.1

Raise awareness and competency on security risks

PROCESS 5.1.1

TRADITIONAL AND GENERATIVE AI

Provide adequate training and guidance on the security risks of AI to all personnel, including developers, system owners and senior leaders

EVIDENCE

Internal documentation

Documentary evidence that team members have relevant security knowledge and training on threats, vulnerabilities, impact, and mitigation measures relevant to securing AI systems. This can include, where applicable:

- Training records
- Attendance records
- Assessments
- Certifications

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.2

Conduct security risk assessments

PROCESS 5.2.1

TRADITIONAL AND GENERATIVE AI

Apply a holistic process to model threats and risks to an AI system, in accordance with relevant industry standards/best practices. This includes identifying metrics that reflect the effectiveness of security measures

EVIDENCE

Internal documentation (e.g., risk assessment)

Documentary evidence that risk assessment has been done in accordance with the relevant industry standards/guidelines/best practices, with risk mitigation steps and factors taken.

Useful references can include:
- US NIST AI Risk Management Framework
- UK NCSC guidance on secure development and deployment of software applications
- OWASP Secure Software Development Lifecycle (SSDLC)
- CIA triad

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.3

Secure the supply chain

PROCESS 5.3.1

TRADITIONAL AND GENERATIVE AI

Assess and monitor potential security risks of the AI system's supply chain across its life cycle Ensure that suppliers adhere to policy and security standards, or that risks are otherwise appropriately managed. Consider evaluating supply chain components (e.g. through code checking, or against vulnerability databases)

EVIDENCE

Internal documentation

Documentary evidence that supply chain security has been done, these can include:
- Applying secure software development lifecycle practices
- Assessment of the integrity of acquired datasets that is used for training the model or referenced by the model.
- Considering risks of using untrusted 3rd party models.
- Scanning and/or Sandboxing untrusted models where relevant.
- Limiting sensitive data from being provided or uploaded.
- Evaluation of dependent software libraries.

Useful references may include:
- UK NCSC supply chain security guidance
- MITRE supply chain securityframework
- ETSI GR SAI 002 Securing AI Data Supply Chain Security

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.4

Consider security benefits and trade-offs when selecting the appropriate model to use

PROCESS 5.4.1

TRADITIONAL AND GENERATIVE AI

When developing or selecting an appropriate AI model for your system, consider factors which may affect its security (such as complexity, explainability, interpretability, and sensitivity of training data)

EVIDENCE

Internal documentation

Documentary evidence of considering security trade-offs when selecting the appropriate model to use:
- Consideration of model complexity
- Consider explainability of the model
- Assessing the need to use sensitive data
- Model hardening, if appropriate.

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.5

Identify, track and protect AI-related assets

PROCESS 5.5.1

TRADITIONAL AND GENERATIVE AI

Understand the value of AI-related assets, including models, data, prompts, logs and assessments. Have processes to track, authenticate, version control, and secure assets

EVIDENCE

Internal documentation (e.g., asset management document)

Documentary evidence that AI-related assets have been identified, tracked and protected. This can include:
- Documentation and backup of the data, codes, test cases and model, including any changes made and by whom.
- Secure data at rest, and data in transit.
- Have regular backups in event of compromise.
- Implement controls to limit what AI can access and generate, based on sensitivity of the data.

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.6

Secure the AI development environment

PROCESS 5.6.1

TRADITIONAL AND GENERATIVE AI

Apply standard infrastructure security principles, such as implementing appropriate access controls and logging/monitoring, segregation of environments, and secure-by-default configurations

EVIDENCE

Internal documentation (e.g., access control management document)

Documentary evidence that the development environment has been secured. This can include:
- Appropriate access controls to APIs, models and data, logs, and the environments that they are in, following the principle of least privilege.
- Access logging and monitoring
- Configurations secure by default

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.7

Secure the deployment infrastructure and environment of AI systems

PROCESS 5.7.1

TRADITIONAL AND GENERATIVE AI

Apply standard infrastructure security principles, such as access controls and logging/monitoring, segregation of environments, secure-by-default configurations, and firewalls

EVIDENCE

Internal documentation (e.g., access control management document)

Documentary evidence that the deployment environment has been secured. This can include:
- Ensuring contingency plans are in place to mitigate disruption or failure of AI services.
- Appropriate access controls to APIs, models and data, logs, and the environments that they are in, following the principle of least privilege.
- Access logging and monitoring
- Configurations secure by default
- Implementing Firewalls

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.8

Establish incident management procedures

PROCESS 5.8.1

TRADITIONAL AND GENERATIVE AI

Put in place appropriate incident response, escalation and remediation plans

EVIDENCE

Internal documentation (e.g. Incident response playbook)

Documentary evidence of incident management procedures, including:
- Having plans to address different attack and outage scenarios. Implement measures to assist investigation.
- Following appropriate guidance when applying logging and auditing logs
- Regularly reassess plans as the system changes.
- Reporting to the relevant stakeholders and authority when an alert has been raised or an investigation has concluded that a cyber incident has occurred
- Have regular backups in event of compromise.

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.9

Release AI systems responsibly

PROCESS 5.9.1

TRADITIONAL AND GENERATIVE AI

Release models, applications or systems only after subjecting them to appropriate and effective security checks and evaluation

EVIDENCE

Internal documentation (e.g. Test performance documents)

Documentary evidence of responsible release of AI, including:
- Verifying model/data singnature and hashes before deployment and periodically.
- Benchmark and test AI before release.
- Security testing on AI systems.

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.10

Monitor AI system inputs

PROCESS 5.10.1

TRADITIONAL AND GENERATIVE AI

Monitor and log inputs to the AI system, such as queries, prompts and requests. Proper logging allows for compliance, audit, investigation and remediation

EVIDENCE

Internal documentation (e.g., log files)

Documentary evidence of monitoring AI system inputs, including:
- Validating/Monitoring inputs to the model and system for possible attacks and suspicious activity.
- Monitoring/Limiting rate of queries

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.11

Monitor AI system outputs and behaviour

PROCESS 5.11.1

TRADITIONAL AND GENERATIVE AI

Monitoring models after deployment to make sure they are performing as intended, and alert system owners to potential issues (whether caused by adversarial attacks or otherwise)

EVIDENCE

Internal documentation (e.g., log files)

Documentary evidence of monitoring AI system outputs and behaviour, including:
- Monitoring model outputs and model performance.
- Ensuring adequate human oversight to verify model output, when viable or appropriate, such as monitoring for anomalous behaviour that might indicate intrusions, compromise, or data drift.

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.12

Adopt a secure-by-design approach to updates and continuous learning

PROCESS 5.12.1

TRADITIONAL AND GENERATIVE AI

Ensure risks associated to model updates have been considered and appropriately managed

EVIDENCE

Internal documentation (e.g., risk management document)

Documentary evidence of adopting a secure-by-design approach to continuous learning and updates, including:
- Treating major updates as new versions and integrate software updates with model updates and renewal.
- Treating new input data used for training as new data.

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.13

Establish a vulnerability disclosure process

PROCESS 5.13.1

TRADITIONAL AND GENERATIVE AI

Put in place a feedback process for users to share any findings of concern, which might uncover potential vulnerabilities to the system

EVIDENCE

Internal documentation (e.g. Feedback channels and procedure)

Documentary evidence of a vulnerability disclosure process, this can include:
- Open lines of communication.
- Sharing findings with appropriate stakeholders and authorities.

Useful references/resources may include:
- SingCERT Vulnerability disclosure policy
- UK NCSC Vulnerability disclosure Toolkit

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 5.14

Ensure proper data and model disposal

PROCESS 5.14.1

TRADITIONAL AND GENERATIVE AI

Implement proper and secure disposal/destruction of data and models in accordance with relevant industry standards or regulations

EVIDENCE

Internal documentation

Documentary evidence of ensuring proper and secure disposal/destruction of data and models in accordance with data privacy standards and/or relevant rules and regulations.

Examples include crypto shredding or degaussing

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

AI VERIFY FOUNDATION

Robustness

AI system should be resilient against attacks and attempts at manipulation by third party malicious actors, and can still function without producing undesirable output despite unexpected input

Robustness requires that AI systems maintain its level of performance under any circumstances, including potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner. This section focuses on the technical robustness of the AI system throughout its AI life cycle, to assess the proper operation of a system as intended by the system owner. This section should be carried out alongside established cybersecurity testing regimes for AI systems, to ensure overall system robustness

OUTCOME 6.1

Put in place a series of tests to document and monitor the AI system's performance (e.g., accuracy)

PROCESS 6.1.1

TECHNICAL TESTING | TRADITIONAL AND GENERATIVE AI

Traditional AI - Calculate accuracy metrics at testing time. Add noise to dataset and compare the 2 accuracy metrics (between the dataset with and dataset without noise). The accuracy metric is calculated based on the matching ground truth labels against the predictions.

Generative AI - Benchmarking (to compare AI model's performance against time / another model), with datasets that are relevant to the use case. For example, adding noise / perturbate (invariance testing such as synonym, typo, punctuation or adversarial perturbation) to the original prompts in the benchmark and tested against the same metric

METRICS

Values obtained from technical tools

Documented testing results from use of testing tools such as AI Verify and Project Moonshot

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 6.2

Put in place measures to ensure the quality of data used to develop the AI system

PROCESS 6.2.1

TRADITIONAL AND GENERATIVE AI

- Implement measures to ensure data is up-to-date, complete, and representative of the environment the system will be deployed in
- Log training run metadata to do comparison in production, e.g., parameters, and version model to monitor model staleness
- Monitor production versus training data characteristics at production stage e.g., statistical distribution, data types, and validation constraints, to detect data and concept drift

EVIDENCE

Internal documentation of physical testing

Evidence of measures implemented that documents:
- Performance metrics (e.g., accuracy, AUROC, AUPR)
- Prediction confidence score, variation ratio for the original prediction, predictive entropy
- That data is of high quality, up-to-date, complete, and representative of the environment the system will be deployed in

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 6.3

Review factors that may lead to a low level of accuracy of the AI system and assess if it can result in critical, adversarial, or damaging consequences

PROCESS 6.3.1

TRADITIONAL AND GENERATIVE AI

Document intended use cases, risks, limitations (e.g., in model cards)

EVIDENCE

Internal documentation

Documentary evidence of intended use cases, risks, and limitations in model cards

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 6.4

Consider whether the AI system's operation can invalidate the data or assumptions it was trained on e.g., feedback loops, user adaptation, and adversarial attacks

PROCESS 6.4.1

TRADITIONAL AND GENERATIVE AI

Document intended use cases, risks, limitations (e.g., in model cards)

EVIDENCE

Internal documentation

Documentary evidence of intended use cases, risks, and limitations in model cards

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 6.5

Put in place a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness

PROCESS 6.5.1

TRADITIONAL AND GENERATIVE AI

Implement a review process that highlights changes in code (e.g., training, serving), input data (e.g., raw data, features), and output data (e.g., inference results, performance metrics)

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of mechanism to evaluate when an AI system has been changed to merit a new review of its technical robustness

Mechanism should include a review process that highlights changes in:
- code (training, serving);
- input data (e.g., raw data, features); and
- output data (e.g., inference results, performance metrics)

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 6.6

Establish a strategy to monitor and mitigate the risk of black box attacks on live AI systems

PROCESS 6.6.1

TRADITIONAL AND GENERATIVE AI

Implement methods to mitigate known adversarial attacks at training time, including decisions whether to adopt / not adopt the methods.

Note: It may not be possible for all models (e.g., if the model is deterministic or with a model with high level of interactivity with users)

EVIDENCE

Internal documentation of physical testing

Documentary evidence of implementing methods to mitigate adversarial attacks at training time, including decisions on whether to adopt / not adopt the methods

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 6.6

Establish a strategy to monitor and mitigate the risk of black box attacks on live AI systems

PROCESS 6.6.2

TRADITIONAL AND GENERATIVE AI

Monitor requests made to live AI system, e.g., frequency and feature distribution of queries, in order to detect whether it is being used suspiciously

EVIDENCE

Internal documentation of physical testing

Documentary evidence of monitoring requests made to live AI system, e.g., frequency and feature distribution of queries, in order to detect whether it is being used suspiciously

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 6.6

Establish a strategy to monitor and mitigate the risk of black box attacks on live AI systems

PROCESS 6.6.3

TRADITIONAL AND GENERATIVE AI

Take action on users who exhibit suspicious activity, e.g., flag for review, rate-limit or block further requests, suspend user accounts

EVIDENCE

Internal documentation of physical testing

Documentary evidence of action taken on users who exhibit suspicious activity.

Possible actions include to:
- flag for review;
- rate-limit or block further requests; and
- suspend user accounts

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 6.7

Establish a strategy to ensure that the AI system is valid and reliable

PROCESS 6.7.1

TRADITIONAL AND GENERATIVE AI

Establish policy to evaluate, continuous monitor and conduct periodic validation of the quality of the model's output, capabilities and robustness of safety measures before and after models have gone live. This includes:
- Model performance, e.g., monitor feature drift, inference drift, accuracy against ground truth
- Application performance, e.g., latency, throughput, error rates, near-misses, and negative impacts
- Continuously assess the quality of the output(s) of the AI system and ensure that the operating conditions of a live AI system match the thesis under which it was originally developed
- Evaluate AI system performance in real-world scenarios to observe its behaviour in practical environments and reveal issues that might not surface in controlled and optimized testing environments
- Regular review of safety of the AI system (e.g., implement and assess guardrails)

EVIDENCE

Internal documentation of physical testing

Documentary evidence of the conduct of continuous monitoring and periodic validation even after models have gone live.

This can include:
- Notifications to admins when a model/system exceeds some thresholds and the system is paused (if safe to do so) until the model can be improved. Any decisions that have been made/ implemented while the AI was below a threshold should be flagged for re-evaluation and potentially redress/remediation if harm occurred
- Model performance (e.g., monitor feature drift, inference drift, accuracy against ground truth)
- Application performance (e.g., latency, throughput, error rates)

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 6.7

Establish a strategy to ensure that the AI system is valid and reliable

PROCESS 6.7.2

GENERATIVE AI

Establish policy that:

- States extent of human domain knowledge used to improve the AI system (e.g., RLHF, fine tuning, retrieval-augmented generation)
- Review and verify sources and citations in generated output
- Track instances of anthropomorphising in AI system (where applicable)

EVIDENCE

Internal documentation

Documentary evidence that policy is established and implemented:
- Implementation of structured prompt logging and validation, ensuring domain knowledge sources are explicitly tracked
- Use of automated citation verification tools to ensure LLM responses match external references
- Linking of specific content chunks to questions and answers, providing sentence-level context correlation to ensure traceable answer origins

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

AI VERIFY FOUNDATION

Fairness

AI should not result in unintended and inappropriate discrimination against individuals or groups.

Fairness is about designing AI systems that avoid creating or reinforcing unfair bias in the AI system, based on the intended definition of fairness for individuals or groups, that is aligned with the desired outcomes of the AI system. This section focuses on testing the ability of the AI system to align with the intended fairness outcomes, throughout the AI lifecycle.

OUTCOME 7.1

Carry out a quantitative analysis to measure algorithmic fairness of a model's predictions against the ground truth

PROCESS 7.1.1

TECHNICAL TESTING | TRADITIONAL AI

AI system owner decides on the fairness metrics for the sensitive feature. To measures the disparity of error rates across groups, based on the selected fairness metrics

METRICS

Values obtained from technical tools

Documented testing results from use of testing tools such as AI Verify

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 7.2

Carry out tests to evaluate for safety and possible harms against individuals or groups

PROCESS 7.2.1

TECHNICAL TESTING | GENERATIVE AI

Identify relevant and/or use-case appropriate benchmarks - prepare datasets/prompts that include questions or statements that can solicit possible harms including toxic output against individuals or groups
Document instructions given to AI red-teamers
Run benchmarking and redteaming
Identify relevant stakeholders (e.g., AI system owner, product manager, risk team) to determine if the test scores are acceptable for the use case
Where applicable, share results with relevant stakeholders

METRICS

Values obtained from technical tools

Documented testing results from use of testing tools such as Project Moonshot

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 7.3

Assess within-group fairness (also known as individual fairness)

PROCESS 7.3.1

TRADITIONAL AI

Apply counterfactual fairness assessment

EVIDENCE

Internal documentation

Documentary evidence of counterfactual fairness assessment

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 7.4

Put in place processes to test for potential biases during the entire lifecycle of the AI system, so that practitioners can act to mitigate biases based on feedback (e.g., biases due to possible limitations stemming from the composition of the used data sets such as a lack of diversity and non-representativeness)

PROCESS 7.4.1

TRADITIONAL AND GENERATIVE AI

Monitor the changes in fairness metric values in the lifecycle of the AI system.

EVIDENCE

Internal documentation of physical testing

Documentary evidence of implemented processes to test for potential biases during the entire lifecycle of the AI system

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 7.5

Establish a strategy for the selection of fairness metrics that are aligned with the desired outcomes of the AI system's intended application

PROCESS 7.5.1

TRADITIONAL AI

Consider using Fairness Decision Tree (e.g., AI Verify, Aequitas) to select the appropriate metric(s) for your application

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of strategy/process undertaken to select fairness metrics that align with the desired outcomes of the AI system's intended application. For example, Binary and Multiclass Classification
- Equal Parity
- Disparate Impact
- False Negative Rate Parity
- False Positive Rate Parity
- False Omission Rate Parity
- False Discovery Rate Parity
- True Positive Rate Parity
- True Negative Rate Parity
- Negative Predictive Value Parity
- Positive Predictive Value Parity

Regression
- Mean Absolute Error Parity
- Mean Square Error Parity

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 7.6

Define sensitive features for the organisation that are consistent with the legislation and corporate values

PROCESS 7.6.1

TRADITIONAL AI

Identify the sensitive features and their privileged and unprivileged groups where feasible

EVIDENCE

Internal documentation

Documentary evidence of identification of sensitive features and its privileged and unprivileged groups. Examples of sensitive features could include religion, nationality, birthplace, gender, and race. Also refer to country-specific guidelines e.g., Singapore's Tripartite Guidelines on Fair Employment Practices and UK Equality Act

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 7.7

Define sensitive features for the organisation that are consistent with the legislation and corporate values

PROCESS 7.7.1

TRADITIONAL AI

Where feasible, consult the impacted communities on the correct definition of fairness (e.g., representatives of elderly persons or persons with disabilities), values and considerations of those impacted (e.g., individual's preference)

EVIDENCE

External / internal correspondence

Documentary evidence of consultations conducted with impacted communities on the correct definition of fairness

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 7.8

Establish a strategy or a set of procedures to check that the data used in the training of the AI model, is representative of the population who make up the end-users of the AI model

PROCESS 7.8.1

TRADITIONAL AND GENERATIVE AI

Perform exploratory data analysis. Test the representation of each group in the data. To mitigate concerns of model collapse, resample data or collect more data if:
- a particular group is severely underrepresented or
- training data is overly homogenous or
- GAI-produced

EVIDENCE

Internal documentation of physical testing

Documentary evidence of the establishment of a strategy or a set of procedures to check that the data used in the training of the AI model, is representative of the population who make up the end-users of the AI model

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 7.9

Put in place a mechanism that allows for the flagging of issues related to bias, discrimination, or poor performance of the AI system

PROCESS 7.9.1

TRADITIONAL AND GENERATIVE AI

Monitor threshold violations of fairness metrics post-deployment and for actual harms

EVIDENCE

Internal documentation of physical testing

Documentary evidence of
- monitoring of threshold violations of fairness metrics
- obtaining feedback from those impacted by the AI system, offering redress and remediation option if feasible

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 7.10

Put in place appropriate mechanisms to ensure fairness in your AI system

PROCESS 7.10.1

TRADITIONAL AND GENERATIVE AI

Monitor metrics for the latest set of data for the model currently being deployed on an ongoing basis.

EVIDENCE

Internal documentation of physical testing

Documentary evidence of monitoring metrics for the latest set of data for the model currently being deployed on an ongoing basis

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 7.11

Address the risk of biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness), by applying appropriate adjustments on data samples of minorities

PROCESS 7.11.1

TRADITIONAL AND GENERATIVE AI

Where possible, handle imbalanced training sets with minorities.
Examples:
- Oversample minority class
- Undersample majority class
- Generate synthetic samples (SMOTE)

EVIDENCE

Internal documentation of physical testing

Documentary evidence of addressing the risk of biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness), by applying appropriate adjustments on data samples of minorities

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

AI VERIFY FOUNDATION

Data Governance

Governing data used in AI systems, including putting in place good governance practices for data quality, lineage, and compliance.

Data Governance ensures that data is properly managed over time throughout the enterprise, including establishing authority, management, and decision-making parameters related to the data used in AI systems.

OUTCOME 8.1

Put in place measures to ensure data quality over time

PROCESS 8.1.1

TRADITIONAL AND GENERATIVE AI

Verify the quality of data used in the AI system, including intellectual property and privacy risks as well as unsafe content (e.g., CBRN). This may include the following:
- accuracy in terms of how well the values in the dataset match the true characteristics of the entity described by the dataset
- completeness in terms of attributes and items e.g., checking for missing values, duplicate records
- veracity in terms of how credible the data is, including whether the data originated from a reliable source
- How recently the dataset was compiled or updated
- Relevance for the intended purpose
- Integrity in terms of how well extraction and transformation have been performed if multiple datasets are joined;
- Usability in terms of how the data are tracked and stored in a consistent, human-readable format
- Providing distribution analysis e.g., feature distributions of input data
- Seek feedback from relevant stakeholders

EVIDENCE

Internal documentation

Documentary evidence that proves due diligence has been done to ensure the quality of data, including intellectual property and privacy risks to ensure use of proprietary or sensitive data is consistent with applicable laws. This can include the use of relevant processes or software that:
- Conducts validation schema checks
- Identifies possible errors and inconsistencies at the exploratory data analysis stage before training the dataset
- Assigns roles to the entire data pipeline to trace who manipulated data and by which rule
- Allows for review before a change is made
- Unit tests to validate that each data operation is performed correctly prior to deployment
- Allow for periodic reviewing and update of datasets
- Allow for continuous assessment of the quality of the input data to the AI system, including drift parameters and thresholds, where applicable

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 8.2

Put in place measures to understand the lineage of data, including knowing where the data originally came from, how it was collected, curated, and moved within the organisation over time

PROCESS 8.2.1

TRADITIONAL AND GENERATIVE AI

Maintain a data provenance record to ascertain the quality of the data based on its origin and subsequent transformation. This could include the following:
- Take steps to understand the meaning of and how data was collected, including known assumptions and practices
- Evaluate data and content flow within AI systems
- Document data usage and related concerns
- Ensure any data labelling is done by a representative group of labellers and instructions are provided
- Document the procedure for assessing labels for bias
- Trace potential sources of errors
- Update to data and monitoring other changes that may impact the verifiability of content origins
- Attribute data to their sources
- Anonymise data, leverage privacy output filters, remove personally identifiable information (PII)

EVIDENCE

Internal documentation

Documentary evidence of a data provenance record that includes the following info, where applicable:
- clear explanations of what data is used, how it is collected and why
- source of data and its labels
- who the labellers were and whether bias tests were conducted to assess if the labelled data was biased (e.g., bias assessment)
- how data is transformed over time (e.g., data deletions, rectification requests)
- risk management if the origin of data is difficult to be established

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 8.3

Ensure data practices comply with relevant regulatory requirements or industry standards

PROCESS 8.3.1

TRADITIONAL AND GENERATIVE AI

Ensure that data protection, consent, collection, use, withdrawal, retention and assessment has been carried out in accordance with the relevant regulatory requirements and/or industry standards. Disclose policies and practices where appropriate. Mitigation steps have been taken. For example:
- review training data for CBRN information, and intellectual property, and where appropriate, remove it.
- Implement measures to prevent, flag, or take other action in response to outputs that reproduce particular training data (e.g., plagiarized, trademarked, patented, licensed content or trade secret material)

EVIDENCE

1) Internal documentation 2) Assessment documentation or certification(s)

Documentary evidence that assessment has been done in accordance with the relevant data protection, copyright and intellectual property laws/ standards/guidelines/best practices. For example:
- applicable data protection laws and regulations such as Singapore's Personal Data Protection Act, European Data Governance Act
- Singapore's Data Protection Trustmark
- Asia Pacific Economic Cooperation Cross Border Privacy Rules and Privacy Recognition for Processors
- OECD Privacy Principles
- Recognised data governance standards from international standard bodies (e.g., ISO, US NIST, IEEE)

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 8.3

Ensure data practices comply with relevant regulatory requirements or industry standards

PROCESS 8.3.2

GENERATIVE AI

Detect presence of personally identifiable information or sensitive data in generated output Implement techniques such as anonymisation, synthetic data generation and privacy enhancing technologies to minimise risks associated with linking AI-generated content back to humans

EVIDENCE

Internal documentation

Documentary evidence of technical tests or detection results, software/developer documentation. Technical tools could include Project Moonshot for red teaming

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 8.4

Ensure team competency in data governance

PROCESS 8.4.1

TRADITIONAL AND GENERATIVE AI

Ensure that relevant team members are knowledgeable about their roles and responsibilities for data governance. Relevant team members include any employee that is involved in managing and using the data for the AI system. For example, having a data policy team to manage the tracking of data lineage with proper controls

EVIDENCE

Internal documentation

Documentary evidence that team members have relevant knowledge and training on data governance. This can include, where applicable:
- Training records
- Attendance records
- Assessments
- Certifications
- Feedback forms

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 8.5

Address AI risks associated with third party entities including risks of infringement of a third party's rights

PROCESS 8.5.1

TRADITIONAL AND GENERATIVE AI

Categorise different types of content associated with third party rights (e.g., copyright, intellectual property) to be informed on the use of external data
Implement practices on how third party intellectual property and training data will be used, stored and collected
Conduct monitoring of output for privacy risk and data disclosure
Implement process to respond to potential intellectual property infringement claims or other rights
Educate relevant stakeholders such as third parties on best practices for managing risks
Obtain feedback and recommendations from organizational boards or committees when using third-party pre-trained models

EVIDENCE

Internal documentation

Documentary evidence of:
- Risk assessment and monitoring related to third party entities have been conducted
- Process to respond to third parties on various issues such as infringement claims or other rights
- Feedback been incorporated when using third-party pre-trained models

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

AI VERIFY FOUNDATION

Accountability

AI systems should have organisational structures and actors accountable for the proper functioning of AI systems.

Accountability is about having clear internal governance mechanisms for proper management oversight of the AI system's development and deployment.

OUTCOME 9.1

Establish policies on the development and use of AI that is aligned with other organisational policies

PROCESS 9.1.1

TRADITIONAL AND GENERATIVE AI

Identify acceptable, unacceptable and illegal uses of AI model and output, including criteria for the AI model not respond to (e.g., queries to chatbots)
Process to engage end users on their expectations and needs (e.g. providing general usage agreements that scope its use)
Incorporate relevant stakeholders such as end users' expectations/ needs in the responsible development and use of the AI system
Implement mechanisms for recourse
Compile violations, take-down requests, intellectual property infringement
AI policy reviewed regularly to ensure its continued suitability, adequacy and effectiveness, proportional to the identified risks, including making adjustments to organisational roles and components

EVIDENCE

External / internal correspondence

Documentary evidence of
- the overall AI policy of the organisation, e.g., principles that guide AI-related activities, processes for handling deviations and exceptions to policy. acceptable use policy, including mechanism for resource
- policies that address proprietary and open source technologies and data and third party personnel
- attempts to understand end users needs and mitigate risks related to its misuse
- reports and statistics of violations
- regular review of the AI policy

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 9.2

Intended purposes, potentially beneficial uses, contexts specific laws, norms and expectations, interactions with the AI systems, and prospective settings in which the AI system will be deployed are understood and documented

PROCESS 9.2.1

TRADITIONAL AND GENERATIVE AI

Refer to 4.2 and 4.3. In addition, document the use, norms and expectations in which the AI system will be deployed. Considerations to include relevant stakeholders (e.g., users, domain and socio-cultural experts), and how they interact with the AI system

EVIDENCE

Internal documentation

Documentary evidence of norms, expectations and users' interactions with the AI system

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 9.3

Establish processes to ensure that its usage of services, products or materials are in support of the responsible development and use of the AI system

PROCESS 9.3.1

TRADITIONAL AND GENERATIVE AI

Select AI system suppliers which align with organisation's approach.

When AI system suppliers do not perform as intended, there are processes to take remedial actions.

EVIDENCE

Internal documentation/ External correspondence

Documentary evidence of processes to select suppliers which align with the organisation's approach.

Documentary evidence of correspondence with suppliers to take corrective action.

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 9.4

Organisation has in place a policy to deactivate, decommission and phase out AI systems safely and in a manner that does not affect the organisation

PROCESS 9.4.1

TRADITIONAL AND GENERATIVE AI

Put in place a:
- manual on how to deactivate AI model while considering factors such as data security, retention and leakage, as well as users' dependency on the AI model
- communications plan to inform AI stakeholders as part of deactivation or disengagement process

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of the manual and the considerations / criteria to deactivate or phase out AI models

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.5

Establish clear internal governance mechanisms to ensure clear roles and responsibilities for the use of AI by the organisation

PROCESS 9.5.1

TRADITIONAL AND GENERATIVE AI

Adapt existing structures, communication lines, procedures, and rules (e.g., three lines of defence risk management model) or implement new ones

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of adaptation or new implementation of structures, communication lines, procedures, and rules (e.g., three lines of defence risk management model)

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 9.5

Establish clear internal governance mechanisms to ensure clear roles and responsibilities for the use of AI by the organisation

PROCESS 9.5.2

TRADITIONAL AND GENERATIVE AI

For organisations who are using AI across departments, establish teams or AI governance committee that comprises representatives from data science, technology, risk, and product to facilitate cross-departmental oversight for the lifecycle governance of AI systems
If it's not practical to have cross-department oversight, integrate accountability of the use of AI into existing risk or compliance structures
For systems with national security risks, involve relevant stakeholders

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of the establishment of an AI governance committee.

This committee should be sufficiently representative. One way to achieve this is by having representatives from:
- data science;
- technology;
- legal and compliance;
- risk and product; and
- user experience research, ethics, and psychology

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 9.5

Establish clear internal governance mechanisms to ensure clear roles and responsibilities for the use of AI by the organisation

PROCESS 9.5.3

TRADITIONAL AND GENERATIVE AI

Develop and implement policies, procedures, and training to ensure that staff are familiar with their duties and the organisation's risk management practices

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of:
- policies and procedures outlining the organisation's AI policies (e.g. content provenance) and risks management practices, including how regular the policies and responsibilities are being reviewed
- list of training courses and staff attendance for courses

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.5

Establish clear internal governance mechanisms to ensure clear roles and responsibilities for the use of AI by the organisation

PROCESS 9.5.4

TRADITIONAL AND GENERATIVE AI

Enable an incident management process:
- processes and mechanisms to report on actions or decisions that affect the AI system's outcome in compliance with legal and regulatory requirements, and a corresponding process for the accountable party to respond to the consequences of such an outcome
- the above processes and mechanisms to be reviewed and updated at appropriate junctures
- for incident monitoring and reporting, including verifying information and having a minimum set of criteria for reporting such as stakeholders impacted
- to verify teams that conduct incident response demonstrate appropriate skills and training as well as have a diverse composition and responsibilities based on the particular incident type
- to measure the rate at which recommendations from security checks and incidents are implemented, and assess how quickly the AI system can adapt and improve
- that safeguards intellectual property rights

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence that outlines roles, responsibilities, and key processes for

- the reporting on actions or decisions that affect the AI system's outcome
- the corresponding process for the accountable party to respond to the consequences of such an outcome
- monitoring and after action review of incident response and disclosure
- incident response plan and preventive measures

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.6

Define the policy mechanism for enforcing access rights and permissions for the various roles of users

PROCESS 9.6.1

TRADITIONAL AND GENERATIVE AI

Implement fine-grained access control that aligns with various roles for users:
- Access to code and data for training AI models
- Access to code and data for deploying AI models
- Access to different execution environments
- Permission to perform various actions (e.g., launch training job, review model, deploy model server)
- Permission to define access control rules and perform other administrative functions

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of the implementation of fine-grained access control that aligns with various roles for users, which include:

- Access to code and data for training AI models
- Access to code and data for deploying AI models
- Access to different execution environments
- Permission to perform various actions (e.g., launch training job, review model, deploy model server)
- Permission to define access control rules and perform other administrative functions

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.7

Establish clear responsibilities between different parties within the broader supply chain – partners, suppliers, customers, third parties

PROCESS 9.7.1

TRADITIONAL AND GENERATIVE AI

Responsibilities and obligations are clearly communicated to all parties related to the AI system
Implement service level agreements and contracts with third parties that specify requirements, expectation and evaluation e.g. content provenance, usage rights, security requirements
Implement a risk assessment framework to evaluate and monitor third parties' performance (e.g., content provenance standards)

Conduct assessment on third parties or vendors, including AI risks and benefits arising from use of third party models or resources such as:
- application of organisational risk tolerance
- value chain risks
- transparency artifacts (e.g., model and system cards) for third party models

EVIDENCE

External / internal correspondence

Documentary evidence of
- communication with relevant parties on responsibilities and obligations relating to the AI system
- assessment of third parties or vendors. Assessment to have due diligence process for acquisition and procurement vendor assessment, including considerations for data privacy and other risks

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 9.8

Establish the appropriate process or governance-by-design technology to automate or facilitate the AI system's auditability throughout its lifecycle

PROCESS 9.8.1

TRADITIONAL AND GENERATIVE AI

Process or technology should handle:
- Version control of code and model
- Version data or maintain immutable data
- Audit trail of deployment history, log inputs/outputs, associate server predictions with the originating model

EVIDENCE

Internal documentation of physical testing

Documentary evidence of the establishment of the appropriate process or governance-by-design technology to automate or facilitate the AI system's auditability throughout its lifecycle.

The process or technology should handle:
- Version control of code and model;
- Version data or maintain immutable data; and
- Audit trail of deployment history, log inputs/outputs, associate server predictions with the originating model

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.9

If you are using third-party 'black box' models, assess the suitability and limits of the model for your use case

PROCESS 9.9.1

TRADITIONAL AND GENERATIVE AI

Evaluate the necessity of third-party models e.g., they are trained on data otherwise not accessible to your organisation ,or you do not have the requisite capability to build AI systems in-house

EVIDENCE

Internal documentation

Documentary evidence of evaluation completed regarding the necessity of third-party models

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.9

If you are using third-party 'black box' models, assess the suitability and limits of the model for your use case

PROCESS 9.9.2

TRADITIONAL AND GENERATIVE AI

Demonstrate effort to understand how the third-party models were built, including 1) what data was used to train the models, 2) how the models are assessed for effectiveness and explainability 3) under what circumstances does the AI system perform poorly

EVIDENCE

Internal documentation

Documentary evidence of effort undertaken to understand how the third-party models were built, which includes:

- what data was used to train the models;
- how the models are assessed for effectiveness and explainability; and
- under what circumstances does the AI system perform poorly

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.10

Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, unacceptable and unanticipated impacts as well as best practices are in place
Note: “AI Actors” is defined as “those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI”

PROCESS 9.10.1

TRADITIONAL AND GENERATIVE AI

In addition to 7.7.1,
- Put in place a process to engage external AI actors for feedback pre-deployment and post-deployment
- Define use case where feedback exercise would be beneficial
- Ensure regular outreach and feedback exercises are representative and diverse. The exercises can include studies, prototyping and testing activities (e.g. test of AI capabilities, human proficiency tests), how users perceive the output from the AI model, to identify or review unanticipated impacts, or to detect subtle shifts in quality or alignment of AI model output with community and societal values.
- Allocate time and resources for outreach, feedback and recourse processes, including prioritising feedback based on risk assessment
- Verify adequacy of user instructions through using testing
- Develop metrics to evaluate the feedback
- Identify emerging best practices and technologies in measuring and managing identified risks

Where applicable, verify those conducting feedback are not directly involved in system development task for the AI model

EVIDENCE

Internal documentation

Documentary evidence of engagement and feedback from relevant AI actors before and after deployment, as well as incorporation of feedback

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 9.11

Resources required to manage AI risks are taken into account – along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts

PROCESS 9.11.1

TRADITIONAL AND GENERATIVE AI

Conduct impact assessment on the use of AI verses non-AI alternative systems, approaches, or methods, and the resources required to manage the risk of using AI

EVIDENCE

Internal documentation

Documentary evidence of impact assessment

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 9.12

Mechanisms are in place to inventory AI systems including ensuring that they are properly resourced in areas such as (a) data; (b) tooling; (c) system & computing; and (d) in human resources according to organizational risk priorities

PROCESS 9.12.1

TRADITIONAL AND GENERATIVE AI

Put in place guided flow for documenting (i) the inventory of AI systems and necessary resources (e.g., data, tooling, system & computing and human resources), (ii) risk priorities and (iii) inventory exemptions (e.g. GAI systems embedded into application software). This include inventory of all third party entities.
Data include data provenance information such as watermark, signatures, versioning
System include bug tracking or external information sharing resource

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of considerations of resources (e.g., data, tooling, system & computing and human resources) and risk priorities

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

AI VERIFY FOUNDATION

Human Agency & Oversight

Ability to implement appropriate oversight and control measures with humans-in-the-loop at the appropriate juncture.

AI systems can be used to support or influence humans in decision-making processes. AI systems that 'act' like humans also have an effect on human perception, expectation, and functionality. Human agency and oversight ensure that the human has the ability to self-assess and intervene where necessary to ensure that the AI system is used to achieve the intended goals. The human should also have the ability to improve and override the operation of the system when the AI system results in a negative outcome.

OUTCOME 10.1

Establish a strategy for maintaining independent oversight over the development and deployment of AI systems

PROCESS 10.1.1

TRADITIONAL AND GENERATIVE AI

Reviewers should be distinct from those who are training and deploying models. However, it is acceptable to have the same individuals training and deploying models

EVIDENCE

Internal documentation (e.g., log, register or database)

Documentary evidence of strategy for maintaining independent oversight over the development and deployment of AI systems

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 10.2

Ensure that the various parties involved in using, reviewing, and sponsoring the AI system are adequately trained and equipped with the necessary tools and information for proper oversight to:
- Obtain the needed information to conduct inquiries into past decisions made and actions taken throughout the AI lifecycle
- Record information on training and deploying models as part of the workflow process

PROCESS 10.2.1

TRADITIONAL AND GENERATIVE AI

Put in place guided flow for documenting (i) important info via model cards, forms, SDK library; and (ii) important processes that provide objective criteria for decision-making (e.g., fairness metrics selection)

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of guided flow for documenting:

- important info via model cards, forms, SDK library; and
- important processes that provide objective criteria for decision-making (e.g., fairness metrics selection)

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 10.2

Ensure that the various parties involved in using, reviewing, and sponsoring the AI system are adequately trained and equipped with the necessary tools and information for proper oversight to:
- Obtain the needed information to conduct inquiries into past decisions made and actions taken throughout the AI lifecycle
- Record information on training and deploying models as part of the workflow process

PROCESS 10.2.2

TRADITIONAL AND GENERATIVE AI

Implement a data management system to gather and organise relevant information based on the needs of different user roles (e.g., reviewing models, and monitoring live systems)

EVIDENCE

Internal documentation (e.g., procedure manual, log, register, or database)

Documentary evidence of data management system to gather and organise relevant information based on the needs of different user roles

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 10.3

Ensure specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system

PROCESS 10.3.1

TRADITIONAL AND GENERATIVE AI

Define the role of the human in its oversight and control of the AI system (e.g., human-in-the-loop, human-out-the-loop, human-over-the-loop)

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of the definition of the role of human in oversight and control of the AI system. This can include several escalation pathways depending on the criticality of the risks of the application. For example, escalation of high-risk AI systems to Management/ AI Governance Committee for review and approval prior to commercial deployment, and escalation of high-risk control gaps to Management/ AI Governance Committee for deliberation of follow-up actions

Escalation process is also needed when there is any dispute with AI decision/ output and a tracking mechanism can help to log all historical dispute and resolutions

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 10.3

Ensure specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system

PROCESS 10.3.2

TRADITIONAL AND GENERATIVE AI

When the AI model is making a decision for which it is significantly unsure of the answer/prediction, consider designing the system to be able to flag these cases and triage them for a human to review
Monitor and evaluate the instances where human operators or other systems override the GAI's decisions

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of:
- consideration made in the design of the AI system on its ability to flag instances when it is making a decision for which it is significantly unsure of the answer/prediction, in order that such cases be triaged for a human to review
- instances where human operators or other systems override the GAI's decisions

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 10.3

Ensure specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system

PROCESS 10.3.3

TRADITIONAL AND GENERATIVE AI

Implement mechanisms to detect if model input represents an outlier in terms of training data (e.g., return some "data outlier score" with predictions)

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of implementation of mechanisms to detect if model input represents an outlier in terms of training data

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 10.4

Put in place a review process before AI models are put into production, where key features and properties of the AI model are shared and visualised in a way that is accessible to decision-makers within the organisation

PROCESS 10.4.1

TRADITIONAL AND GENERATIVE AI

Implement a systematic review process to present performance, explainability, and fairness metrics in a way that is understandable by data science, product, legal and risk, experience research, and ethics teams

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of the implementation of a systematic review process to present performance, explainability, and fairness metrics in a way that is understandable by relevant teams (e.g., data science, product, legal and risk, experience research, and ethics teams)

PROCESS CHECKS COMPLETED

- ☐ **Yes**
- ☐ **No**
- ☐ **N/A**

ELABORATION

OUTCOME 10.5

Establish a frequency and process for testing and re-evaluating AI systems

PROCESS 10.5.1

TRADITIONAL AND GENERATIVE AI

After models are put into production, put in place mechanisms to review the performance of the models on an ongoing basis, either continuously or at regular intervals.
Criteria could be time-based (e.g., every 2 years) or event-based (before the launch of a new AI product, after the introduction of new data, operating context has changed due to external circumstances), or when the AI system has undergone substantial modification

EVIDENCE

Internal documentation of physical testing

Documentary evidence of the establishment of a frequency and process for testing and re-evaluating AI systems

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

OUTCOME 10.6

Ensure the appropriate parties who are accountable for the AI system (e.g., AI governance committee, AI system owner, and reviewers) have considered how the AI system is used to benefit humans in decision-making processes

PROCESS 10.6.1

TRADITIONAL AND GENERATIVE AI

Declaration of transparency on how and where in the decision-making process the AI system is used to complement or replace the human

EVIDENCE

1) Internal documentation (e.g., procedure manual) 2) External / internal correspondence

Documentary evidence of the declaration of transparency on how and where in the decision-making process the AI system is used to complement or replace the human

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

AI VERIFY FOUNDATION

Inclusive Growth, Societal, & Environmental Well-being

This Principle highlights the potential for trustworthy AI to contribute to overall growth and prosperity for all – individuals, society, and the planet – and advance global development objectives

Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender, and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development, and well-being.

OUTCOME 11.1

Ensure that the development of AI system is for the beneficial outcomes for the environment

PROCESS 11.1.1

TRADITIONAL AND GENERATIVE AI

Put in place a process to determine that the development and deployment of the AI system is for the benefit of the environment, where applicable.

For example:

- conduct impact assessment on the natural environment, and take measures to minimise negative impact (e.g., pollution)
- address environmental concerns (e.g., energy and water consumption, carbon capture and green washing concerns)
- monitor energy and waste consumptions

The above could be carried out through interviews with relevant stakeholders and impact studies

EVIDENCE

Internal documentation (e.g., procedure manual) / External correspondence

Documentary evidence of consideration and measurement of AI system's impact on environment, which may include (where applicable):

- Energy, waste and water consumption
- Carbon capture
- Green washing concerns

PROCESS CHECKS COMPLETED

- [ ] Yes
- [ ] No
- [ ] N/A

ELABORATION

OUTCOME 11.2

Ensure that the development of AI system is for the beneficial outcomes for individuals and society

PROCESS 11.2.1

TRADITIONAL AND GENERATIVE AI

Put in place a process to determine that the development and deployment of the AI system is for the benefit of people and society, where applicable. For example, conduct impact assessment on societal implications such as life, health, work and skills and take measures to minimise negative impact

The above could be carried out through interviews with relevant stakeholders and impact studies

EVIDENCE

Internal documentation (e.g., procedure manual)

Documentary evidence of consideration of AI system's impact on individuals and society, which may include (where applicable):
- Human capabilities to learn and make decisions
- Skills, jobs, and/or job quality
- Creative economies
- Discriminatory and/or exclusionary norms

PROCESS CHECKS COMPLETED

- [ ] **Yes**
- [ ] **No**
- [ ] **N/A**

ELABORATION

BROUGHT TO YOU BY

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

AI VERIFY FOUNDATION