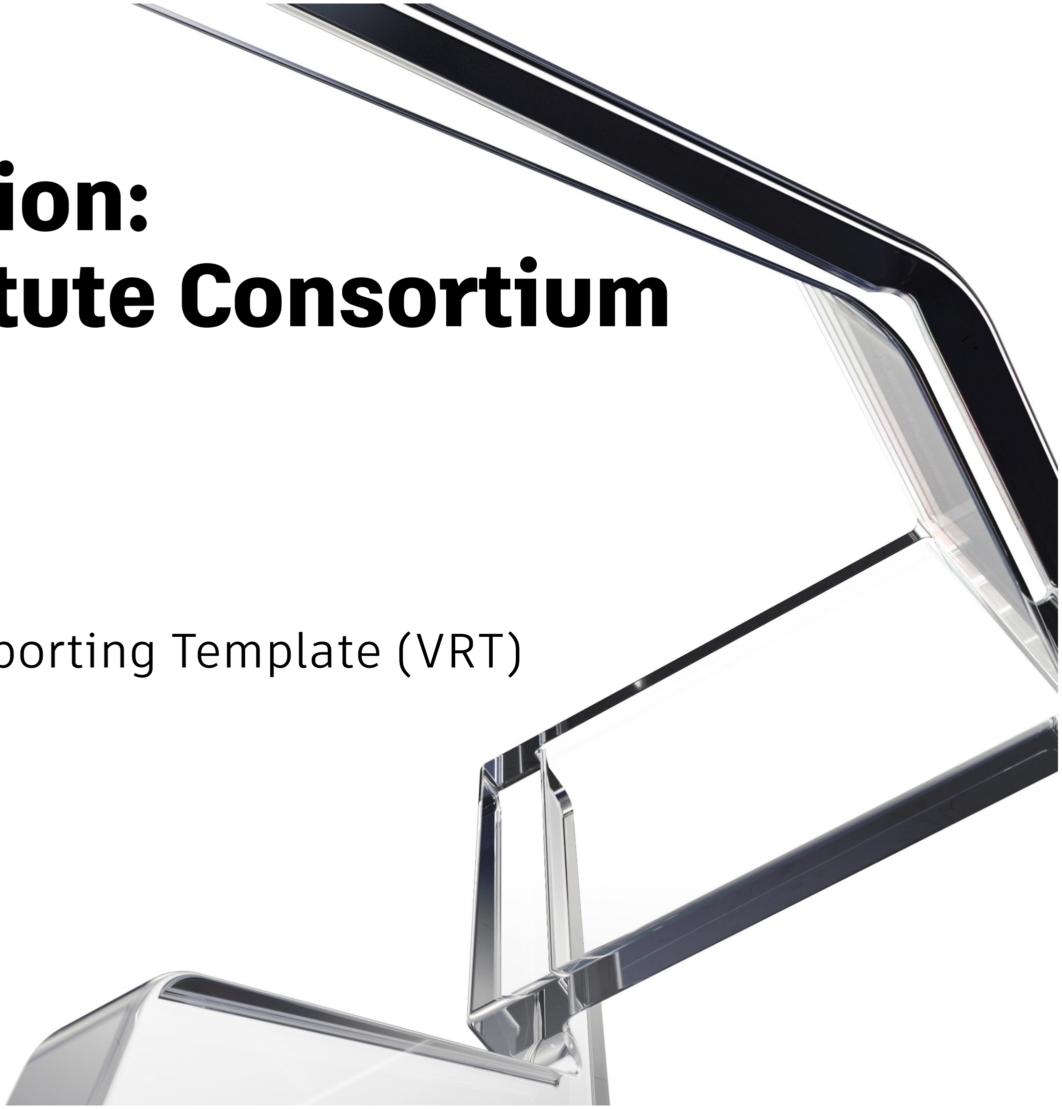
AUTODESK

# Autodesk Contribution: NIST AI Safety Institute Consortium Working Group

**Autodesk - Confidential**

WG 1.3 GenAI Profile – Voluntary Reporting Template (VRT)



Ryan Broulette
Trusted AI Consultant

# CONTENT DISCLAIMER

*The content in this document is intended for the NIST AISIC Working Group 1.3.*

*This does not represent the actual documentation and/or processes performed by Autodesk.*

*This document is conceptual in nature and intended to illustrate the ways which the stated objectives can be met (a voluntary reporting template could be created).*

***Any attribution to Autodesk in AISIC's public materials requires prior approval from Autodesk.***

# 1. Objective

Overall objective:

Provide NIST's Working Group 1.3 with considerations to inform how voluntary reporting templates of organizations' GenAI profiles can be designed, and, how organizations can operationalize processes to complete them.

This is depicted through illustrative examples of potential content for a subset of subcategories, organized in two areas:

1. Considerations for VRT format and elements with detail.
2. Considerations for evaluating select NIST Subcategories (from AI RMF) to complete a template and Gen AI Profile.

# 1. VRT Reporting Format

Recommendations to consider when developing the format to enable the completion of the GenAI profile.

**Level and scope of the report template (Govern 1.3)**: Consider a dynamic template to capture details relevant to either the organization as a whole or their AI use cases. Each level may be informed by distinct/separate risk management processes that are appropriate for their organization:

- *Company level:* Activities to evaluate whether processes are adequately designed to effectively map, measure, and manage AI development and use across the organization.
- *AI use case level (product/service):* Specific activities to obtain relevant information about the AI use and whether the risks specific to the AI use are adequately managed. This will also support whether the governance level activities are operating effectively.
- Organizations should use the risk management mechanisms as an enabler to create a GenAI profile and demonstrate responsible development and use.

**Scope, roles and nature of the AI (Map 1, Govern 2)**:

- Enable the template to be customized based on the organization's role, and, third parties and their role, and industry specific.
- Enable the ability to define the profile based on the applicable scope of the AI System and its components, including models, data, uses, actors, technology, and align with potential impacts of each.
- Enable the template to be customized (dynamic) based on the orientation of the AI (informed by the AI actor role), focused on (a) whether the AI is developed internally or acquired externally, (b) the data type, and ( c) whether the intended users are internal employees or external customers

# 1. VRT Reporting Elements

## Generative AI Profile – Elements for consideration in GenAI Profile

*General (MAP 1.1):*

- Name of product/service with AI
- AI description (type)
- AI techniques and tasks performed (Map 2.1)
- Intended purpose of use (Map 1.1)
- Intended users of the AI
- Data type and classification
- Assumptions
- Limitations
- AI system components in scope

*Assessment status: (Completed / Not completed)*

- Compliance with laws/regulations (GOVERN 1.1)
- Compliance with internal requirements
- Security
- Privacy
- Resource management
- Ethics
- Responsible AI Development
- Responsible AI Use

*Accountability status (GOVERN 2.1):*

Designation of accountability:

- AI datasets
- AI models
- AI performance and metrics
- AI testing and training
- AI impacts to users, organizations, society
- Human involvement
- Third party relationship owners

*Transparency status:*

- Inventory of technology and tools (Govern 1.6)
- Explainable and easily understood communications about the AI and decisions
- User notification when they use AI
- Data outputs meet intended purposes and are explainable.
- Mechanisms are implemented to receive and act upon feedback from various AI actors. (GOVERN 5.1)

*Other status of GenAI Profile actions by characteristic*

# 1. VRT Reporting Elements

## Risk and impact – identify, assess

Inherent Risks of AI systems, based on the context, each component, and the status of actions:

- Model risk
- Data risk
- AI use risk
- Security risks
- Privacy risks
- Resilience risks
- Governance risks
- Legal & Compliance risks
- Environmental Risks

Evaluate processes and controls in areas:

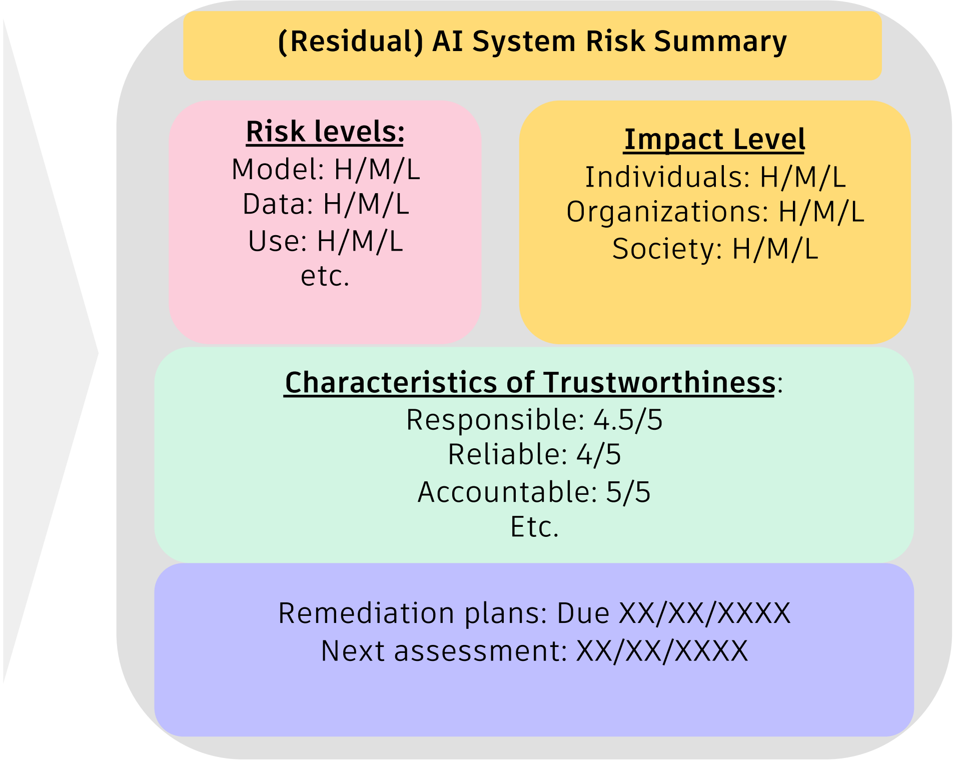
- Data:
    - Acquisition
    - Preparation
    - Labelling
    - Data quality
    - Training and testing
    - Deployment
    - Operation
    - Retention and decommissioning
- Model:
    - Design
    - Development – training, testing
    - Deployment
    - Operation and Monitoring
- AI Use
    - Performance
    - Incidents
    - Outputs

# 1. VRT Reporting Elements

## Risk and impact – respond, monitor

Measuring mitigations based on defined rating scales (examples)

- Safety, fairness, bias
    - Human involvement in decisions
    - Managed bias in data selection and fit
    - Managed bias in rules
    - Managed bias of output
    - Testing adequacy
    - Robustness of data
- Transparency
    - Explainable comms
    - Etc.
- Accountable
- Responsible
- Etc.



**(Residual) AI System Risk Summary**

**Risk levels:**
Model: H/M/L
Data: H/M/L
Use: H/M/L
etc.

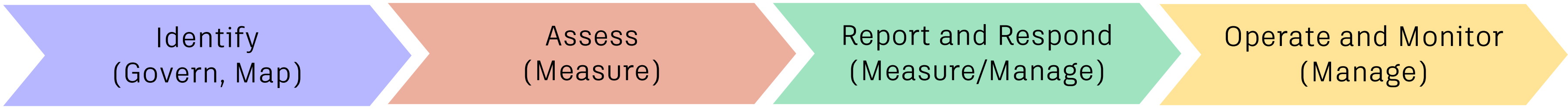
**Impact Level**
Individuals: H/M/L
Organizations: H/M/L
Society: H/M/L

**Characteristics of Trustworthiness**:
Responsible: 4.5/5
Reliable: 4/5
Accountable: 5/5
Etc.

Remediation plans: Due XX/XX/XXXX
Next assessment: XX/XX/XXXX

# 2. Evaluation and creation of the the GenAI Profile

**Process overview:** This describes the workflow with tasks to consider when developing the GenAI profile and performing assessments, which are aligned to various NIST subcategories. Following this process will enable the ability to complete the profile and track implementation.



Identify (Govern, Map)

Assess (Measure)

Report and Respond (Measure/Manage)

Operate and Monitor (Manage)

**What:** Determine the scope of the AI System and identify the potential risks and impacts to individuals, organizations, and communities.

**How:**

- Register AI use, data, and model details
- Inquiries, observations, and reviews of related artifacts.
- Use the AI risk taxonomy

**What:** Assess the level of control for whether the measures reasonably mitigate risks and minimize impacts.

**How:**

- Review evidence of controls and how the impacts are minimized according to the risk methodology.
- Impacts to individuals, organizations, and society are measured.

**What:** Validate and communicate the results and recommendations. Establish responses and action plans to treat risks.

**How:**

- Communicate results and recommendations to responsible owners.
- Document risk treatment

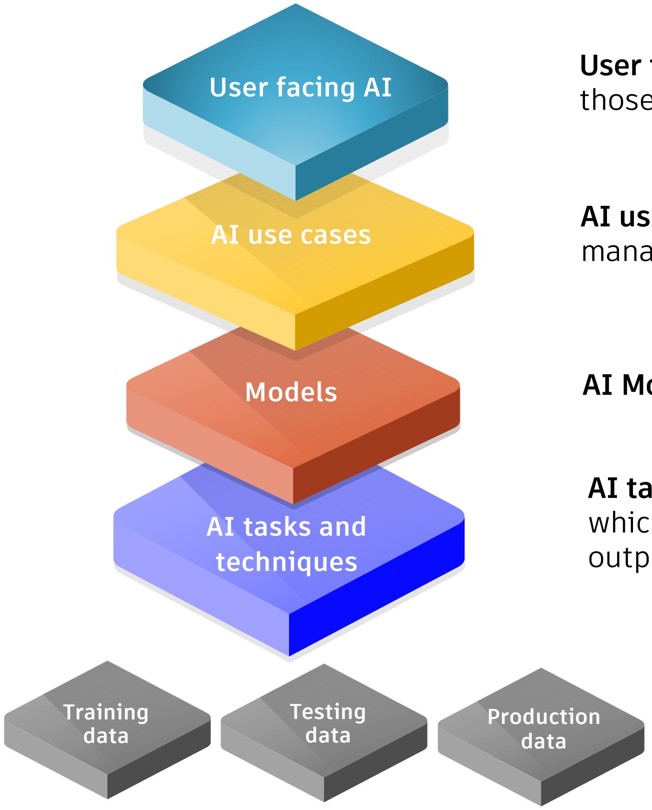
**What:** Perform action plans, monitor for remediation, conduct periodic reviews.

**How:**

- Risk treatment and action plans are managed and reported Leadership through governance channels and committees.
- Continuous monitoring is performed
- Future assessment are scheduled

# Further detail of “Identify:” Scoping the assessment

The illustration depicts considerations to make when scoping an AI system assessment



**User facing AI:** Identify who uses the AI and how, including the products, users, and those that may be impacted.

**AI use cases:** Capture the details of how the AI will be used and how the organization is managing relevant projects and initiatives.

**AI Models:** Identify relevant attributes of the models, integrations, and impacts

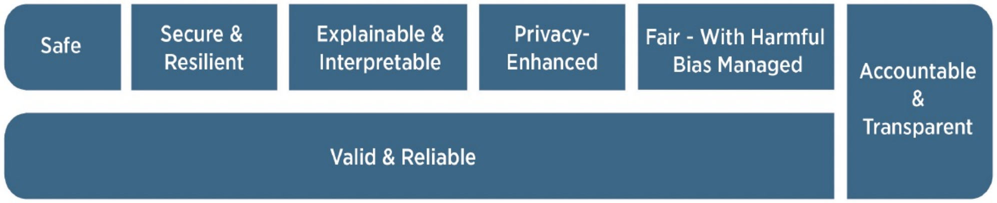
**AI tasks and techniques:** Define what AI/ML techniques are being performed, which tasks it is completing (i.e. classify, recommend, etc.) how it may impact the outputs and users.

**Datasets:** Identify and evaluate all data inputs and outputs through formal categorization and classification.

# 2. Evaluating NIST’s Characteristics of Trustworthy to create a profile

**Step 1:** Define and communicate the characteristics of trustworthiness as principles in policies/processes.

*NIST’s Characteristics of Trustworthy AI Systems*


Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable

**Step 2**: Define objectives for the responsible development and use at each phase of the AI system lifecycle that align with the characteristics of trustworthiness (principles) and implement processes to ensure objectives are met. Metrics should be measurable and tested.

**Step 3:** Evaluate the adequacy of the metrics and results of testing, for mitigating the risks and impacts and alignment with the principles (characteristics of trustworthiness).

- NIST 600.1 Risks
- Data risks
- Model risks
- Risks of use of AI
- Governance, legal, compliance risks

## 2. Evaluating NIST’s Characteristics of Trustworthy to create a profile

*Examples of potential considerations for evaluating the NIST Subcategories of Gen AI Profile, which are operationalized through governance and assessment mechanisms. (Note: This is not an exhaustive list and is only meant as select examples*

| NIST AI RMF Subcategories | Evaluation guidance |
|---|---|
| • GOVERN 1.2: The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices. | • Determine to what extent the assessment process is designed to assess characteristics of trustworthiness in each phase of the AI system lifecycle. (See process previous slide) |
| • MAP 1: Context is established and understood.<br>• MAP 5: Impacts to individuals, groups, communities, organizations, and society are characterized. | • Determine whether the context of use and those impacted have been identified and can be organized consistently according to defined taxonomies.<br>• Use qualitative/quantitative measurement scales to assess the levels of impact from the AI. |
| • MEASURE 1: Appropriate methods and metric are identified and applied.<br>• MEASURE 2: AI systems are evaluated for trustworthy characteristics. | • Review whether risk owners have defined metrics and completed testing of the AI for each characteristic of trustworthiness, for the applicable data, model, and other system components in scope of the AI.<br>• The conclusions could be measured following a risk-based approach, with both quantifiable and qualifiable scales. |
| • MEASURE 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.<br>• MANAGE 1.1: A determination is made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed. | • The context of the AI use and the information about the model, data, technology, use and other system components, could be part of the evaluation, as determined in the scope.<br>• Results are validated with experts of the product/service, such as whether the actual outputs are within defined thresholds of acceptable outputs and meet its intended purposes.<br>• Acceptance criteria for deployment is defined and validated to be met/not met and for meeting applicable characteristics. If certain characteristics of responsible use are not met, the risk and impact levels |

AUTODESK