AI RMF PLAYBOOK
# TABLE OF CONTENTS

**GOVERN**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Govern 1</td>
<td>1</td>
</tr>
<tr>
<td>Govern 1.1</td>
<td>2</td>
</tr>
<tr>
<td>Govern 1.2</td>
<td>2</td>
</tr>
<tr>
<td>Govern 1.3</td>
<td>4</td>
</tr>
<tr>
<td>Govern 1.4</td>
<td>7</td>
</tr>
<tr>
<td>Govern 1.5</td>
<td>9</td>
</tr>
<tr>
<td>Govern 1.6</td>
<td>12</td>
</tr>
<tr>
<td>Govern 1.7</td>
<td>14</td>
</tr>
<tr>
<td>Govern 2</td>
<td>16</td>
</tr>
<tr>
<td>Govern 2.1</td>
<td>18</td>
</tr>
<tr>
<td>Govern 2.2</td>
<td>18</td>
</tr>
<tr>
<td>Govern 2.3</td>
<td>21</td>
</tr>
<tr>
<td>Govern 3</td>
<td>23</td>
</tr>
<tr>
<td>Govern 3.1</td>
<td>25</td>
</tr>
<tr>
<td>Govern 3.2</td>
<td>25</td>
</tr>
<tr>
<td>Govern 4</td>
<td>28</td>
</tr>
<tr>
<td>Govern 4.1</td>
<td>31</td>
</tr>
<tr>
<td>Govern 4.2</td>
<td>31</td>
</tr>
<tr>
<td>Govern 4.3</td>
<td>34</td>
</tr>
<tr>
<td>Govern 5</td>
<td>37</td>
</tr>
<tr>
<td>Govern 5.1</td>
<td>39</td>
</tr>
<tr>
<td>Govern 5.2</td>
<td>39</td>
</tr>
<tr>
<td>Govern 6</td>
<td>42</td>
</tr>
<tr>
<td>Govern 6.1</td>
<td>44</td>
</tr>
<tr>
<td>Govern 6.2</td>
<td>44</td>
</tr>
<tr>
<td>MAP</td>
<td>47</td>
</tr>
<tr>
<td>Map 1</td>
<td>49</td>
</tr>
<tr>
<td>Map 1.1</td>
<td>50</td>
</tr>
<tr>
<td>Map 1.2</td>
<td>50</td>
</tr>
<tr>
<td>Map 1.3</td>
<td>56</td>
</tr>
<tr>
<td>Map 1.4</td>
<td>59</td>
</tr>
<tr>
<td>Map 1.5</td>
<td>61</td>
</tr>
<tr>
<td>Map 1.6</td>
<td>63</td>
</tr>
<tr>
<td>Map 1.7</td>
<td>66</td>
</tr>
<tr>
<td>Map 2</td>
<td>69</td>
</tr>
<tr>
<td>Map 2.1</td>
<td>69</td>
</tr>
<tr>
<td>Map 2.2</td>
<td>71</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>Manage 1.2</td>
<td>176</td>
</tr>
<tr>
<td>Manage 1.3</td>
<td>178</td>
</tr>
<tr>
<td>Manage 1.4</td>
<td>180</td>
</tr>
<tr>
<td>Manage 2</td>
<td>182</td>
</tr>
<tr>
<td>Manage 2.1</td>
<td>182</td>
</tr>
<tr>
<td>Manage 2.2</td>
<td>184</td>
</tr>
<tr>
<td>Manage 2.3</td>
<td>190</td>
</tr>
<tr>
<td>Manage 2.4</td>
<td>192</td>
</tr>
<tr>
<td>Manage 3</td>
<td>195</td>
</tr>
<tr>
<td>Manage 3.1</td>
<td>195</td>
</tr>
<tr>
<td>Manage 3.2</td>
<td>198</td>
</tr>
<tr>
<td>Manage 4</td>
<td>200</td>
</tr>
<tr>
<td>Manage 4.1</td>
<td>200</td>
</tr>
<tr>
<td>Manage 4.2</td>
<td>202</td>
</tr>
<tr>
<td>Manage 4.3</td>
<td>204</td>
</tr>
</tbody>
</table>
FORWARD

The Playbook provides suggested actions for achieving the outcomes laid out in the AI Risk Management Framework (AI RMF) Core (Tables 1 – 4 in AI RMF 1.0). Suggestions are aligned to each sub-category within the four AI RMF functions (Govern, Map, Measure, Manage).

The Playbook is neither a checklist nor set of steps to be followed in its entirety.

Playbook suggestions are voluntary. Organizations may utilize this information by borrowing as many – or as few – suggestions as apply to their industry use case or interests.
GOVERN

Govern
A culture of risk management is cultivated and present
GOVERN 1
Policies, processes, procedures and practices across the organization related to the mapping, measuring and managing of AI risks are in place, transparent, and implemented effectively.

GOVERN 1.1
Legal and regulatory requirements involving AI are understood, managed, and documented.

About
AI systems may be subject to specific applicable legal and regulatory requirements. Some legal requirements can mandate (e.g., nondiscrimination, data privacy and security controls) documentation, disclosure, and increased AI system transparency. These requirements are complex and may not be applicable or differ across applications and contexts.

For example, AI system testing processes for bias measurement, such as disparate impact, are not applied uniformly within the legal context. Disparate impact is broadly defined as a facially neutral policy or practice that disproportionately harms a group based on a protected trait. Notably, some modeling algorithms or debiasing techniques that rely on demographic information, could also come into tension with legal prohibitions on disparate treatment (i.e., intentional discrimination).

Additionally, some intended users of AI systems may not have consistent or reliable access to fundamental internet technologies (a phenomenon widely described as the “digital divide”) or may experience difficulties interacting with AI systems due to disabilities or impairments. Such factors may mean different communities experience bias or other negative impacts when trying to access AI systems. Failure to address such design issues may pose legal risks, for example in employment related activities affecting persons with disabilities.

Suggested Actions
- Maintain awareness of the applicable legal and regulatory considerations and requirements specific to industry, sector, and business purpose, as well as the application context of the deployed AI system.
- Align risk management efforts with applicable legal standards.
- Maintain policies for training (and re-training) organizational staff about necessary legal or regulatory considerations that may impact AI-related design, development and deployment activities.

**Transparency & Documentation**

*Organizations can document the following:*

- To what extent has the entity defined and documented the regulatory environment—including minimum requirements in laws and regulations?
- Has the system been reviewed for its compliance to applicable laws, regulations, standards, and guidance?
- To what extent has the entity defined and documented the regulatory environment—including applicable requirements in laws and regulations?
- Has the system been reviewed for its compliance to relevant applicable laws, regulations, standards, and guidance?

**AI Transparency Resources**

*GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.*

**References**

*Andrew Smith, "Using Artificial Intelligence and Algorithms," FTC Business Blog (2020).*


*Patrick Hall, Benjamin Cox, Steven Dickerson, Arjun Ravi Kannan, Raghu Kulkarni, and Nicholas Schmidt, "A United States fair lending perspective on machine learning," Frontiers in Artificial Intelligence 4 (2021).*

*AI Hiring Tools and the Law, Partnership on Employment & Accessible Technology (PEAT, peatworks.org).*
GOVERN 1.2
The characteristics of trustworthy AI are integrated into organizational policies, processes, and procedures.

About
Policies, processes, and procedures are central components of effective AI risk management and fundamental to individual and organizational accountability. All stakeholders benefit from policies, processes, and procedures which require preventing harm by design and default.

Organizational policies and procedures will vary based on available resources and risk profiles but can help systematize AI actor roles and responsibilities throughout the AI lifecycle. Without such policies, risk management can be subjective across the organization, and exacerbate rather than minimize risks over time. Policies, or summaries thereof, are understandable to relevant AI actors. Policies reflect an understanding of the underlying metrics, measurements, and tests that are necessary to support policy and AI system design, development, deployment and use.

Lack of clear information about responsibilities and chains of command will limit the effectiveness of risk management.

Suggested Actions
Organizational AI risk management policies should be designed to:
- Define key terms and concepts related to AI systems and the scope of their purposes and intended uses.
- Connect AI governance to existing organizational governance and risk controls.
- Align to broader data governance policies and practices, particularly the use of sensitive or otherwise risky data.
- Detail standards for experimental design, data quality, and model training.
- Outline and document risk mapping and measurement processes and standards.
• Detail model testing and validation processes.
• Detail review processes for legal and risk functions.
• Establish the frequency of and detail for monitoring, auditing and review processes.
• Outline change management requirements.
• Outline processes for internal and external stakeholder engagement.
• Establish whistleblower policies to facilitate reporting of serious AI system concerns.
• Detail and test incident response plans.
• Verify that formal AI risk management policies align to existing legal standards, and industry best practices and norms.
• Establish AI risk management policies that broadly align to AI system trustworthy characteristics.
• Verify that formal AI risk management policies include currently deployed and third-party AI systems.

**Transparency & Documentation**

Organizations can document the following:

- To what extent do these policies foster public trust and confidence in the use of the AI system?
- What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles?
- What policies and documentation has the entity developed to encourage the use of its AI system as intended?
- To what extent are the model outputs consistent with the entity’s values and principles to foster public trust and equity?

**AI Transparency Resources**

[GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities](#).
References


GAO, “Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities,” GAO@100 (GAO-21-519SP), June 2021.

NIST, "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools".


SAS Institute, “The SAS® Data Governance Framework: A Blueprint for Success”. URL


“Play 5: Create a formal policy,” Partnership on Employment & Accessible Technology (PEAT, peatworks.org).


GOVERN 1.3
Processes and procedures are in place to determine the needed level of risk management activities based on the organization's risk tolerance.

About
Risk management resources are finite in any organization. Adequate AI governance policies delineate the mapping, measurement, and prioritization of risks to allocate resources toward the most material issues for an AI system to ensure effective risk management. Policies may specify systematic processes for assigning mapped and measured risks to standardized risk scales.

AI risk tolerances range from negligible to critical – from, respectively, almost no risk to risks that can result in irredeemable human, reputational, financial, or environmental losses. Risk tolerance rating policies consider different sources of risk, (e.g., financial, operational, safety and wellbeing, business, reputational, or model risks). A typical risk measurement approach entails the multiplication, or qualitative combination, of measured or estimated impact and likelihood of impacts into a risk score \( \text{risk} \approx \text{impact} \times \text{likelihood} \). This score is then placed on a risk scale. Scales for risk may be qualitative, such as red-amber-green (RAG), or may entail simulations or econometric approaches. Impact assessments are a common tool for understanding the severity of mapped risks. In the most fulsome AI risk management approaches, all models are assigned to a risk level.

Suggested Actions
- Establish policies to define mechanisms for measuring or understanding an AI system’s potential impacts, e.g., via regular impact assessments at key stages in the AI lifecycle, connected to system impacts and frequency of system updates.
- Establish policies to define mechanisms for measuring or understanding the likelihood of an AI system’s impacts and their magnitude at key stages in the AI lifecycle.
- Establish policies that define assessment scales for measuring potential AI system impact. Scales may be qualitative, such as red-amber-green (RAG), or may entail simulations or econometric approaches.
- Establish policies for assigning an overall risk measurement approach for an AI system, or its important components, e.g., via multiplication or combination of a mapped risk’s impact and likelihood (risk \( \approx \) impact \( \times \) likelihood).
- Establish policies to assign systems to uniform risk scales that are valid across the organization’s AI portfolio (e.g., documentation templates) and acknowledge risk tolerance and risk levels may change over the lifecycle of an AI system.

**Transparency & Documentation**

Organizations can document the following:
- How do system performance metrics inform risk tolerance decisions?
- What policies has the entity developed to ensure the use of the AI system is consistent with organizational risk tolerance?
- How do the entity’s data security and privacy assessments inform risk tolerance decisions?

**AI Transparency Resources**

[GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.](URL)

**References**


GAO-17-63: Enterprise Risk Management: Selected Agencies’ Experiences Illustrate Good Practices in Managing Risk. URL
GOVERN 1.4
The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.

About
Clear policies and procedures relating to documentation and transparency facilitate and enhance efforts to communicate roles and responsibilities for the Map, Measure and Manage functions across the AI lifecycle. Standardized documentation can help organizations systematically integrate AI risk management processes and enhance accountability efforts. For example, by adding their contact information to a work product document, AI actors can improve communication, increase ownership of work products, and potentially enhance consideration of product quality. Documentation may generate downstream benefits related to improved system replicability and robustness. Proper documentation storage and access procedures allow for quick retrieval of critical information during a negative incident. Explainable machine learning efforts (models and explanatory methods) may bolster technical documentation practices by introducing additional information for review and interpretation by AI Actors.

Suggested Actions
- Establish and regularly review documentation policies that, among others, address information related to:
  - AI actors contact information
  - Business justification
  - Scope and usages
  - Expected and potential risks and impacts
  - Assumptions and limitations
  - Description and characterization of training data
  - Algorithmic methodology
  - Evaluated alternative approaches
  - Description of output data
  - Testing and validation results (including explanatory visualizations and information)
  - Down- and up-stream dependencies
  - Plans for deployment, monitoring, and change management
  - Stakeholder engagement plans
Verify documentation policies for AI systems are standardized across the organization and remain current.
Establish policies for a model documentation inventory system and regularly review its completeness, usability, and efficacy.
Establish mechanisms to regularly review the efficacy of risk management processes.
Identify AI actors responsible for evaluating efficacy of risk management processes and approaches, and for course-correction based on results.
Establish policies and processes regarding public disclosure of the use of AI and risk management material such as impact assessments, audits, model documentation and validation and testing results.
Document and review the use and efficacy of different types of transparency tools and follow industry standards at the time a model is in use.

Transparency & Documentation
Organizations can document the following:

- To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?
- What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed? How much distributional shift or model drift from baseline performance is acceptable?

AI Transparency Resources
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
References


Christoph Molnar, Interpretable Machine Learning, lulu.com.


GOVERN 1.5
Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, organizational roles and responsibilities are clearly defined, including determining the frequency of periodic review.

About
AI systems are dynamic and may perform in unexpected ways once deployed or after deployment. Continuous monitoring is a risk management process for tracking unexpected issues and performance changes, in real-time or at a specific frequency, across the AI system lifecycle.

Incident response and “appeal and override” are commonly used processes in information technology management. These processes enable real-time flagging of potential incidents, and human adjudication of system outcomes.

Establishing and maintaining incident response plans can reduce the likelihood of additive impacts during an AI incident. Smaller organizations which may not have fulsome governance programs, can utilize incident response plans for addressing system failures, abuse or misuse.

Suggested Actions
- Establish policies to allocate appropriate resources and capacity for assessing impacts of AI systems on individuals, communities and society.
- Establish policies and procedures for monitoring and addressing AI system performance and trustworthiness, including bias and security problems, across the lifecycle of the system.
- Establish policies for AI system incident response or confirm that existing incident response policies apply to AI systems.
- Establish policies to define organizational functions and personnel responsible for AI system monitoring and incident response activities.
- Establish mechanisms to enable the sharing of feedback from impacted individuals or communities about negative impacts from AI systems.
- Establish mechanisms to provide recourse for impacted individuals or communities to contest problematic AI system outcomes.
- Establish opt-out mechanisms.
Transparency & Documentation

Organizations can document the following:

- To what extent does the system/entity consistently measure progress towards stated goals and objectives?
- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?
- Did your organization address usability problems and test whether user interfaces served their intended purposes?

AI Transparency Resources


References


GOVERN 1.6
Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.

About
An AI system inventory is an organized database of artifacts relating to an AI system or model. It may include system documentation, incident response plans, data dictionaries, links to implementation software or source code, names and contact information for relevant AI actors, or other information that may be helpful for model or system maintenance and incident response purposes. AI system inventories also enable a holistic view of organizational AI assets. A serviceable AI system inventory may allow for the quick resolution of:
- specific queries for single models, such as “when was this model last refreshed?”
- high-level queries across all models, such as, “how many models are currently deployed within our organization?” or “how many users are impacted by our models?”

AI system inventories are a common element of traditional model risk management approaches and can provide technical, business and risk management benefits. Typically inventories capture all organizational models or systems, as partial inventories may not provide the value of a full inventory.

Suggested Actions
- Establish policies that define the creation and maintenance of AI system inventories.
- Establish policies that define a specific individual or team that is responsible for maintaining the inventory.
- Establish policies that define which models or systems are inventoried, with preference to inventorying all models or systems, or minimally, to high-risk models or systems, or systems deployed in high-stakes settings.
- Establish policies that define model or system attributes to be inventoried, e.g., documentation, links to source code, incident response plans, data dictionaries, AI actor contact information.
Transparency & Documentation

Organizations can document the following:

- Who is responsible for documenting and maintaining the AI system inventory details?
- What processes exist for data generation, acquisition/collection, ingestion, staging/storage, transformations, security, maintenance, and dissemination?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?

AI Transparency Resources


References

GOVERN 1.7
Processes and procedures are in place for decommissioning and phasing out of AI systems safely and in a manner that does not increase risks or decrease the organization’s trustworthiness.

About
Irregular or indiscriminate termination or deletion of models or AI systems may be inappropriate and increase organizational risk. For example, AI systems may be subject to regulatory requirements or implicated in future security or legal investigations. To maintain trust, organizations may consider establishing policies and processes for the systematic and deliberate decommissioning of AI systems. Typically, such policies consider user and community concerns, risks in dependent and linked systems, and security, legal or regulatory concerns. Decommissioned models or systems may be stored in a model inventory along with active models, for an established length of time.

Suggested Actions
- Establish policies for decommissioning AI systems. Such policies typically address:
  - User and community concerns, and reputational risks.
  - Business continuity and financial risks.
  - Up and downstream system dependencies.
  - Regulatory requirements (e.g., data retention).
  - Potential future legal, regulatory, security or forensic investigations.
  - Migration to the replacement system, if appropriate.
- Establish policies that delineate where and for how long decommissioned systems, models and related artifacts are stored.
- Establish policies that address ancillary data or artifacts that must be preserved for fulsome understanding or execution of the decommissioned AI system, e.g., predictions, explanations, intermediate input feature representations, usernames and passwords, etc.
**Transparency & Documentation**

**Organizations can document the following:**

- What processes exist for data generation, acquisition/collection, ingestion, staging/storage, transformations, security, maintenance, and dissemination?
- To what extent do these policies foster public trust and confidence in the use of the AI system?
- If anyone believes that the AI no longer meets this ethical framework, who will be responsible for receiving the concern and as appropriate investigating and remediating the issue? Do they have authority to modify, limit, or stop the use of the AI?
- If it relates to people, were there any ethical review applications/reviews/approvals? (e.g., Institutional Review Board applications)

**AI Transparency Resources**

- [Intel.gov: AI Ethics Framework for Intelligence Community - 2020.](https://www.intel.gov/)
- [Datasheets for Datasets.](https://www.datascienceinsider.com)

**References**


GOVERN 2
Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.

GOVERN 2.1
Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.

About
The development of a risk-aware organizational culture starts with defining responsibilities. For example, under some risk management structures, professionals carrying out test and evaluation tasks are independent from AI system developers and report through risk management functions or directly to executives. This kind of structure may help counter implicit biases such as groupthink or sunk cost fallacy and bolster risk management functions, so efforts are not easily bypassed or ignored.

Instilling a culture where AI system design and implementation decisions can be questioned and course-corrected by empowered AI actors can enhance organizations’ abilities to anticipate and effectively manage risks before they become ingrained.

Suggested Actions
- Establish policies that define the AI risk management roles and responsibilities for positions directly and indirectly related to AI systems, including, but not limited to - Boards of directors or advisory committees - Senior management - AI audit functions - Product management - Project management - AI design - AI development - Human-AI interaction - AI testing and evaluation - AI acquisition and procurement - Impact assessment functions - Oversight functions
- Establish policies that promote regular communication among AI actors participating in AI risk management efforts.
• Establish policies that separate management of AI system development functions from AI system testing functions, to enable independent course-correction of AI systems.
• Establish policies to identify, increase the transparency of, and prevent conflicts of interest in AI risk management efforts.
• Establish policies to counteract confirmation bias and market incentives that may hinder AI risk management efforts.
• Establish policies that incentivize AI actors to collaborate with existing legal, oversight, compliance, or enterprise risk functions in their AI risk management activities.

### Transparency & Documentation

**Organizations can document the following:**

• To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?
• Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?
• Are the responsibilities of the personnel involved in the various AI governance processes clearly defined?
• What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
• Did your organization implement accountability-based practices in data management and protection (e.g. the PDPA and OECD Privacy Principles)?

**AI Transparency Resources**

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities
References


GOVERN 2.2
The organization’s personnel and partners receive AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements.

About
To enhance AI risk management adoption and effectiveness, organizations are encouraged to identify and integrate appropriate training curricula into enterprise learning requirements. Through regular training, AI actors can maintain awareness of:
  - AI risk management goals and their role in achieving them.
  - Organizational policies, applicable laws and regulations, and industry best practices and norms.

See MAP 3.4 and 3.5 for additional relevant information.

Suggested Actions
- Establish policies for personnel addressing ongoing education about:
  - Applicable laws and regulations for AI systems.
  - Potential negative impacts that may arise from AI systems.
  - Organizational AI policies.
  - Trustworthy AI characteristics.
- Ensure that trainings are suitable across AI actor sub-groups - for AI actors carrying out technical tasks (e.g., developers, operators, etc.) as compared to AI actors in oversight roles (e.g., legal, compliance, audit, etc.).
- Ensure that trainings comprehensively address technical and socio-technical aspects of AI risk management.
- Verify that organizational AI policies include mechanisms for internal AI personnel to acknowledge and commit to their roles and responsibilities.
- Verify that organizational policies address change management and include mechanisms to communicate and acknowledge substantial AI system changes.
- Define paths along internal and external chains of accountability to escalate risk concerns.
Transparency & Documentation

Organizations can document the following:

- Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?
- How does the entity determine the necessary skills and experience needed to design, develop, deploy, assess, and monitor the AI system?
- How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?
- What efforts has the entity undertaken to recruit, develop, and retain a workforce with backgrounds, experience, and perspectives that reflect the community impacted by the AI system?

AI Transparency Resources

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities

References


“Developing Staff Trainings for Equitable AI,” Partnership on Employment & Accessible Technology (PEAT, peatworks.org). URL
GOVERN 2.3
Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.

About
Senior leadership and members of the C-Suite in organizations that maintain an AI portfolio, should maintain awareness of AI risks, affirm the organizational appetite for such risks, and be responsible for managing those risks.

Accountability ensures that a specific team and individual is responsible for AI risk management efforts. Some organizations grant authority and resources (human and budgetary) to a designated officer who ensures adequate performance of the institution’s AI portfolio (e.g., predictive modeling, machine learning).

Suggested Actions
Organizational management can:
- Declare risk tolerances for developing or using AI systems.
- Support AI risk management efforts and play an active role in such efforts.
- Integrate a risk and harm prevention mindset throughout the AI lifecycle as part of organizational culture.
- Support competent risk management executives.
- Delegate the power, resources, and authorization to perform risk management to each appropriate level throughout the management chain.

Organizations can establish board committees for AI risk management and oversight functions and integrate those functions within the organization’s broader enterprise risk management approaches.
**Transparency & Documentation**

Organizations can document the following:

- Did your organization’s board and/or senior management sponsor, support and participate in your organization’s AI governance?
- What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
- Do AI solutions provide sufficient information to assist the personnel to make an informed decision and take actions accordingly?
- To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?

**AI Transparency Resources**

- [WEF Companion to the Model AI Governance Framework- 2020.](https://www.weforum.org)

**References**


GOVERN 3
Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle.

GOVERN 3.1
Decision-makings related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds).

About
A diverse team that includes AI actors with diversity of experience, disciplines, and backgrounds to enhance organizational capacity and capability for anticipating risks is better equipped to carry out risk management. Consultation with external personnel may be necessary when internal teams lack a diverse range of lived experiences or disciplinary expertise.

To extend the benefits of diversity, equity, and inclusion to both the users and AI actors, it is recommended that teams are composed of a diverse group of individuals who reflect a range of backgrounds, perspectives and expertise.

Without commitment from senior leadership, beneficial aspects of team diversity and inclusion can be overridden by unstated organizational incentives that inadvertently conflict with the broader values of a diverse workforce.

Suggested Actions
Organizational management can:

- Define policies and hiring practices at the outset that promote interdisciplinary roles, competencies, skills, and capacity for AI efforts.
- Define policies and hiring practices that lead to demographic and domain expertise diversity; empower staff with necessary resources and support, and facilitate the contribution of staff feedback and concerns without fear of reprisal.
- Establish policies that facilitate inclusivity and the integration of new insights into existing practice.
- Seek external expertise to supplement organizational diversity, equity, inclusion, and accessibility where internal expertise is lacking.
- Establish policies that incentivize AI actors to collaborate with existing nondiscrimination, accessibility and accommodation, and human resource functions, employee resource group (ERGs), and diversity, equity, inclusion, and accessibility (DEIA) initiatives.

**Transparency & Documentation**

**Organizations can document the following:**
- Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?
- Entities include diverse perspectives from technical and non-technical communities throughout the AI life cycle to anticipate and mitigate unintended consequences including potential bias and discrimination.
- Stakeholder involvement: Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.
- Strategies to incorporate diverse perspectives include establishing collaborative processes and multidisciplinary teams that involve subject matter experts in data science, software development, civil liberties, privacy and security, legal counsel, and risk management.
- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?

**AI Transparency Resources**
- Datasheets for Datasets.
References


Mary L. Cummings and Songpo Li, 2021a. Sources of subjectivity in machine learning models. ACM Journal of Data and Information Quality, 13(2), 1–9

GOVERN 3.2
Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.

About
Identifying and managing AI risks and impacts are enhanced when a broad set of perspectives and actors across the AI lifecycle, including technical, legal, compliance, social science, and human factors expertise is engaged. AI actors include those who operate, use, or interact with AI systems for downstream tasks, or monitor AI system performance. Effective risk management efforts include:

- clear definitions and differentiation of the various human roles and responsibilities for AI system oversight and governance
- recognizing and clarifying differences between AI system overseers and those using or interacting with AI systems.

Suggested Actions
- Establish policies and procedures that define and differentiate the various human roles and responsibilities when using, interacting with, or monitoring AI systems.
- Establish procedures for capturing and tracking risk information related to human-AI configurations and associated outcomes.
- Establish policies for the development of proficiency standards for AI actors carrying out system operation tasks and system oversight tasks.
- Establish specified risk management training protocols for AI actors carrying out system operation tasks and system oversight tasks.
- Establish policies and procedures regarding AI actor roles, and responsibilities for human oversight of deployed systems.
- Establish policies and procedures defining human-AI configurations (configurations where AI systems are explicitly designated and treated as team members in primarily human teams) in relation to organizational risk tolerances, and associated documentation.
- Establish policies to enhance the explanation, interpretation, and overall transparency of AI systems.
- Establish policies for managing risks regarding known difficulties in human-AI configurations, human-AI teaming, and AI system user experience and user interactions (UI/UX).

## Transparency & Documentation

**Organizations can document the following:**

- What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?
- To what extent has the entity documented the appropriate level of human involvement in AI-augmented decision-making?
- How will the accountable human(s) address changes in accuracy and precision due to either an adversary’s attempts to disrupt the AI or unrelated changes in operational/business environment, which may impact the accuracy of the AI?
- To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?
- How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?

## AI Transparency Resources

- [Intel.gov: AI Ethics Framework for Intelligence Community - 2020](https://www.intel.gov/).
References


GOVERN 4
Organizational teams are committed to a culture that considers and communicates AI risk.

GOVERN 4.1
Organizational policies, and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize negative impacts.

About
A risk culture and accompanying practices can help organizations effectively triage the most critical risks. Organizations in some industries implement three (or more) “lines of defense,” where separate teams are held accountable for different aspects of the system lifecycle, such as development, risk management, and auditing. While a traditional three-lines approach may be impractical for smaller organizations, leadership can commit to cultivating a strong risk culture through other means. For example, “effective challenge,” is a culture- based practice that encourages critical thinking and questioning of important design and implementation decisions by experts with the authority and stature to make such changes.

Red-teaming is another risk measurement and management approach. This practice consists of adversarial testing of AI systems under stress conditions to seek out failure modes or vulnerabilities in the system. Red-teams are composed of external experts or personnel who are independent from internal AI actors.

Suggested Actions
- Establish policies that require inclusion of oversight functions (legal, compliance, risk management) from the outset of the system design process.
• Establish policies that promote effective challenge of AI system design, implementation, and deployment decisions, via mechanisms such as the three lines of defense, model audits, or red-teaming – to minimize workplace risks such as groupthink.
• Establish policies that incentivize safety-first mindset and general critical thinking and review at an organizational and procedural level.
• Establish whistleblower protections for insiders who report on perceived serious problems with AI systems.
• Establish policies to integrate a harm and risk prevention mindset throughout the AI lifecycle.

Transparency & Documentation
Organizations can document the following:
• To what extent has the entity documented the AI system’s development, testing methodology, metrics, and performance outcomes?
• Are organizational information sharing practices widely followed and transparent, such that related past failed designs can be avoided?
• Are training manuals and other resources for carrying out incident response documented and available?
• Are processes for operator reporting of incidents and near-misses documented and available?

AI Transparency Resources
• Datasheets for Datasets.
• GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
• WEF Model AI Governance Framework Assessment 2020.

References


**GAO, “Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities,” GAO@100 (GAO-21-519SP), June 2021.**

**Donald Sull, Stefano Turconi, and Charles Sull, “When It Comes to Culture, Does Your Company Walk the Talk?” MIT Sloan Mgmt. Rev., 2020.**

**Kathy Baxter, AI Ethics Maturity Model, Salesforce.**
GOVERN 4.2
Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate and use, and communicate about the impacts more broadly.

About
Impact assessments are one approach for driving responsible technology development practices. And, within a specific use case, these assessments can provide a high-level structure for organizations to frame risks of a given algorithm or deployment. Impact assessments can also serve as a mechanism for organizations to articulate risks and generate documentation for managing and oversight activities when harms do arise.

Impact assessments may:
- be applied at the beginning of a process but also iteratively and regularly since goals and outcomes can evolve over time.
- include perspectives from AI actors, including operators, users, and potentially impacted communities (including historically marginalized communities, those with disabilities, and individuals impacted by the digital divide),
- assist in “go/no-go” decisions for an AI system.
- consider conflicts of interest, or undue influence, related to the organizational team being assessed.

See the MAP function playbook guidance for more information relating to impact assessments.

Suggested Actions
- Establish impact assessment policies and processes for AI systems used by the organization.
- Align organizational impact assessment activities with relevant regulatory or legal requirements.
- Verify that impact assessment activities are appropriate to evaluate the potential negative impact of a system and how quickly a system changes, and that assessments are applied on a regular basis.
- Utilize impact assessments to inform broader evaluations of AI system risk.
**Transparency & Documentation**

Organizations can document the following:

- How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?
- How has the entity documented the AI system’s data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata?
- To what extent has the entity clearly defined technical specifications and requirements for the AI system?
- To what extent has the entity documented and communicated the AI system’s development, testing methodology, metrics, and performance outcomes?
- Have you documented and explained that machine errors may differ from human errors?

**AI Transparency Resources**

- [Datasheets for Datasets](https://www.datalos.com/datasheets).

**References**


BSA The Software Alliance (2021) Confronting Bias: BSA’s Framework to Build Trust in AI.


Kathy Baxter, AI Ethics Maturity Model, Salesforce
GOVERN 4.3
Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.

About
Identifying AI system limitations, detecting and tracking negative impacts and incidents, and sharing information about these issues with appropriate AI actors will improve risk management. Issues such as concept drift, AI bias and discrimination, shortcut learning or under specification are difficult to identify using current standard AI testing processes. Organizations can institute in-house use and testing policies and procedures to identify and manage such issues. Efforts can take the form of pre-alpha or pre-beta testing, or deploying internally developed systems or products within the organization. Testing may entail limited and controlled in-house, or publicly available, AI system testbeds, and accessibility of AI system interfaces and outputs.

Without policies and procedures that enable consistent testing practices, risk management efforts may be bypassed or ignored, exacerbating risks or leading to inconsistent risk management activities.

Information sharing about impacts or incidents detected during testing or deployment can:
- draw attention to AI system risks, failures, abuses or misuses,
- allow organizations to benefit from insights based on a wide range of AI applications and implementations, and
- allow organizations to be more proactive in avoiding known failure modes.

Organizations may consider sharing incident information with the AI Incident Database, the AIAAIC, users, impacted communities, or with traditional cyber vulnerability databases, such as the MITRE CVE list.

Suggested Actions
- Establish policies and procedures to facilitate and equip AI system testing.
- Establish organizational commitment to identifying AI system limitations and sharing of insights about limitations within appropriate AI actor groups.
• Establish policies for reporting and documenting incident response.
• Establish policies and processes regarding public disclosure of incidents and information sharing.
• Establish guidelines for incident handling related to AI system risks and performance.

**Transparency & Documentation**

Organizations can document the following:

• Did your organization address usability problems and test whether user interfaces served their intended purposes? Consulting the community or end users at the earliest stages of development to ensure there is transparency on the technology used and how it is deployed.

• Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?

• To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?

**AI Transparency Resources**

• [WEF Model AI Governance Framework Assessment 2020](#).
• [WEF Companion to the Model AI Governance Framework- 2020](#).

**References**


BSA The Software Alliance (2021) Confronting Bias: BSA’s Framework to Build Trust in AI.

“Using Combined Expertise to Evaluate Web Accessibility,” W3C Web Accessibility Initiative. [URL](#).
GOVERN 5
Processes are in place for robust engagement with relevant AI actors.

GOVERN 5.1
Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.

About
Beyond internal and laboratory-based system testing, organizational policies and practices may consider AI system fitness-for-purpose related to the intended context of use.

Participatory stakeholder engagement is one type of qualitative activity to help AI actors answer questions such as whether to pursue a project or how to design with impact in mind. This type of feedback, with domain expert input, can also assist AI actors to identify emergent scenarios and risks in certain AI applications. The consideration of when and how to convene a group and the kinds of individuals, groups, or community organizations to include is an iterative process connected to the system's purpose and its level of risk. Other factors relate to how to collaboratively and respectfully capture stakeholder feedback and insight that is useful, without being a solely perfunctory exercise.

These activities are best carried out by personnel with expertise in participatory practices, qualitative methods, and translation of contextual feedback for technical audiences.

Participatory engagement is not a one-time exercise and is best carried out from the very beginning of AI system commissioning through the end of the lifecycle. Organizations can consider how to incorporate engagement when beginning a project and as part of their monitoring of systems. Engagement is often utilized as a consultative practice, but this perspective may inadvertently lead to “participation washing.” Organizational transparency about the purpose and goal of the engagement can help mitigate that possibility.
Organizations may also consider targeted consultation with subject matter experts as a complement to participatory findings. Experts may assist internal staff in identifying and conceptualizing potential negative impacts that were previously not considered.

**Suggested Actions**

- Establish AI risk management policies that explicitly address mechanisms for collecting, evaluating, and incorporating stakeholder and user feedback that could include:
  - Recourse mechanisms for faulty AI system outputs.
  - Bug bounties.
  - Human-centered design.
  - User-interaction and experience research.
  - Participatory stakeholder engagement with individuals and communities that may experience negative impacts.

- Verify that stakeholder feedback is considered and addressed, including environmental concerns, and across the entire population of intended users, including historically excluded populations, people with disabilities, older people, and those with limited access to the internet and other basic technologies.

- Clarify the organization’s principles as they apply to AI systems – considering those which have been proposed publicly – to inform external stakeholders of the organization’s values. Consider publishing or adopting AI principles.

**Transparency & Documentation**

Organizations can document the following:

- What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?
- To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?
- How easily accessible and current is the information available to external stakeholders?
• What was done to mitigate or reduce the potential for harm?
• Stakeholder involvement: Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.

**AI Transparency Resources**

- [Datasheets for Datasets](#).
- [GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities](#).
- [AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019](#).
- [Stakeholders in Explainable AI, Sep. 2018](#).

**References**


Rumman Chowdhury and Jutta Williams, "Introducing Twitter’s first algorithmic bias bounty challenge."


Microsoft Community Jury, Azure Application Architecture Guide.

GOVERN 5.2
Mechanisms are established to enable AI actors to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation.

About
Organizational policies and procedures that equip AI actors with the processes, knowledge, and expertise needed to inform collaborative decisions about system deployment improve risk management. These decisions are closely tied to AI systems and organizational risk tolerance.

Risk tolerance, established by organizational leadership, reflects the level and type of risk the organization will accept while conducting its mission and carrying out its strategy. When risks arise, resources are allocated based on the assessed risk of a given AI system. Organizations typically apply a risk tolerance approach where higher risk systems receive larger allocations of risk management resources and lower risk systems receive less resources.

Suggested Actions
- Explicitly acknowledge that AI systems, and the use of AI, present inherent costs and risks along with potential benefits.
- Define reasonable risk tolerances for AI systems informed by laws, regulation, best practices, or industry standards.
- Establish policies that ensure all relevant AI actors are provided with meaningful opportunities to provide feedback on system design and implementation.
- Establish policies that define how to assign AI systems to established risk tolerance levels by combining system impact assessments with the likelihood that an impact occurs. Such assessment often entails some combination of:
  - Econometric evaluations of impacts and impact likelihoods to assess AI system risk.
  - Red-amber-green (RAG) scales for impact severity and likelihood to assess AI system risk.
Transparency & Documentation
Organizations can document the following:

- Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?
- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- Who is accountable for the ethical considerations during all stages of the AI lifecycle?
- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?
- Does the AI solution provide sufficient information to assist the personnel to make an informed decision and take actions accordingly?

AI Transparency Resources

- Stakeholders in Explainable AI, Sep. 2018.
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019.

References


Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.

Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third party’s intellectual property or other rights.

About
Risk measurement and management can be complicated by how customers use or integrate third-party data or systems into AI products or services, particularly without sufficient internal governance structures and technical safeguards.

Organizations usually engage multiple third parties for external expertise, data, software packages (both open source and commercial), and software and hardware platforms across the AI lifecycle. This engagement has beneficial uses and can increase complexities of risk management efforts.

Organizational approaches to managing third-party (positive and negative) risk may be tailored to the resources, risk profile, and use case for each system. Organizations can apply governance approaches to third-party AI systems and data as they would for internal resources — including open-source software, publicly available data, and commercially available models.

Suggested Actions
• Collaboratively establish policies that address third-party AI systems and data.
• Establish policies related to:
  ○ Transparency into third-party system functions, including knowledge about training data, training and inference algorithms, and assumptions and limitations.
  ○ Thorough testing of third-party AI systems. (See MEASURE for more detail)
- Requirements for clear and complete instructions for third-party system usage. Evaluate policies for third-party technology.
- Establish policies that address supply chain, full product lifecycle and associated processes, including legal, ethical, and other issues concerning procurement and use of third-party software or hardware systems and data.

**Transparency & Documentation**

Organizations can document the following:
- Did you establish mechanisms that facilitate the AI system’s auditability (e.g., traceability of the development process, the sourcing of training data and the logging of the AI system’s processes, outcomes, positive and negative impact)?
- If a third party created the AI, how will you ensure a level of explainability or interpretability?
- Did you ensure that the AI system can be audited by independent third parties?
- Did you establish a process for third parties (e.g., suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?
- To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?

**AI Transparency Resources**
References


GOVERN 6.2
Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

About
To mitigate the potential harms of third-party system failures, organizations may implement policies and procedures that include redundancies for covering third-party functions.

Suggested Actions
- Establish policies for handling third-party system failures to include consideration of redundancy mechanisms for vital third-party AI systems.
- Verify that incident response plans address third-party AI systems.

Transparency & Documentation
Organizations can document the following:
- To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?
- Did you establish a process for third parties (e.g., suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?
- If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?

AI Transparency Resources
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019.
References


MAP

Map
Context is recognized and risks related to context are identified
MAP 1
Context is established and understood.

MAP 1.1
Intended purpose, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes; uses and risks across the development or product AI lifecycle; TEVV and system metrics.

About
Highly accurate and optimized systems can cause harm. Relatedly, organizations should expect broadly deployed AI tools to be reused, repurposed, and potentially misused regardless of intentions.

AI actors can work collaboratively, and with external parties such as community groups, to help delineate the bounds of acceptable deployment, consider preferable alternatives, and identify principles and strategies to manage likely risks. Context mapping is the first step in this effort, and may include examination of the following:

- intended purpose and impact of system use.
- concept of operations.
- intended, prospective, and actual deployment setting.
- requirements for system deployment and operation.
- end user and operator expectations.
- specific set or types of end users.
- potential negative impacts to individuals, groups, communities, organizations, and society – or context-specific impacts such as legal requirements or impacts to the environment.
- unanticipated, downstream, or other unknown contextual factors.
- how AI system changes connect to impacts.
These types of processes can assist AI actors in understanding how limitations, constraints, and other realities associated with the deployment and use of AI technology can create impacts once they are deployed or operate in the real world. When coupled with the enhanced organizational culture resulting from the established policies and procedures in the Govern function, the Map function can provide opportunities to foster and instill new perspectives, activities, and skills for approaching risks and impacts.

Context mapping also includes discussion and consideration of non-AI or non-technology alternatives especially as related to whether the given context is narrow enough to manage AI and its potential negative impacts. Non-AI alternatives may include capturing and evaluating information using semi-autonomous or mostly manual methods.

**Suggested Actions**

- Maintain awareness of industry, technical, and applicable legal standards.
- Examine trustworthiness of AI system design and consider, non-AI solutions.
- Consider intended AI system design tasks along with unanticipated purposes in collaboration with human factors and socio-technical domain experts.
- Define and document the task, purpose, minimum functionality, and benefits of the AI system to inform considerations about whether the utility of the project or its lack of.
- Identify whether there are non-AI or non-technology alternatives that will lead to more trustworthy outcomes.
- Examine how changes in system performance affect downstream events such as decision-making (e.g.: changes in an AI model objective function create what types of impacts in how many candidates do/do not get a job interview).
- Determine the end user and organizational requirements, including business and technical requirements.
• Determine and delineate the expected and acceptable AI system context of use, including:
  ○ social norms
  ○ impacted individuals, groups, and communities
  ○ potential positive and negative impacts to individuals, groups, communities, organizations, and society
  ○ operational environment
• Perform context analysis related to time frame, safety concerns, geographic area, physical environment, ecosystems, social environment, and cultural norms within the intended setting (or conditions that closely approximate the intended setting).
• Gain and maintain awareness about evaluating scientific claims related to AI system performance and benefits before launching into system design.
• Identify human-AI interaction and/or roles, such as whether the application will support or replace human decision making.
• Plan for risks related to human-AI configurations, and document requirements, roles, and responsibilities for human oversight of deployed systems.

Transparency & Documentation
Organizations can document the following:
• To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?
• Did you establish a process for third parties (e.g., suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?
• If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?
AI Transparency Resources

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities,
- "Microsoft Responsible AI Standard, v2"

References

Socio-technical systems

Problem formation


Context mapping


Identification of harms


Understanding and documenting limitations in ML


Understanding and documenting limitations in ML


Meta AI. System Cards, a new resource for understanding how AI systems work, 2021.

When not to deploy
Statistical balance

Assessment of science in AI
Arvind Narayanan. How to recognize AI snake oil. URL
Emily M. Bender. 2022. On NYT Magazine on AI: Resist the Urge to be Impressed. (April 17, 2022).

Emily M. Bender. 2022. On NYT Magazine on AI: Resist the Urge to be Impressed. (April 17, 2022).
MAP 1.2
Inter-disciplinary AI actors, competencies, skills and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.

About
Successfully mapping context requires a team of AI actors with a diversity of experience, expertise, abilities and backgrounds, and with the resources and independence to engage in critical inquiry.

Having a diverse team contributes to more broad and open sharing of ideas and assumptions about the purpose and function of the technology being designed and developed – making these implicit aspects more explicit. The benefit of a diverse staff in managing AI risks is not the beliefs or presumed beliefs of individual workers, but the behavior that results from a collective perspective. An environment which fosters critical inquiry creates opportunities to surface problems and identify existing and emergent risks.

Suggested Actions
- Establish interdisciplinary teams to reflect a wide range of skills, competencies, and capabilities for AI efforts. Verify that team membership includes demographic diversity, broad domain expertise, and lived experiences. Document team composition.
- Create and empower interdisciplinary expert teams to capture, learn, and engage the interdependencies of deployed AI systems and related terminologies and concepts from disciplines outside of AI practice such as law, sociology, psychology, anthropology, public policy, systems design, and engineering.

Transparency & Documentation
Organizations can document the following:
- To what extent do the teams responsible for developing and maintaining the AI system reflect diverse opinions, backgrounds, experiences, and perspectives?
• Did the entity document the demographics of those involved in the design and development of the AI system to capture and communicate potential biases inherent to the development process, according to forum participants?
• What specific perspectives did stakeholders share, and how were they integrated across the design, development, deployment, assessment, and monitoring of the AI system?
• To what extent has the entity addressed stakeholder perspectives on the potential negative impacts of the AI system on end users and impacted populations?
• What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?
• Did your organization address usability problems and test whether user interfaces served their intended purposes? Consulting the community or end users at the earliest stages of development to ensure there is transparency on the technology used and how it is deployed.

AI Transparency Resources
• GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
• WEF Model AI Governance Framework Assessment 2020.
• WEF Companion to the Model AI Governance Framework- 2020.
• AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019.

References

Microsoft Community Jury., Azure Application Architecture Guide.


MAP 1.3

The organization’s mission and relevant goals for the AI technology are understood and documented.

About

Defining and documenting the specific business purpose of an AI system in a broader context of societal values helps teams to evaluate risks and increases the clarity of “go/no-go” decisions about whether to deploy.

Trustworthy AI technologies may present a demonstrable business benefit beyond implicit or explicit costs, provide added value, and don’t lead to wasted resources. Organizations can feel confident in performing risk avoidance if the implicit or explicit risks outweigh the advantages of AI systems, and not implementing an AI solution whose risks surpass potential benefits.

For example, making AI systems more equitable can result in better managed risk, and can help enhance consideration of the business value of making inclusively designed, accessible and more equitable AI systems.

Suggested Actions

- Build transparent practices into AI system development processes.
- Review the documented system purpose from a socio-technical perspective and in consideration of societal values.
- Determine possible misalignment between societal values and stated organizational principles and code of ethics.
- Flag latent incentives that may contribute to negative impacts.
- Evaluate AI system purpose in consideration of potential risks, societal values, and stated organizational principles.

Transparency & Documentation

Organizations can document the following:

- How does the AI system help the entity meet its goals and objectives?
- How do the technical specifications and requirements align with the AI system’s goals and objectives?
- To what extent is the output appropriate for the operational context?
Al Transparency Resources

- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019, URL, LINK
- Including Insights from the Comptroller General’s Forum on the Oversight of Artificial Intelligence An Accountability Framework for Federal Agencies and Other Entities, 2021, URL, PDF.

References


PEAT “Business Case for Equitable AI”.
MAP 1.4
The business value or context of business use has been clearly defined or – in the case of assessing existing AI systems – re-evaluated.

About
Socio-technical AI risks emerge from the interplay between technical development decisions and how a system is used, who operates it, and the social context into which it is deployed. Addressing these risks is complex and requires a commitment to understanding how contextual factors may interact with AI lifecycle actions. One such contextual factor is how organizational mission and identified system purpose create incentives within AI system design, development, and deployment tasks that may result in positive and negative impacts. By establishing comprehensive and explicit enumeration of AI systems’ context of business use and expectations, organizations can identify and manage these types of risks.

Suggested Actions
- Document business value or context of business use
- Reconcile documented concerns about the system’s purpose within the business context of use compared to the organization’s stated values, mission statements, social responsibility commitments, and AI principles.
- Reconsider the design, implementation strategy, or deployment of AI systems with potential impacts that do not reflect institutional values.

Transparency & Documentation
Organizations can document the following:
- What goals and objectives does the entity expect to achieve by designing, developing, and/or deploying the AI system?
- To what extent are the system outputs consistent with the entity’s values and principles to foster public trust and equity?
- To what extent are the metrics consistent with system goals, objectives, and constraints, including ethical and compliance considerations?
AI Transparency Resources

- WEF Model AI Governance Framework Assessment 2020. URL

References

Algorithm Watch. AI Ethics Guidelines Global Inventory.

Ethical OS toolkit.


Organizational risk tolerances are determined and documented.

**About**

Risk tolerance reflects the level and type of risk the organization is willing to accept while conducting its mission and carrying out its strategy.

Organizations can follow existing regulations and guidelines for risk criteria, tolerance and response established by organizational, domain, discipline, sector, or professional requirements. Some sectors or industries may have established definitions of harm or may have established documentation, reporting, and disclosure requirements.

Within sectors, risk management may depend on existing guidelines for specific applications and use case settings. Where established guidelines do not exist, organizations will want to define reasonable risk tolerance in consideration of different sources of risk (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).

Risk tolerances inform and support decisions about whether to continue with development or deployment - termed “go/no-go”. Go/no-go decisions related to AI system risks can take stakeholder feedback into account but remain independent from stakeholders’ vested financial or reputational interests.

If mapping risk is prohibitively difficult, a "no-go" decision may be considered for the specific system.

**Suggested Actions**

- Utilize existing regulations and guidelines for risk criteria, tolerance and response established by organizational, domain, discipline, sector, or professional requirements.
- Establish risk tolerance levels for AI systems and allocate the appropriate oversight resources to each level.
• Establish risk criteria in consideration of different sources of risk, (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).

• Identify maximum allowable risk tolerance above which the system will not be deployed, or will need to be prematurely decommissioned, within the contextual or application setting.

• Articulate and analyze tradeoffs across trustworthiness characteristics as relevant to proposed context of use. When tradeoffs arise, document them and plan for traceable actions (e.g.: impact mitigation, removal of system from development or use) to inform management decisions.

• Review uses of AI systems for “off-label” purposes, especially in settings that organizations have deemed as high-risk. Document decisions, risk-related trade-offs, and system limitations.

Transparency & Documentation

Organizations can document the following:
• Which existing regulations and guidelines apply, and the entity has followed, in the development of system risk tolerances?
• What criteria and assumptions has the entity utilized when developing system risk tolerances?
• How has the entity identified maximum allowable risk tolerance?
• What conditions and purposes are considered “off-label” for system use?

AI Transparency Resources
• GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
• WEF Model AI Governance Framework Assessment 2020.
• WEF Companion to the Model AI Governance Framework- 2020.
References


GAO-17-63: Enterprise Risk Management: Selected Agencies’ Experiences Illustrate Good Practices in Managing Risk. (See Table 3).

NIST Risk Management Framework.
System requirements (e.g., “the system shall respect the privacy of its users”) are elicited from and understood by relevant AI actors. Design decisions take socio-technical implications into account to address AI risks.

About

AI system development requirements may outpace documentation processes for traditional software. When written requirements are unavailable or incomplete, AI actors may inadvertently overlook business and stakeholder needs, over-rely on implicit human biases such as confirmation bias and groupthink, and maintain exclusive focus on computational requirements.

Eliciting system requirements, designing for end users, and considering societal impacts early in the design phase is a priority that can enhance AI systems’ trustworthiness.

Suggested Actions

- Proactively incorporate trustworthy characteristics into system requirements.
- Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to system design or deployment decisions.
- Develop and standardize practices to assess potential impacts at all stages of the AI lifecycle, and in collaboration with interdisciplinary experts, actors external to the team that developed or deployed the AI system, and potentially impacted communities.
- Include potentially impacted groups, communities and external entities (e.g., civil society organizations, research institutes, local community groups, and trade associations) in the formulation of priorities, definitions and outcomes during impact assessment activities.
- Conduct qualitative interviews with end user(s) to regularly evaluate expectations and design plans related to Human-AI configurations and tasks.
Suggested Actions

- Analyze dependencies between contextual factors and system requirements. List potential impacts that may arise from not fully considering the importance of trustworthiness characteristics in any decision making.
- Follow responsible design techniques in tasks such as software engineering, product management, and participatory engagement. Some examples for eliciting and documenting stakeholder requirements include product requirement documents (PRDs), user stories, user interaction/user experience (UI/UX) research, systems engineering, ethnography and related field methods.
- Conduct user research to understand individuals, groups and communities that will be impacted by the AI, their values & context, and the role of systemic and historical biases. Integrate learnings into decisions about data selection and representation.

Transparency & Documentation

Organizations can document the following:

- What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?
- To what extent is this information sufficient and appropriate to promote transparency? Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.
- To what extent has relevant information been disclosed regarding the use of AI systems, such as (a) what the system is for, (b) what it is not for, (c) how it was designed, and (d) what its limitations are? (Documentation and external communication can offer a way for entities to provide transparency.)
- How will the relevant AI actor(s) address changes in accuracy and precision due to either an adversary’s attempts to disrupt the AI system or unrelated changes in the operational/business environment, which may impact the accuracy of the AI system?
• What metrics has the entity developed to measure performance of the AI system?
• What justifications, if any, has the entity provided for the assumptions, boundaries, and limitations of the AI system?

**AI Transparency Resources**

- **GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.**
- **Stakeholders in Explainable AI, Sep. 2018.**
- **High-Level Expert Group on Artificial Intelligence set up by the European Commission, Ethics Guidelines for Trustworthy AI. URL, PDF**

**References**


MAP 2
Categorization of the AI system is performed.

MAP 2.1
The specific task, and methods used to implement the task, that the AI system will support is defined (e.g., classifiers, generative models, recommenders).

About
AI actors define the technical learning or decision-making task(s) an AI system is designed to accomplish, or the benefits that the system will provide. The clearer and narrower the task definition, the easier it is to map its benefits and risks, leading to more fulsome risk management.

Suggested Actions
- Define and document AI system’s existing and potential learning task(s) along with known assumptions and limitations.

Transparency & Documentation
Organizations can document the following:
- To what extent has the entity clearly defined technical specifications and requirements for the AI system?
- To what extent has the entity documented the AI system’s development, testing methodology, metrics, and performance outcomes?
- How do the technical specifications and requirements align with the AI system’s goals and objectives?
- Did your organization implement accountability-based practices in data management and protection (e.g., the PDPA and OECD Privacy Principles)?
- How are outputs marked to clearly show that they came from an AI?

AI Transparency Resources
- Datasheets for Datasets.
- WEF Companion to the Model AI Governance Framework- 2020. URL
References


Information about the AI system’s knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making informed decisions and taking subsequent actions.

About
An AI lifecycle consists of many interdependent activities involving a diverse set of actors that often do not have full visibility or control over other parts of the lifecycle and its associated contexts or risks. The interdependencies between these activities, and among the relevant AI actors and organizations, can make it difficult to reliably anticipate potential impacts of AI systems. For example, early decisions in identifying the purpose and objective of an AI system can alter its behavior and capabilities, and the dynamics of deployment setting (such as end users or impacted individuals) can shape the positive or negative impacts of AI system decisions. As a result, the best intentions within one dimension of the AI lifecycle can be undermined via interactions with decisions and conditions in other, later activities. This complexity and varying levels of visibility can introduce uncertainty. And, once deployed and in use, AI systems may sometimes perform poorly, manifest unanticipated negative impacts, or violate legal or ethical norms. These risks and incidents can result from a variety of factors. For example, downstream decisions can be influenced by end user over-trust or under-trust, and other complexities related to AI-supported decision-making.

Anticipating, articulating, assessing and documenting AI systems’ knowledge limits and how system output may be utilized and overseen by humans can help mitigate the uncertainty associated with the realities of AI system deployments. Rigorous design processes include defining system knowledge limits, which are confirmed and refined based on TEVV processes.

Suggested Actions
- Document settings, environments and conditions that are outside the AI system’s intended use.
• Design for end user workflows and toolsets, concept of operations, and explainability and interpretability criteria in conjunction with end user(s) and associated qualitative feedback.

• Plan and test human-AI configurations under close to real-world conditions and document results.

• Follow stakeholder feedback processes to determine whether a system achieved its documented purpose within a given use context, and whether end users can correctly comprehend system outputs or results.

• Document dependencies on upstream data and other AI systems, including if the specified system is an upstream dependency for another AI system or other data.

• Document connections the AI system or data will have to external networks (including the internet), financial markets, and critical infrastructure that have potential for negative externalities. Identify and document negative impacts as part of considering the broader risk thresholds and subsequent go/no-go deployment as well as post-deployment decommissioning decisions.

**Transparency & Documentation**

Organizations can document the following:

• Does the AI system provide sufficient information to assist the personnel to make an informed decision and take actions accordingly?

• What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

• Based on the assessment, did your organization implement the appropriate level of human involvement in AI-augmented decision-making?

**AI Transparency Resources**

• [Datasheets for Datasets. URL](#)

• [WEF Model AI Governance Framework Assessment 2020.](#)

• [WEF Companion to the Model AI Governance Framework- 2020.](#)
References

Context of use


Human-AI interaction


Human Readiness Level Scale in the System Development Process, American National Standards Institute and Human Factors and Ergonomics Society, ANSI/HFES 400-2021

Microsoft Responsible AI Standard, v2.


Scientific integrity and 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.

About
Standard testing and evaluation protocols provide a basis to confirm assurance in a system that it is operating as designed and claimed. AI systems’ complexities create challenges for traditional testing and evaluation methodologies, which tend to be designed for static or isolated system performance. Opportunities for risk continue well beyond design and deployment, into system operation and application of system-enabled decisions. Testing and evaluation methodologies and metrics therefore address a continuum of activities. TEVV is enhanced when key metrics for performance, safety, and reliability are interpreted in a socio-technical context and not confined to the boundaries of the AI system pipeline.

Other challenges for managing AI risks relate to dependence on large scale datasets, which can impact data quality and validity concerns. The difficulty of finding the “right” data may lead AI actors to select datasets based more on accessibility and availability than on suitability for operationalizing the phenomenon that the AI system intends to support or inform. Such decisions could contribute to an environment where the data used in processes is not fully representative of the populations or phenomena that are being modeled, introducing downstream risks. Practices such as dataset reuse may also lead to disconnect from the social contexts and time periods of their creation. This contributes to issues of validity of the underlying dataset for providing proxies, measures, or predictors within the model.

Suggested Actions
- Identify and document experiment design and statistical techniques that are valid for testing complex socio-technical systems like AI, which involve human factors, emergent properties, and dynamic context(s) of use.
• Develop and apply TEVV protocols for models, system and its subcomponents, deployment, and operation.

• Demonstrate and document that AI system performance and validation metrics are interpretable and unambiguous for downstream decision making tasks and take socio-technical factors such as context of use into consideration.

• Identify and document assumptions, techniques, and metrics used for testing and evaluation throughout the AI lifecycle including experimental design techniques for data collection, selection, and management practices in accordance with data governance policies established in GOVERN.

• Identify testing modules that can be incorporated throughout the AI lifecycle and verify that processes enable corroboration by independent evaluators.

• Establish mechanisms for regular communication and feedback among relevant AI actors and internal or external stakeholders related to the validity of design and deployment assumptions.

• Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to the development of TEVV approaches throughout the lifecycle to detect and assess potentially harmful impacts.

• Document assumptions made and techniques used in data selection, curation, preparation and analysis, including:
  - identification of constructs and proxy targets,
  - development of indices – especially those operationalizing concepts that are inherently unobservable (e.g., “hireability,” “criminality,” “lendability”).

• Map adherence to policies that address data and construct validity, bias, privacy and security for AI systems and verify documentation, oversight, and processes.

• Identify and document transparent methods (e.g., causal discovery methods) for inferring causal relationships between constructs being modeled and dataset attributes or proxies.
• Identify and document processes to understand and trace test and training data lineage and its metadata resources for mapping risks.
• Document known limitations, risk mitigation efforts associated with, and methods used for, training data collection, selection, labeling, cleaning, and analysis (e.g., treatment of missing, spurious, or outlier data; biased estimators).
• Establish and document practices to check for capabilities that are in excess of those that are planned for, such as emergent properties, and to revisit prior risk management steps in light of any new capabilities.
• Establish processes to test and verify that design assumptions about the set of deployment contexts continue to be accurate and sufficiently complete.
• Work with domain experts and other external AI actors to:
  ◦ Gain and maintain contextual awareness and knowledge about how human behavior, organizational factors and dynamics, and society influence, and are represented in, datasets, processes, models, and system output.
  ◦ Identify participatory approaches for responsible Human-AI configurations and oversight tasks, taking into account sources of cognitive bias.
  ◦ Identify techniques to manage and mitigate sources of bias (systemic, computational, human- cognitive) in computational models and systems, and the assumptions and decisions in their development.
• Investigate and document potential negative impacts due related to the full product lifecycle and associated processes that may conflict with organizational values and principles.

Transparency & Documentation
Organizations can document the following:
• Are there any known errors, sources of noise, or redundancies in the data?
• Over what timeframe was the data collected? Does the collection timeframe match the creation time frame
What is the variable selection and evaluation process?
How was the data collected? Who was involved in the data collection process? If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)
As time passes and conditions change, is the training data still representative of the operational environment?
Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)
How does the entity ensure that the data collected are adequate, relevant, and not excessive in relation to the intended purpose?

AI Transparency Resources
- Datasheets for Datasets.

References

Challenges with dataset selection


Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. arXiv:1608.07836. URL
Dataset and test, evaluation, validation and verification (TEVV) processes in AI system development


Statistical balance


Measurement and evaluation


Existing frameworks


MAP 3
Al capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.

MAP 3.1
Potential benefits of intended AI system functionality and performance are examined and documented.

About
AI systems have enormous potential to improve quality of life, enhance economic prosperity and security costs. Organizations are encouraged to define and document system purpose and utility, and its potential positive impacts. benefits beyond current known performance benchmarks.

It is encouraged that risk management and assessment of benefits and impacts include processes for regular and meaningful communication with potentially affected groups and communities. These stakeholders can provide valuable input related to systems’ benefits and possible limitations. Organizations may differ in the types and number of stakeholders with which they engage.

Other approaches such as human-centered design (HCD) and value-sensitive design (VSD) can help AI teams to engage broadly with individuals and communities. This type of engagement can enable AI teams to learn about how a given technology may cause positive or negative impacts, that were not originally considered or intended.

Suggested Actions
- Utilize participatory approaches and engage with system end users to understand and document AI systems’ potential benefits, efficacy and interpretability of AI task output.
- Maintain awareness and documentation of the individuals, groups, or communities who make up the system’s internal and external stakeholders.
• Verify that appropriate skills and practices are available in-house for carrying out participatory activities such as eliciting, capturing, and synthesizing user, operator and external feedback, and translating it for AI design and development functions.
• Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to system design or deployment decisions.
• Consider performance to human baseline metrics or other standard benchmarks.
• Incorporate feedback from end users, and potentially impacted individuals and communities about perceived system benefits.

**Transparency & Documentation**
**Organizations can document the following:**
• Have the benefits of the AI system been communicated to end users?
• Have the appropriate training material and disclaimers about how to adequately use the AI system been provided to end users?
• Has your organization implemented a risk management system to address risks involved in deploying the identified AI system (e.g., personnel risk or changes to commercial objectives)?

**AI Transparency Resources**
• [Intel.gov: AI Ethics Framework for Intelligence Community - 2020.](#)
• [GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.](#)
• [Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019.](#)

**References**


MAP 3.2
Potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness - as connected to organizational risk tolerance - are examined and documented.

About
Anticipating negative impacts of AI systems is a difficult task. Negative impacts can be due to many factors, such as system non-functionality or use outside of its operational limits, and may range from minor annoyance to serious injury, financial losses, or regulatory enforcement actions. AI actors can work with a broad set of stakeholders to improve their capacity for understanding systems’ potential impacts – and subsequently – systems’ risks.

Suggested Actions
- Perform context analysis to map potential negative impacts arising from not integrating trustworthiness characteristics. When negative impacts are not direct or obvious, AI actors can engage with stakeholders external to the team that developed or deployed the AI system, and potentially impacted communities, to examine and document:
  - Who could be harmed?
  - What could be harmed?
  - When could harm arise?
  - How could harm arise?
- Identify and implement procedures for regularly evaluating the qualitative and quantitative costs of internal and external AI system failures. Develop actions to prevent, detect, and/or correct potential risks and related impacts. Regularly evaluate failure costs to inform go/no-go deployment decisions throughout the AI system lifecycle.
Transparency & Documentation
Organizations can document the following:

- To what extent does the system/entity consistently measure progress towards stated goals and objectives?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- Have you documented and explained that machine errors may differ from human errors?

AI Transparency Resources

- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019. LINK.

References


MAP 3.3
Targeted application scope is specified and documented based on the system’s capability, established context, and AI system categorization.

About
Systems that function in a narrow scope tend to enable better mapping, measurement, and management of risks in the learning or decision-making tasks and the system context. A narrow application scope also helps ease TEVV functions and related resources within an organization.

For example, large language models or open-ended chatbot systems that interact with the public on the internet have a large number of risks that may be difficult to map, measure, and manage due to the variability from both the decision-making task and the operational context. Instead, a task-specific chatbot utilizing templated responses that follow a defined “user journey” is a scope that can be more easily mapped, measured and managed.

Suggested Actions
- Consider narrowing contexts for system deployment, including factors related to: - How outcomes may directly or indirectly affect users, groups, communities and the environment. - Length of time the system is deployed in between re-trainings. - Geographical regions in which the system operates. - Dynamics related to community standards or likelihood of system misuse or abuses (either purposeful or unanticipated). - How AI system features and capabilities can be utilized within other applications, or in place of other existing processes.
- Engage AI actors from legal and procurement functions when specifying target application scope.

Transparency & Documentation
Organizations can document the following:
- To what extent has the entity clearly defined technical specifications and requirements for the AI system?
How do the technical specifications and requirements align with the AI system’s goals and objectives?

**AI Transparency Resources**
- [GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.](#)
- [Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019.](#)

**References**


UK Centre for Data Ethics and Innovation, “The roadmap to an effective AI assurance ecosystem”.

---

86
**MAP 3.4**
Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed and documented.

**About**
Human-AI configurations can span from fully autonomous to fully manual. AI systems can autonomously make decisions, defer decision-making to a human expert, or be used by a human decision-maker as an additional opinion. In some scenarios, professionals with expertise in a specific domain work in conjunction with an AI system towards a specific end goal—for example, a decision about another individual(s). Depending on the purpose of the system, the expert may interact with the AI system but is rarely part of the design or development of the system itself. These experts are not necessarily familiar with machine learning, data science, computer science, or other fields traditionally associated with AI design or development and - depending on the application - will likely not require such familiarity. For example, for AI systems that are deployed in health care delivery the experts are the physicians and bring their expertise about medicine—not data science, data modeling and engineering, or other computational factors. The challenge in these settings is not educating the end user about AI system capabilities, but rather leveraging, and not replacing, practitioner domain expertise.

Questions remain about how to configure humans and automation for managing AI risks. Risk management is enhanced when organizations that design, develop or deploy AI systems for use by professional operators and practitioners:
- are aware of these knowledge limitations and strive to identify risks in human-AI interactions and configurations across all contexts, and the potential resulting impacts,
- define and differentiate the various human roles and responsibilities when using or interacting with AI systems, and
- determine proficiency standards for AI system operation in proposed context of use, as enumerated in MAP-1 and established in GOVERN-3.2.
Suggested Actions

- Identify and declare AI system features and capabilities that may affect downstream AI actors’ decision-making in deployment and operational settings for example how system features and capabilities may activate known risks in various human-AI configurations, such as selective adherence.
- Identify skills and proficiency requirements for operators, practitioners and other domain experts that interact with AI systems, Develop AI system operational documentation for AI actors in deployed and operational environments, including information about known risks, mitigation criteria, and trustworthy characteristics enumerated in Map-1.
- Define and develop training materials for proposed end users, practitioners and operators about AI system use and known limitations.
- Define and develop certification procedures for operating AI systems within defined contexts of use, and information about what exceeds operational boundaries.
- Include operators, practitioners and end users in AI system prototyping and testing activities to help inform operational boundaries and acceptable performance. Conduct testing activities under scenarios similar to deployment conditions.
- Verify model output provided to AI system operators, practitioners and end users is interactive, and specified to context and user requirements defined in Map-1.
- Verify AI system output is interpretable and unambiguous for downstream decision-making tasks.
- Design AI system explanation complexity to match the level of problem and context complexity.
- Verify that design principles are in place for safe operation by AI actors in decision-making environments.
- Develop approaches to track human-AI configurations, operator, and practitioner outcomes for integration into continual improvement.
Transparency & Documentation

Organizations can document the following:

- What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles?
- How will the accountable human(s) address changes in accuracy and precision due to either an adversary’s attempts to disrupt the AI or unrelated changes in operational/business environment, which may impact the accuracy of the AI?
- How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?
- Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?
- What metrics has the entity developed to measure performance of various components?

AI Transparency Resources


References


MAP 3.5
Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from GOVERN function.

About
As AI systems have evolved in accuracy and precision, computational systems have moved from being used purely for decision support—or for explicit use by and under the control of a human operator—to automated decision making with limited input from humans. Computational decision support systems augment another, typically human, system in making decisions. These types of configurations increase the likelihood of outputs being produced with little human involvement.

Defining and differentiating various human roles and responsibilities for AI systems’ governance, and differentiating AI system overseers and those using or interacting with AI systems can enhance AI risk management activities.

In critical systems, high-stakes settings, and systems deemed high-risk it is of vital importance to evaluate risks and effectiveness of oversight procedures before an AI system is deployed.

Ultimately, AI system oversight is a shared responsibility, and attempts to properly authorize or govern oversight practices will not be effective without organizational buy-in and accountability mechanisms, for example those suggested in the GOVERN function.

Suggested Actions
- Identify and document AI systems’ features and capabilities that require human oversight, in relation to operational and societal contexts, trustworthy characteristics, and risks identified in MAP-1.
- Establish practices for AI systems’ oversight in accordance with policies developed in GOVERN-1.
- Define and develop training materials for relevant AI Actors about AI system performance, context of use, known limitations and negative impacts, and suggested warning labels.
• Include relevant AI Actors in AI system prototyping and testing activities. Conduct testing activities under scenarios similar to deployment conditions.
• Evaluate AI system oversight practices for validity and reliability. When oversight practices undergo extensive updates or adaptations, retest, evaluate results, and course correct as necessary.
• Verify that model documents contain interpretable descriptions of system mechanisms, enabling oversight personnel to make informed, risk-based decisions about system risks.

**Transparency & Documentation**

Organizations can document the following:
• What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
• How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?
• Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?
• To what extent has the entity documented the AI system’s development, testing methodology, metrics, and performance outcomes?

**AI Transparency Resources**

**References**


**MAP 4**
Risks and benefits are mapped for all components of the AI system including third-party software and data.

**MAP 4.1**
Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party’s intellectual property or other rights.

**About**
Technologies and personnel from third parties are another potential sources of risk to consider during AI risk management activities. Such risks may be difficult to map since risk priorities or tolerances may not be the same as the deployer organization.

For example, the use of pre-trained models, which tend to rely on large uncurated dataset or often have undisclosed origins, has raised concerns about privacy, bias, and unanticipated effects along with possible introduction of increased levels of statistical uncertainty, difficulty with reproducibility, and issues with scientific validity.

**Suggested Actions**
- Review audit reports, testing results, product roadmaps, warranties, terms of service, end user license agreements, contracts, and other documentation related to third-party entities to assist in value assessment and risk management activities.
- Review third-party software release schedules and software change management plans (hotfixes, patches, updates, forward- and backward- compatibility guarantees) for irregularities that may contribute to AI system risks.
- Inventory third-party material (hardware, open-source software, foundation models, open-source data, proprietary software, proprietary data, etc.) required for system implementation and maintenance.
• Review redundancies related to third-party technology and personnel to assess potential risks due to lack of adequate support.

**Transparency & Documentation**

Organizations can document the following:

- Did you establish a process for third parties (e.g., suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?
- If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?
- How will the results be independently verified?

**AI Transparency Resources**

- [GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.](#)
- [Intel.gov: AI Ethics Framework for Intelligence Community - 2020.](#)
- [WEF Model AI Governance Framework Assessment 2020.](#)

**References**

**Language models**


MAP 4.2
Internal risk controls for components of the AI system including third-party AI technologies are identified and documented.

About
In the course of their work, AI actors often utilize open-source, or otherwise freely available, third-party technologies – some of which may have privacy, bias, and security risks. Organizations may consider internal risk controls for these technology sources and build up practices for evaluating third-party material prior to deployment.

Suggested Actions
- Track third parties preventing or hampering risk-mapping as indications of increased risk.
- Supply resources such as model documentation templates and software safelists to assist in third-party technology inventory and approval activities.
- Review third-party material (including data and models) for risks related to bias, data privacy, and security vulnerabilities.
- Apply traditional technology risk controls – such as procurement, security, and data privacy controls – to all acquired third-party technologies.

Transparency & Documentation
Organizations can document the following:
- Can the AI system be audited by independent third parties?
- To what extent do these policies foster public trust and confidence in the use of the AI system?
- Are mechanisms established to facilitate the AI system’s auditability (e.g., traceability of the development process, the sourcing of training data and the logging of the AI system’s processes, outcomes, positive and negative impact)?

AI Transparency Resources
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019. LINK

References


MAP 5
Impacts to individuals, groups, communities, organizations, and society are characterized.

MAP 5.1
Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.

About
AI actors can evaluate, document and triage the likelihood of AI system impacts identified in Map 5.1. Likelihood estimates may then be assessed and judged for go/no-go decisions about deploying an AI system. If an organization decides to proceed with deploying the system, the likelihood and magnitude estimates can be used to assign TEVV resources appropriate for the risk level.

Suggested Actions
- Establish assessment scales for measuring AI systems’ impact. Scales may be qualitative, such as red-amber-green (RAG), or may entail simulations or econometric approaches. Document and apply scales uniformly across the organization’s AI portfolio.
- Apply TEVV regularly at key stages in the AI lifecycle, connected to system impacts and frequency of system updates.
- Identify and document likelihood and magnitude of system benefits and negative impacts in relation to trustworthiness characteristics.

Transparency & Documentation
Organizations can document the following:
- Which population(s) does the AI system impact?
- What assessments has the entity conducted on trustworthiness characteristics for example data security and privacy impacts associated with the AI system?
- Can the AI system be tested by independent third parties?
AI Transparency Resources

- Datasheets for Datasets.
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019. URL
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019. LINK

References


Artificial Intelligence Incident Database. 2022.


MAP 5.2
Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.

About
AI systems are socio-technical in nature and can have positive, neutral, or negative implications that extend beyond their stated purpose. Negative impacts can be wide-ranging and affect individuals, groups, communities, organizations, and society, as well as the environment and national security.

Organizations can create a baseline for system monitoring to increase opportunities for detecting emergent risks. After an AI system is deployed, engaging different stakeholder groups – who may be aware of, or experience, benefits or negative impacts that are unknown to AI actors involved in the design, development and deployment activities – allows organizations to understand and monitor system benefits and potential negative impacts more readily.

Suggested Actions
- Establish and document stakeholder engagement processes at the earliest stages of system formulation to identify potential impacts from the AI system on individuals, groups, communities, organizations, and society.
- Employ methods such as value sensitive design (VSD) to identify misalignments between organizational and societal values, and system implementation and impact.
- Identify approaches to engage, capture, and incorporate input from system end users and other key stakeholders to assist with continuous monitoring for potential impacts and emergent risks.
- Incorporate quantitative, qualitative, and mixed methods in the assessment and documentation of potential impacts to individuals, groups, communities, organizations, and society.
Identify a team (internal or external) that is independent of AI design and development functions to assess AI system benefits, positive and negative impacts and their likelihood and magnitude.

Evaluate and document stakeholder feedback to assess potential impacts for actionable insights regarding trustworthiness characteristics and changes in design approaches and principles.

Develop TEVV procedures that incorporate socio-technical elements and methods and plan to normalize across organizational culture. Regularly review and refine TEVV processes.

**Transparency & Documentation**

Organizations can document the following:

- If the AI system relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this managed?
- If the AI system relates to other ethically protected groups, have appropriate obligations been met? (e.g., medical data might include information collected from animals)
- If the AI system relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

**AI Transparency Resources**

- [Datasheets for Datasets](#).
- [GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities](#).
- [AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019](#).
- [Intel.gov: AI Ethics Framework for Intelligence Community - 2020](#).
- [Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019](#).
References


Microsoft Responsible AI Standard, v2.

Microsoft Research AI Fairness Checklist.

PEAT AI & Disability Inclusion Toolkit – Risks of Bias and Discrimination in AI Hiring Tools.
MEASURE

Identified risks are assessed, analyzed, or tracked
Measure 1
Appropriate methods and metrics are identified and applied.

Measure 1.1
Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.

About
The development and utility of trustworthy AI systems depends on reliable measurements and evaluations of underlying technologies and their use. Compared with traditional software systems, AI technologies bring new failure modes, inherent dependence on training data and methods which directly tie to data quality and representativeness. Additionally, AI systems are inherently socio-technical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks – and benefits – can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed. In other words, what should be measured depends on the purpose, audience, and needs of the evaluations.

These two factors influence selection of approaches and metrics for measurement of AI risks enumerated during the Map function. The AI landscape is evolving and so are the methods and metrics for AI measurement. The evolution of metrics is key to maintaining efficacy of the measures.

Suggested Actions
- Establish approaches for detecting, tracking and measuring known risks, errors, incidents or negative impacts.
- Identify testing procedures and metrics to demonstrate whether or not the system is fit for purpose and functioning as claimed.
- Identify testing procedures and metrics to demonstrate AI system trustworthiness.
• Define acceptable limits for system performance (e.g., distribution of errors) and include course correction suggestions if/when the system performs beyond acceptable limits.
• Define metrics for, and regularly assess, AI actor competency for effective system operation,
• Identify transparency metrics to assess whether stakeholders have access to necessary information about system design, development, deployment, use, and evaluation.
• Utilize accountability metrics to determine whether AI designers, developers, and deployers maintain clear and transparent lines of responsibility and are open to inquiries.
• Document metric selection criteria and include considered but unused metrics.
• Monitor AI system external inputs including training data, models developed for other contexts, system components reused from other contexts, and third-party tools and resources.
• Report metrics to inform assessments of system generalizability and reliability.
• Assess and document pre- vs post-deployment system performance. Include existing and emergent risks.
• Document risks or trustworthiness characteristics identified in the Map function that will not be measured, including justification for non-measurement.

Transparency & Documentation
Organizations can document the following:
• How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?
• What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
• Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)
• Did your organization address usability problems and test whether user interfaces served their intended purposes?
• What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e., manual vs automated, adversarial and stress testing)?

**AI Transparency Resources**
- [Artificial Intelligence Ethics Framework for the Intelligence Community.](https://www.jhu.edu/intelligence-ethics-framework/)
- [Datasheets for Datasets.](https://www.nist.gov/)

**References**


Measure 1.2
Appropriateness of AI metrics and effectiveness of existing controls is regularly assessed and updated including reports of errors and impacts on affected communities.

About
Different AI tasks, such as neural networks or natural language processing, benefit from different evaluation techniques. Use-case and particular settings in which the AI system is used also affects appropriateness of the evaluation techniques. Changes in the operational settings, data drift, model drift are among factors that suggest regularly assessing and updating appropriateness of AI metrics and their effectiveness can enhance reliability of AI system measurements.

Suggested Actions
- Assess external validity of all measurements (e.g., the degree to which measurements taken in one context can generalize to other contexts).
- Assess effectiveness of existing metrics and controls on a regular basis throughout the AI system lifecycle.
- Document reports of errors, incidents and negative impacts and assess sufficiency and efficacy of existing metrics for repairs, and upgrades.
- Develop new metrics when existing metrics are insufficient or ineffective for implementing repairs and upgrades.
- Develop and utilize metrics to monitor, characterize and track external inputs, including any third-party tools.
- Determine frequency and scope for sharing metrics and related information with stakeholders and impacted communities.
- Utilize stakeholder feedback processes established in the Map function to capture, act upon and share feedback from end users and potentially impacted communities.
- Collect and report software quality metrics such as rates of bug occurrence and severity, time to response, and time to repair (See Manage 4.3).
Transparency & Documentation
Organizations can document the following:

- What metrics has the entity developed to measure performance of the AI system?
- To what extent do the metrics provide accurate and useful measure of performance?
- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- How will the accuracy or appropriate performance metrics be assessed?
- What is the justification for the metrics selected?

AI Transparency Resources

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework for the Intelligence Community.

References


Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009.


Measure 1.3
Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.

About
The current AI systems are brittle, the failure modes are not well described, and the systems are dependent on the context in which they were developed and do not transfer well outside of the training environment. A reliance on local evaluations will be necessary along with a continuous monitoring of these systems. Measurements that extend beyond classical measures (which average across test cases) or expand to focus on pockets of failures where there are potentially significant costs can improve the reliability of risk management activities. Feedback from affected communities about how AI systems are being used can make AI evaluation purposeful. Involving internal experts who did not serve as front-line developers for the system and/or independent assessor’s regular assessments of AI systems helps a fulsome characterization of AI systems’ performance and trustworthiness.

Suggested Actions
- Evaluate TEVV processes regarding incentives to identify risks and impacts.
- Utilize separate testing teams established in the Govern function (2.1 and 4.1) to enable independent decisions and course-correction for AI systems. Track processes and measure and document change in performance.
- Plan and evaluate AI system prototypes with end user populations early and continuously in the AI lifecycle. Document test outcomes and course correct.
- Assess independence and stature of TEVV and oversight AI actors, to ensure they have the required levels of independence and resources to perform assurance, compliance, and feedback tasks effectively.
• Evaluate interdisciplinary and demographically diverse internal team established in Map 1.2.
• Evaluate effectiveness of external stakeholder feedback mechanisms, specifically related to processes for eliciting, evaluating and integrating input from diverse groups.
• Evaluate effectiveness of external stakeholder feedback mechanisms for enhancing AI actor visibility and decision-making regarding AI system risks and trustworthy characteristics.

**Transparency & Documentation**

Organizations can document the following:

• What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
• How easily accessible and current is the information available to external stakeholders?
• To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders?
• To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
• To what extent is this information sufficient and appropriate to promote transparency? Do external stakeholders have access to information on the design, operation, and limitations of the AI system?
• What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

**AI Transparency Resources**

• GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.
• Artificial Intelligence Ethics Framework for the Intelligence Community.
References


Measure 2
AI systems are evaluated for trustworthy characteristics.

Measure 2.1
Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.

About
Documenting measurement approaches, test sets, metrics, processes and materials used, and associated details builds foundation upon which to build a valid, reliable measurement process. Documentation enables repeatability and consistency, and can enhance AI risk management decisions.

Suggested Actions
- Leverage existing industry best practices for transparency and documentation of all possible aspects of measurements. Examples include: data sheet for data sets, model cards, [commenters provided examples].
- Regularly assess the effectiveness of tools used to document measurement approaches, test sets, metrics, processes and materials used.
- Update the tools as needed.

Transparency & Documentation
Organizations can document the following:
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- To what extent has the entity documented the AI system’s development, testing methodology, metrics, and performance outcomes?

AI Transparency Resources

References


Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting.” FAT *19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, 220–29.


Measure 2.2
Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.

About
Measurement and evaluation of AI systems often involves testing with human subjects or using data captured from human subjects. Protection of human subjects is required by law when carrying out federally funded research and is a domain specific requirement for some disciplines. Standard human subjects protection procedures include protecting the welfare and interests of human subjects, designing evaluations to minimize risks to subjects, and completion of mandatory training regarding legal requirements and expectations.

Evaluations of AI system performance that utilize human subjects or human subject data should reflect the population within the context of use. AI system activities utilizing non-representative data may lead to inaccurate assessments or negative and harmful outcomes. It is often difficult – and sometimes impossible, to collect data or perform evaluation tasks that reflect the full operational purview of an AI system. Methods for collecting, annotating, or using these data can also contribute to the challenge. To counteract these challenges, organizations can connect human subjects' data collection, and dataset practices, to AI system contexts and purposes and do so in close collaboration with AI Actors from the relevant domains.

Suggested Actions
- Follow human subjects research requirements as established by organizational and disciplinary requirements, including informed consent and compensation, during dataset collection activities.
- Analyze differences between intended and actual population of users or data subjects, including likelihood for errors, incidents or negative impacts.
- Utilize disaggregated evaluation methods (e.g., by race, age, gender, ethnicity, ability, region) to improve AI system performance when deployed in real world settings.
Establish thresholds and alert procedures for dataset representativeness within the context of use.

Construct datasets in close collaboration with experts with knowledge of the context of use.

Follow intellectual property and privacy rights related to datasets and their use, including for the subjects represented in the data.

Evaluate data representativeness through
  ○ investigating known failure modes,
  ○ assessing data quality and diverse sourcing,
  ○ applying public benchmarks,
  ○ traditional bias testing,
  ○ chaos engineering,
  ○ stakeholder feedback

Use informed consent for individuals providing data used in system testing and evaluation.

Transparency & Documentation

Organizations can document the following:

- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?
- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?
- To what extent has the entity identified and mitigated potential bias—statistical, contextual, and historical—in the data?
- If it relates to people, were they told what the dataset would be used for, and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?
- If human subjects were used in the development or testing of the AI system, what protections were put in place to promote their safety and wellbeing?
AI Transparency Resources

- Artificial Intelligence Ethics Framework for the Intelligence Community.
- Datasheets for Datasets.

References


Andrea Brennen, Ryan Ashley, Ricardo Calix, JJ Ben-Joseph, George Sieniawski, Mona Gogia, and BNH.AI. AI Assurance Audit of RoBERTa, an Open source, Pretrained Large Language Model. IQT Labs, December 2022.
**Measure 2.3**

AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.

**About**

The current risk and impact environment suggests AI system performance estimates are insufficient and require a deeper understanding of deployment context of use. Computationally focused performance testing and evaluation schemes are restricted to test data sets and in silico techniques. These approaches do not directly evaluate risks and impacts in real world environments and can only predict what might create impact based on an approximation of expected AI use. To properly manage risks, more direct information is necessary to understand how and under what conditions deployed AI creates impacts, who is most likely to be impacted, and what that experience is like.

**Suggested Actions**

- Conduct regular and sustained engagement with potentially impacted communities.
- Maintain a demographically diverse and multidisciplinary and collaborative internal team.
- Regularly test and evaluate systems in non-optimized conditions, and in collaboration with AI actors in user interaction and user experience (UI/UX) roles.
- Evaluate feedback from stakeholder engagement activities, in collaboration with human factors and socio-technical experts.
- Collaborate with socio-technical, human factors, and UI/UX experts to identify notable characteristics in context of use that can be translated into system testing scenarios.
- Measure AI systems prior to deployment in conditions similar to expected scenarios.
- Measure and document performance criteria such as validity (false positive rate, false negative rate, etc.) and efficiency (training times, prediction latency, etc.) related to ground truth within the deployment context of use.
• Measure assurance criteria such as AI actor competency and experience.
• Document differences between measurement setting and the deployment environment(s).

Transparency & Documentation
Organizations can document the following:
• What experiments were initially run on this dataset? To what extent have experiments on the AI system been documented?
• To what extent does the system/entity consistently measure progress towards stated goals and objectives?
• How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed? How much distributional shift or model drift from baseline performance is acceptable?
• As time passes and conditions change, is the training data still representative of the operational environment?
• What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e., adversarial or stress testing)?

AI Transparency Resources
• Artificial Intelligence Ethics Framework for the Intelligence Community.
• Datasheets for Datasets.

References
Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009.


**Software Resources**
- Drifter library (performance assessment)
- Manifold library (performance assessment)
- MLextend library (performance assessment)
- PiML library (explainable models, performance assessment)
- SALib library (performance assessment)
- What-If Tool (performance assessment)
Measure 2.4
The functionality and behavior of the AI system and its components – as identified in the MAP function – are monitored when in production.

About
AI systems may encounter new issues and risks while in production as the environment evolves over time. This effect, often referred to as “drift”, means AI systems no longer meet the assumptions and limitations of the original design. Regular monitoring allows AI Actors to monitor the functionality and behavior of the AI system and its components – as identified in the MAP function - and enhance the speed and efficacy of necessary system interventions.

Suggested Actions
- Monitor and document how metrics and performance indicators observed in production differ from the same metrics collected during pre-deployment testing. When differences are observed, consider error propagation and feedback loop risks.
- Utilize hypothesis testing or human domain expertise to measure monitored distribution differences in new input or output data relative to test environments.
- Monitor for anomalies using approaches such as control limits, confidence intervals, integrity constraints and ML algorithms. When anomalies are observed, consider error propagation and feedback loop risks.
- Verify alerts are in place for when distributions in new input data or generated predictions observed in production differ from pre-deployment test outcomes, or when anomalies are detected.
- Assess the accuracy and quality of generated outputs against new collected ground-truth information as it becomes available.
- Utilize human review to track processing of unexpected data and reliability of generated outputs; warn system users when outputs may be unreliable. Verify that human overseers responsible for these processes have clearly defined responsibilities and training for specified tasks.
Collect uses cases from the operational environment for system testing and monitoring activities in accordance with organizational policies and regulatory or disciplinary requirements (e.g., informed consent, institutional review board approval, human research protections),

**Transparency & Documentation**

Organizations can document the following:

- To what extent is the output of each component appropriate for the operational context?
- What justifications, if any, has the entity provided for the assumptions, boundaries, and limitations of the AI system?
- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?
- As time passes and conditions change, is the training data still representative of the operational environment?

**AI Transparency Resources**

- [Artificial Intelligence Ethics Framework for the Intelligence Community](https://www.dni.gov/nodate/cfai-framework).

**References**


Measure 2.5
The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.

About
An AI system that is not validated or that fails validation may be inaccurate or unreliable or may generalize poorly to data and settings beyond its training, creating and increasing AI risks and reducing trustworthiness. AI Actors can improve system validity by creating processes for exploring and documenting system limitations. This includes broad consideration of purposes and uses for which the system was not designed.

Validation risks include the use of proxies or other indicators that are often constructed by AI development teams to operationalize phenomena that are either not directly observable or measurable (e.g., fairness, hireability, honesty, propensity to commit a crime). Teams can mitigate these risks by demonstrating that the indicator is measuring the concept it claims to measure (also known as construct validity). Without this and other types of validation, various negative properties or impacts may go undetected, including the presence of confounding variables, potential spurious correlations, or error propagation and its potential impact on other interconnected systems.

Suggested Actions
- Define the operating conditions and socio-technical context under which the AI system will be validated.
- Define and document processes to establish the system’s operational conditions and limits.
- Establish or identify, and document approaches to measure forms of validity, including:
  - construct validity (the test is measuring the concept it claims to measure)
  - internal validity (relationship being tested is not influenced by other factors or variables)
- external validity (results are generalizable beyond the training condition)
- the use of experimental design principles and statistical analyses and modeling.

- Assess and document system variance. Standard approaches include confidence intervals, standard deviation, standard error, bootstrapping, or cross-validation.
- Establish or identify and document robustness measures.
- Establish or identify and document reliability measures.
- Establish practices to specify and document the assumptions underlying measurement models to ensure proxies accurately reflect the concept being measured.
- Utilize standard software testing approaches (e.g., unit, integration, functional and chaos testing, computer-generated test cases, etc.)
- Utilize standard statistical methods to test bias, inferential associations, correlation, and covariance in adopted measurement models.
- Utilize standard statistical methods to test variance and reliability of system outcomes.
- Monitor operating conditions for system performance outside of defined limits.
- Identify TEVV approaches for exploring AI system limitations, including testing scenarios that differ from the operational environment. Consult experts with knowledge of specific context of use.
- Define post-alert actions. Possible actions may include:
  - alerting other relevant AI actors before action,
  - requesting subsequent human review of action,
  - alerting downstream users and stakeholder that the system is operating outside its defined validity limits,
  - tracking and mitigating possible error propagation
  - action logging
- Log input data and relevant system configuration information whenever there is an attempt to use the system beyond its well-defined range of system validity.
• Modify the system over time to extend its range of system validity to new operating conditions.

**Transparency & Documentation**

Organizations can document the following:

• What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e., adversarial or stress testing)?
• Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
• How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?
• To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?
• What goals and objectives does the entity expect to achieve by designing, developing, and/or deploying the AI system?

**AI Transparency Resources**


**References**


Software Resources

- Drifter library (performance assessment)
- Manifold library (performance assessment)
- MLextend library (performance assessment)
- PiML library (explainable models, performance assessment)
- SALib library (performance assessment)
- What-If Tool (performance assessment)
Measure 2.6
AI system is evaluated regularly for safety risks – as identified in the MAP function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics implicate system reliability and robustness, real-time monitoring, and response times for AI system failures.

About
Many AI systems are being introduced into settings such as transportation, manufacturing or security, where failures may give rise to various physical or environmental harms. AI systems that may endanger human life, health, property or the environment are tested thoroughly prior to deployment, and are regularly evaluated to confirm the system is safe during normal operations, and in settings beyond its proposed use and knowledge limits.

Measuring activities for safety often relate to exhaustive testing in development and deployment contexts, understanding the limits of a system’s reliable, robust, and safe behavior, and real-time monitoring of various aspects of system performance. These activities are typically conducted along with other risk mapping, management, and governance tasks such as avoiding past failed designs, establishing and rehearsing incident response plans that enable quick responses to system problems, the instantiation of redundant functionality to cover failures, and transparent and accountable governance. System safety incidents or failures are frequently reported to be related to organizational dynamics and culture. Independent auditors may bring important independent perspectives for reviewing evidence of AI system safety.

Suggested Actions
- Thoroughly measure system performance in development and deployment contexts, and under stress conditions.
  - Employ test data assessments and simulations before proceeding to production testing. Track multiple performance quality and error metrics.
- Stress-test system performance under likely scenarios (e.g., concept drift, high load) and beyond known limitations, in consultation with domain experts.
- Test the system under conditions similar to those related to past known incidents or near-misses and measure system performance and safety characteristics.
- Apply chaos engineering approaches to test systems in extreme conditions and gauge unexpected responses.
- Document the range of conditions under which the system has been tested and demonstrated to fail safely.

- Measure and monitor system performance in real-time to enable rapid response when AI system incidents are detected.
- Collect pertinent safety statistics (e.g., out-of-range performance, incident response times, system down time, injuries, etc.) in anticipation of potential information sharing with impacted communities or as required by AI system oversight personnel.
- Align measurement to the goal of continuous improvement. Seek to increase the range of conditions under which the system is able to fail safely through system modifications in response to in-production testing and events.
- Document, practice and measure incident response plans for AI system incidents, including measuring response and down times.
- Compare documented safety testing and monitoring information with established risk tolerances on an on-going basis.
- Consult MANAGE for detailed information related to managing safety risks.

Transparency & Documentation
Organizations can document the following:
- What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e., adversarial or stress testing)?
- To what extent has the entity documented the AI system’s development, testing methodology, metrics, and performance outcomes?
Did you establish mechanisms that facilitate the AI system’s auditability (e.g., traceability of the development process, the sourcing of training data and the logging of the AI system’s processes, outcomes, positive and negative impact)?

Did you ensure that the AI system can be audited by independent third parties?

Did you establish a process for third parties (e.g., suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?

**AI Transparency Resources**

- [GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.](#)
- [Artificial Intelligence Ethics Framework for the Intelligence Community.](#)

**References**

- [AI Incident Database. 2022.](#)
- [AIAAIC Repository. 2022.](#)
- [Netflix. Chaos Monkey.](#)
- [IBM. “IBM's Principles of Chaos Engineering,” IBM, n.d.](#)
- [Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub.](#)
Measure 2.7
AI system security and resilience – as identified in the MAP function – are evaluated and documented.

About
AI systems, as well as the ecosystems in which they are deployed, may be said to be resilient if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary. Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data, or other intellectual property through AI system endpoints. 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 and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an unexpected adverse event, security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Resilience relates to robustness and encompasses unexpected or adversarial use (or abuse or misuse) of the model or data.

Suggested Actions
- Establish and track AI system security tests and metrics (e.g., red-teaming activities, frequency and rate of anomalous events, system down-time, incident response times, time-to-bypass, etc.).
- Use red-team exercises to actively test the system under adversarial or stress conditions, measure system response, assess failure modes or determine if system can return to normal function after an unexpected adverse event.
- Document red-team exercise results as part of continuous improvement efforts, including the range of security test conditions and results.
• Use countermeasures (e.g., authentication, throttling, differential privacy, robust ML approaches) to increase the range of security conditions under which the system is able to return to normal function.

• Modify system security procedures and countermeasures to increase robustness and resilience to attacks in response to testing and events experienced in production.

• Verify that information about errors and attack patterns is shared with incident databases, other organizations with similar systems, and system users and stakeholders (MANAGE-4.1).

• Develop and maintain information sharing practices with AI actors from other organizations to learn from common attacks.

• Verify that third party AI resources and personnel undergo security audits and screenings. Risk indicators may include failure of third parties to provide relevant security information.

• Utilize watermarking technologies as a deterrent to data and model extraction attacks.

**Transparency & Documentation**

Organizations can document the following:

• To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?

• What assessments has the entity conducted on data security and privacy impacts associated with the AI system?

• What processes exist for data generation, acquisition/collection, security, maintenance, and dissemination?

• What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e., adversarial or stress testing)?

• If a third party created the AI, how will you ensure a level of explainability or interpretability?
AI Transparency Resources

- **GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.**
- **Artificial Intelligence Ethics Framework for the Intelligence Community.**

References


Software Resources

- **adversarial-robustness-toolbox**
- **counterfit**
- **foolbox**
- **ml_privacy_meter**
- **robustness**
- **tensorflow/privacy**
Measure 2.8
Risks associated with transparency and accountability – as identified in the MAP function – are examined and documented.

About
Transparency enables meaningful visibility into entire AI pipelines, workflows, processes or organizations and decreases information asymmetry between AI developers and operators and other AI Actors and impacted communities. Transparency is a central element of effective AI risk management that enables insight into how an AI system is working, and the ability to address risks if and when they emerge. The ability for system users, individuals, or impacted communities to seek redress for incorrect or problematic AI system outcomes is one control for transparency and accountability. Higher level recourse processes are typically enabled by lower-level implementation efforts directed at explainability and interpretability functionality. See Measure 2.9.

Transparency and accountability across organizations and processes is crucial to reducing AI risks. Accountable leadership – whether individuals or groups – and transparent roles, responsibilities, and lines of communication foster and incentivize quality assurance and risk management activities within organizations.

Lack of transparency complicates measurement of trustworthiness and whether AI systems or organizations are subject to effects of various individual and group biases and design blind spots and could lead to diminished user, organizational and community trust, and decreased overall system value. Enstating accountable and transparent organizational structures along with documenting system risks can enable system improvement and risk management efforts, allowing AI actors along the lifecycle to identify errors, suggest improvements, and figure out new ways to contextualize and generalize AI system features and outcomes.

Suggested Actions
• Instrument the system for measurement and tracking, e.g., by maintaining histories, audit logs and other information that can be used by AI actors to review and evaluate possible sources of error, bias, or vulnerability.
• Calibrate controls for users in close collaboration with experts in user interaction and user experience (UI/UX), human computer interaction (HCI), and/or human-AI teaming.

• Test provided explanations for calibration with different audiences including operators, end users, decision makers and decision subjects (individuals for whom decisions are being made), and to enable recourse for consequential system decisions that affect end users or subjects.

• Measure and document human oversight of AI systems:
  ○ Document the degree of oversight that is provided by specified AI actors regarding AI system output.
  ○ Maintain statistics about downstream actions by end users and operators such as system overrides.
  ○ Maintain statistics about and document reported errors or complaints, time to respond, and response types.
  ○ Maintain and report statistics about adjudication activities.

• Track, document, and measure organizational accountability regarding AI systems via policy exceptions and escalations, and document “go” and “no/go” decisions made by accountable parties.

• Track and audit the effectiveness of organizational mechanisms related to AI risk management, including:
  ○ Lines of communication between AI actors, executive leadership, users and impacted communities.
  ○ Roles and responsibilities for AI actors and executive leadership.
  ○ Organizational accountability roles, e.g., chief model risk officers, AI oversight committees, responsible or ethical AI directors, etc.

Transparency & Documentation

Organizations can document the following:

• To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?

• What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
Who is accountable for the ethical considerations during all stages of the AI lifecycle?
Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
Are the responsibilities of the personnel involved in the various AI governance processes clearly defined?

AI Transparency Resources

- Artificial Intelligence Ethics Framework for the Intelligence Community.

References


Measure 2.9
The AI model is explained, validated, and documented, and an AI system output is interpreted within its context – as identified in the MAP function – and to inform responsible use and governance.

About
Explainability and interpretability assist those operating or overseeing an AI system, as well as users of an AI system, to gain deeper insights into the functionality and trustworthiness of the system, including its outputs. Explainable and interpretable AI systems offer information that help end users understand the purposes and potential impact of an AI system. Risk from lack of explainability may be managed by describing how AI systems function, with descriptions tailored to individual differences such as the user’s role, knowledge, and skill level. Explainable systems can be debugged and monitored more easily, and they lend themselves to more thorough documentation, audit, and governance.

Risks to interpretability often can be addressed by communicating a description of why an AI system made a particular prediction or recommendation.

Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of “what happened”. Explainability can answer the question of “how” a decision was made in the system. Interpretability can answer the question of “why” a decision was made by the system and its meaning or context to the user.

Suggested Actions
• Verify systems are developed to produce explainable models, post-hoc explanations and audit logs.
• When possible or available, utilize approaches that are inherently explainable, such as traditional and penalized generalized linear models, decision trees, nearest-neighbor and prototype-based approaches, rule-based models, generalized additive models, explainable boosting machines and neural additive models.
Test explanation methods and resulting explanations prior to deployment to gain feedback from relevant AI actors, end users, and potentially impacted individuals or groups about whether explanations are accurate, clear, and understandable.

Document AI model details including model type (e.g., convolutional neural network, reinforcement learning, decision tree, random forest, etc.) data features, training algorithms, proposed uses, decision thresholds, training data, evaluation data, and ethical considerations.

Establish, document, and report performance and error metrics across demographic groups and other segments relevant to the deployment context.

Explain systems using a variety of methods, e.g., visualizations, model extraction, feature importance, and others. Since explanations may not accurately summarize complex systems, test explanations according to properties such as fidelity, consistency, robustness, and interpretability.

Assess the characteristics of system explanations according to properties such as fidelity (local and global), ambiguity, interpretability, interactivity, consistency, and resilience to attack/manipulation.

Test the quality of system explanations with end-users and other groups.

Secure model development processes to avoid vulnerability to external manipulation such as gaming explanation processes.

Test for changes in models over time, including for models that adjust in response to production data.

Use transparency tools such as data statements and model cards to document explanatory and validation information.

**Transparency & Documentation**

Organizations can document the following:

- Given the purpose of the AI, what level of explainability or interpretability is required for how the AI made its determination?
• Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
• How has the entity documented the AI system’s data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata?
• What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

AI Transparency Resources
• GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.
• Artificial Intelligence Ethics Framework for the Intelligence Community.

References


Software Resources

- **SHAP**
- **LIME**
- **Interpret**
- **PiML**
- **Iml**
- **Dalex**
Measure 2.10
Privacy risk of the AI system – as identified in the MAP function – is examined and documented.

About
Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals’ agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation).

Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics. Like safety and security, specific technical features of an AI system may promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals.

Privacy-enhancing technologies ("PETs") for AI, as well as data minimizing methods such as de-identification and aggregation for certain model outputs, can support design for privacy-enhanced AI systems. Under certain conditions such as data sparsity, privacy enhancing techniques can result in a loss in accuracy, impacting decisions about fairness and other values in certain domains.

Suggested Actions
- Specify privacy-related values, frameworks, and attributes that are applicable in the context of use through direct engagement with end users and potentially impacted groups and communities.
- Document collection, use, management, and disclosure of personally sensitive information in datasets, in accordance with privacy and data governance policies.
- Quantify privacy-level data aspects such as the ability to identify individuals or groups (e.g., k-anonymity metrics, l-diversity, t-closeness).
- Establish and document protocols (authorization, duration, type) and access controls for training sets or production data containing personally sensitive information, in accordance with privacy and data governance policies.
- Monitor internal queries to production data for detecting patterns that isolate personal records.
- Monitor PSI disclosures and inference of sensitive or legally protected attributes:
  - Assess the risk of manipulation from overly customized content. Evaluate information presented to representative users at various points along axes of difference between individuals (e.g., individuals of different ages, genders, races, political affiliation, etc.).
- Use privacy-enhancing techniques such as differential privacy, when publicly sharing dataset information.
- Collaborate with privacy experts, AI end users and operators, and other domain experts to determine optimal differential privacy metrics within contexts of use.

**Transparency & Documentation**

**Organizations can document the following:**
- Did your organization implement accountability-based practices in data management and protection (e.g., the PDPA and OECD Privacy Principles)?
- What assessments has the entity conducted on data security and privacy impacts associated with the AI system?
- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?
- Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)
- If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial, social or otherwise) What was done to mitigate or reduce the potential for harm?
AI Transparency Resources

- Datasheets for Datasets.

References


Claire McKay Bowen. "Utility Metrics for Differential Privacy: No One-Size-Fits-All." National Institute of Standards and Technology (NIST), November 29, 2021. URL


Measure 2.11
Fairness and bias – as identified in the MAP function – is evaluated and results are documented.

About
Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Organizations’ risk management efforts will be enhanced by recognizing and considering these differences. Systems in which harmful biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases.

Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent.

- Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems.
- Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples.
- Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.

Bias exists in many forms and can become ingrained in the automated systems that help make decisions about our lives. While bias is not always a negative phenomenon, AI systems can potentially increase the speed and scale of biases and perpetuate and amplify harms to individuals, groups, communities, organizations, and society.
Suggested Actions

- Conduct fairness assessments to manage computational and statistical forms of bias which include the following steps:
  - Identify types of harms, including allocational, representational, quality of service, stereotyping, or erasure.
  - Identify across, within, and intersecting groups that might be harmed.
  - Quantify harms using both a general fairness metric, if appropriate (e.g., demographic parity, equalized odds, equal opportunity, statistical hypothesis tests), and custom, context-specific metrics developed in collaboration with affected communities.
  - Analyze quantified harms for contextually significant differences across groups, within groups, and among intersecting groups.
  - Refine identification of within-group and intersectional group disparities.
    - Evaluate underlying data distributions and employ sensitivity analysis during the analysis of quantified harms.
    - Evaluate quality metrics including false positive rates and false negative rates.
    - Consider biases affecting small groups, within-group or intersectional communities, or single individuals.

- Understand and consider sources of bias in training and TEVV data:
  - Differences in distributions of outcomes across and within groups, including intersecting groups.
  - Completeness, representativeness and balance of data sources.
  - Identify input data features that may serve as proxies for demographic group membership (i.e., credit score, ZIP code) or otherwise give rise to emergent bias within AI systems.
  - Forms of systemic bias in images, text (or word embeddings), audio or other complex or unstructured data.

- Leverage impact assessments to identify and classify system impacts and harms to end users, other individuals, and groups with input from potentially impacted communities.
- Identify the classes of individuals, groups, or environmental ecosystems which might be impacted through direct engagement with potentially impacted communities.
- Evaluate systems in regard to disability inclusion, including consideration of disability status in bias testing, and discriminatory screen out processes that may arise from non-inclusive design or deployment decisions.
- Develop objective functions in consideration of systemic biases, in-group/out-group dynamics.
- Use context-specific fairness metrics to examine how system performance varies across groups, within groups, and/or for intersecting groups. Metrics may include statistical parity, error-rate equality, statistical parity difference, equal opportunity difference, average absolute odds difference, standardized mean difference, percentage point differences.
- Customize fairness metrics to specific context of use to examine how system performance and potential harms vary within contextual norms.
- Define acceptable levels of difference in performance in accordance with established organizational governance policies, business requirements, regulatory compliance, legal frameworks, and ethical standards within the context of use.
- Define the actions to be taken if disparity levels rise above acceptable levels.
- Identify groups within the expected population that may require disaggregated analysis, in collaboration with impacted communities.
- Leverage experts with knowledge in the specific context of use to investigate substantial measurement differences and identify root causes for those differences.
- Monitor system outputs for performance or bias issues that exceed established tolerance levels.
- Ensure periodic model updates; test and recalibrate with updated and more representative data to stay within acceptable levels of difference.
• Apply pre-processing data transformations to address factors related to demographic balance and data representativeness.
• Apply in-processing to balance model performance quality with bias considerations.
• Apply post-processing mathematical/computational techniques to model results in close collaboration with impact assessors, socio-technical experts, and other AI actors with expertise in the context of use.
• Apply model selection approaches with transparent and deliberate consideration of bias management and other trustworthy characteristics.
• Collect and share information about differences in outcomes for the identified groups.
• Consider mediations to mitigate differences, especially those that can be traced to past patterns of unfair or biased human decision-making.
• Utilize human-centered design practices to generate deeper focus on societal impacts and counter human-cognitive biases within the AI lifecycle.
• Evaluate practices along the lifecycle to identify potential sources of human-cognitive bias such as availability, observational, and confirmation bias, and to make implicit decision-making processes more explicit and open to investigation.
• Work with human factors experts to evaluate biases in the presentation of system output to end users, operators and practitioners.
• Utilize processes to enhance contextual awareness, such as diverse internal staff and stakeholder engagement.

**Transparency & Documentation**

**Organizations can document the following:**

• To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?
• If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?
• Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
• How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?
• To what extent has the entity identified and mitigated potential bias—statistical, contextual, and historical—in the data?
• Were adversarial machine learning approaches considered or used for measuring bias (e.g.: prompt engineering, adversarial models)

AI Transparency Resources
• GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.
• Artificial Intelligence Ethics Framework for the Intelligence Community.
• Datasheets for Datasets.

References


Microsoft Research. “AI Fairness Checklist.” Microsoft, February 7, 2022. URL


Ben Green. “Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness.” Philosophy and Technology 35, no. 90 (October 8, 2022)


Fairlearn. “Fairness in Machine Learning.” Fairlearn 0.8.0 Documentation, n.d. URL


Software Resources
- aequitas
- AI Fairness 360:
  - Python
  - R
- algofairness
- fairlearn
- fairml
- fairmodels
- fairness
- solas-ai-disparity
- tensorflow/fairness-indicators
- Themis
Measure 2.12
Environmental impact and sustainability of AI model training and management activities – as identified in the MAP function – are assessed and documented.

About
Large-scale, high-performance computational resources used by AI systems for training and operation can contribute to environmental impacts. Direct negative impacts to the environment from these processes are related to energy consumption, water consumption, and greenhouse gas (GHG) emissions. The OECD has identified metrics for each type of negative direct impact.

Indirect negative impacts to the environment reflect the complexity of interactions between human behavior, socio-economic systems, and the environment and can include induced consumption and “rebound effects”, where efficiency gains are offset by accelerated resource consumption.

Other AI related environmental impacts can arise from the production of computational equipment and networks (e.g., mining and extraction of raw materials), transporting hardware, and electronic waste recycling or disposal.

Suggested Actions
- Include environmental impact indicators in AI system design and development plans, including reducing consumption and improving efficiencies.
- Identify and implement key indicators of AI system energy and water consumption and efficiency, and/or GHG emissions.
- Establish measurable baselines for sustainable AI system operation in accordance with organizational policies, regulatory compliance, legal frameworks, and environmental protection and sustainability norms.
- Assess tradeoffs between AI system performance and sustainable operations in accordance with organizational principles and policies, regulatory compliance, legal frameworks, and environmental protection and sustainability norms.
• Identify and establish acceptable resource consumption and efficiency, and GHG emissions levels, along with actions to be taken if indicators rise above acceptable levels.
• Estimate AI system emissions levels throughout the AI lifecycle via carbon calculators or similar process.

Transparency & Documentation

Organizations can document the following:
• Are greenhouse gas emissions, and energy and water consumption and efficiency tracked within the organization?
• Are deployed AI systems evaluated for potential upstream and downstream environmental impacts (e.g., increased consumption, increased emissions, etc.)?
• Could deployed AI systems cause environmental incidents, e.g., air or water pollution incidents, toxic spills, fires or explosions?

AI Transparency Resources
• GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
• Artificial Intelligence Ethics Framework for the Intelligence Community.
• Datasheets for Datasets. URL

References


Measure 2.13
Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated and documented.

About
The development of metrics is a process often considered to be objective but, as a human and organization driven endeavor, can reflect implicit and systemic biases, and may inadvertently reflect factors unrelated to the target function. Measurement approaches can be oversimplified, gamed, lack critical nuance, become used and relied upon in unexpected ways, fail to account for differences in affected groups and contexts.

Revisiting the metrics chosen in Measure 2.1 through 2.12 in a process of continual improvement can help AI actors to evaluate and document metric effectiveness and make necessary course corrections.

Suggested Actions
- Review selected system metrics and associated TEVV processes to determine if they are able to sustain system improvements, including the identification and removal of errors.
- Regularly evaluate system metrics for utility and consider descriptive approaches in place of overly complex methods.
- Review selected system metrics for acceptability within the end user and impacted community of interest.
- Assess effectiveness of metrics for identifying and measuring risks.

Transparency & Documentation
Organizations can document the following:
- To what extent does the system/entity consistently measure progress towards stated goals and objectives?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- To what extent are the model outputs consistent with the entity’s values and principles to foster public trust and equity?
• How will the accuracy or appropriate performance metrics be assessed?

**AI Transparency Resources**

- [GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.](#)
- [Artificial Intelligence Ethics Framework for the Intelligence Community](#)

**References**


Measure 3
Mechanisms for tracking identified AI risks over time are in place.

Measure 3.1
Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.

About
For trustworthy AI systems, regular system monitoring is carried out in accordance with organizational governance policies, AI actor roles and responsibilities, and within a culture of continual improvement. If and when emergent or complex risks arise, it may be necessary to adapt internal risk management procedures, such as regular monitoring, to stay on course. Documentation, resources, and training are part of an overall strategy to support AI actors as they investigate and respond to AI system errors, incidents or negative impacts.

Suggested Actions

- Compare AI system risks with:
  - simpler or traditional models
  - human baseline performance
  - other manual performance benchmarks
- Compare end user and community feedback about deployed AI systems to internal measures of system performance.
- Assess effectiveness of metrics for identifying and measuring emergent risks.
- Measure error response times and track response quality.
- Elicit and track feedback from AI actors in user support roles about the type of metrics, explanations and other system information required for fulsome resolution of system issues. Consider:
  - Instances where explanations are insufficient for investigating possible error sources or identifying responses.
  - System metrics, including system logs and explanations, for identifying and diagnosing sources of system error.
• Elicit and track feedback from AI actors in incident response and support roles about the adequacy of staffing and resources to perform their duties in an effective and timely manner.

**Transparency & Documentation**

**Organizations can document the following:**

• Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?

• To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?

• What metrics has the entity developed to measure performance of the AI system, including error logging?

• To what extent do the metrics provide accurate and useful measure of performance?

**AI Transparency Resources**


• [Artificial Intelligence Ethics Framework for the Intelligence Community.](https://www.saptapadi.org/aiethicsframework)


**References**


Measure 3.2
Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.

About
Risks identified in the Map function may be complex, emerge over time, or difficult to measure. Systematic methods for risk tracking, including novel measurement approaches, can be established as part of regular monitoring and improvement processes.

Suggested Actions
- Establish processes for tracking emergent risks that may not be measurable with current approaches. Some processes may include:
  - Recourse mechanisms for faulty AI system outputs.
  - Bug bounties.
  - Human-centered design approaches.
  - User-interaction and experience research.
  - Participatory stakeholder engagement with affected or potentially impacted individuals and communities.
- Identify AI actors responsible for tracking emergent risks and inventory methods.
- Determine and document the rate of occurrence and severity level for complex or difficult-to-measure risks when:
  - Prioritizing new measurement approaches for deployment tasks.
  - Allocating AI system risk management resources.
  - Evaluating AI system improvements.
  - Making go/no-go decisions for subsequent system iterations.

Transparency & Documentation
Organizations can document the following:
- Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?
- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
• To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders?
• Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
• If anyone believes that the AI no longer meets this ethical framework, who will be responsible for receiving the concern and as appropriate investigating and remediating the issue? Do they have authority to modify, limit, or stop the use of the AI?

AI Transparency Resources
• GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.
• Artificial Intelligence Ethics Framework for the Intelligence Community.

References


Measure 3.3
Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.

About
Assessing impact is a two-way effort. Many AI system outcomes and impacts may not be visible or recognizable to AI actors across the development and deployment dimensions of the AI lifecycle, and may require direct feedback about system outcomes from the perspective of end users and impacted groups.

Feedback can be collected indirectly, via systems that are mechanized to collect errors and other feedback from end users and operators.

Metrics and insights developed in this sub-category feed into Manage 4.1 and 4.2.

Suggested Actions
- Measure efficacy of end user and operator error reporting processes.
- Categorize and analyze type and rate of end user appeal requests and results.
- Measure feedback activity participation rates and awareness of feedback activity availability.
- Utilize feedback to analyze measurement approaches and determine subsequent courses of action.
- Evaluate measurement approaches to determine efficacy for enhancing organizational understanding of real-world impacts.
- Analyze end user and community feedback in close collaboration with domain experts.

Transparency & Documentation
Organizations can document the following:
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- Did your organization address usability problems and test whether user interfaces served their intended purposes?
• How easily accessible and current is the information available to external stakeholders?
• What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

**AI Transparency Resources**

- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations

**References**


Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022


Measure 4
Feedback about efficacy of measurement is gathered and assessed.

Measure 4.1
Measurement approaches for identifying AI risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.

About
AI Actors carrying out TEVV tasks may have difficulty evaluating impacts within the system context of use. AI system risks and impacts are often best described by end users and others who may be affected by output and subsequent decisions. AI Actors can elicit feedback from impacted individuals and communities via participatory engagement processes established in Govern 5.1 and 5.2, and carried out in Map 1.6, 5.1, and 5.2.

Activities described in the Measure function enable AI actors to evaluate feedback from impacted individuals and communities. To increase awareness of insights, feedback can be evaluated in close collaboration with AI actors responsible for impact assessment, human-factors, and governance and oversight tasks, as well as with other socio-technical domain experts and researchers. To gain broader expertise for interpreting evaluation outcomes, organizations may consider collaborating with advocacy groups and civil society organizations.

Insights based on this type of analysis can inform TEVV-based decisions about metrics and related courses of action.

Suggested Actions
- Support mechanisms for capturing feedback from system end users (including domain experts, operators, and practitioners). Successful approaches are:
  - conducted in settings where end users are able to openly share their doubts and insights about AI system output, and in connection to their specific context of use (including setting and task-specific lines of inquiry)
developed and implemented by human-factors and socio-technical domain experts and researchers
- designed to ensure control of interviewer and end user subjectivity and biases

- Identify and document approaches:
  - for evaluating and integrating elicited feedback from system end users
  - in collaboration with human-factors and socio-technical domain experts,
  - to actively inform a process of continual improvement.

- Evaluate feedback from end users alongside evaluated feedback from impacted communities (MEASURE 3.3).

- Utilize end user feedback to investigate how selected metrics and measurement approaches interact with organizational and operational contexts.

- Analyze and document system-internal measurement processes in comparison to collected end user feedback.

- Identify and implement approaches to measure effectiveness and satisfaction with end user elicitation techniques, and document results.

Transparency & Documentation
Organizations can document the following:

- Did your organization address usability problems and test whether user interfaces served their intended purposes?
- How will user and peer engagement be integrated into the model development process and periodic performance review once deployed?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?

AI Transparency Resources
• Artificial Intelligence Ethics Framework for the Intelligence Community.
• WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations

References


Measure 4.2

Measurement results regarding AI system trustworthiness in deployment context(s) and across AI lifecycle are informed by input from domain experts and other relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.

About

Feedback captured from relevant AI Actors can be evaluated in combination with output from Measure 2.5 to 2.11 to determine if the AI system is performing within pre-defined operational limits for validity and reliability, safety, security and resilience, privacy, bias and fairness, explainability and interpretability, and transparency and accountability. This feedback provides an additional layer of insight about AI system performance, including potential misuse or reuse outside of intended settings.

Insights based on this type of analysis can inform TEVV-based decisions about metrics and related courses of action.

Suggested Actions

- Integrate feedback from end users, operators, and affected individuals and communities from Map function as inputs to assess AI system trustworthiness characteristics. Ensure both positive and negative feedback is being assessed.
- Evaluate feedback in connection with AI system trustworthiness characteristics from Measure 2.5 to 2.11.
- Evaluate feedback regarding end user satisfaction with, and confidence in, AI system performance including whether output is considered valid and reliable, and explainable and interpretable.
- Identify mechanisms to confirm/support AI system output (e.g., recommendations), and end user perspectives about that output.
- Measure frequency of AI systems’ override decisions, evaluate and document results, and feed insights back into continual improvement processes.
- Consult AI actors in impact assessment, human factors, and socio-technical tasks to assist with analysis and interpretation of results.
To what extent does the system/entity consistently measure progress towards stated goals and objectives?

What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles?

To what extent are the model outputs consistent with the entity’s values and principles to foster public trust and equity?

Given the purpose of the AI, what level of explainability or interpretability is required for how the AI made its determination?

To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?

**AI Transparency Resources**

- Artificial Intelligence Ethics Framework for the Intelligence Community.

**References**


Measure 4.3
Measurable performance improvements or declines based on consultations with relevant AI actors including affected communities, and field data about context-relevant risks and trustworthiness characteristics, are identified and documented.

About
TEVV activities conducted throughout the AI system lifecycle can provide baseline quantitative measures for trustworthy characteristics. When combined with results from Measure 2.5 to 2.11 and Measure 4.1 and 4.2, TEVV actors can maintain a comprehensive view of system performance. These measures can be augmented through participatory engagement with potentially impacted communities or other forms of stakeholder elicitation about AI systems’ impacts. These sources of information can allow AI actors to explore potential adjustments to system components, adapt operating conditions, or institute performance improvements.

Suggested Actions
- Develop baseline quantitative measures for trustworthy characteristics.
- Delimit and characterize baseline operation values and states.
- Utilize qualitative approaches to augment and complement quantitative baseline measures, in close coordination with impact assessment, human factors and socio-technical AI actors.
- Monitor and assess measurements as part of continual improvement to identify potential system adjustments or modifications.
- Perform and document sensitivity analysis to characterize actual and expected variance in performance after applying system or procedural updates.
- Document decisions related to the sensitivity analysis and record expected influence on system performance and identified risks.

Transparency & Documentation
Organizations can document the following:
- To what extent are the model outputs consistent with the entity’s values and principles to foster public trust and equity?
• How were sensitive variables (e.g., demographic and socioeconomic categories) that may be subject to regulatory compliance specifically selected or not selected for modeling purposes?
• Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?
• How will the accountable human(s) address changes in accuracy and precision due to either an adversary’s attempts to disrupt the AI or unrelated changes in the operational/business environment?
• How will user and peer engagement be integrated into the model development process and periodic performance review once deployed?

**AI Transparency Resources**

- [GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.](#)
- [Artificial Intelligence Ethics Framework for the Intelligence Community.](#)

**References**


MANAGE

Manage
Risks are prioritized and acted upon based on a projected impact
Manage 1
AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.

Manage 1.1
A determination is made as to whether the AI system achieves its intended purpose and stated objectives and whether its development or deployment should proceed.

About
AI systems may not necessarily be the right solution for a given business task or problem. A standard risk management practice is to formally weigh an AI system’s negative risks against its benefits, and to determine if the AI system is an appropriate solution. Tradeoffs among trustworthiness characteristics—such as deciding to deploy a system based on system performance vs system transparency—may require regular assessment throughout the AI lifecycle.

Suggested Actions
- Consider trustworthiness characteristics when evaluating AI systems’ negative risks and benefits.
- Utilize TEVV outputs from map and measure functions when considering risk treatment.
- Regularly track and monitor negative risks and benefits throughout the AI system lifecycle including in post-deployment monitoring.
- Regularly assess and document system performance relative to trustworthiness characteristics and tradeoffs between negative risks and opportunities.
- Evaluate tradeoffs in connection with real-world use cases and impacts and as enumerated in Map function outcomes.

Transparency & Documentation
Organizations can document the following:
- How do the technical specifications and requirements align with the AI system’s goals and objectives?
To what extent are the metrics consistent with system goals, objectives, and constraints, including ethical and compliance considerations?

What goals and objectives does the entity expect to achieve by designing, developing, and/or deploying the AI system?

AI Transparency Resources

- Artificial Intelligence Ethics Framework for the Intelligence Community.
- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations

References


Manage 1.2
Treatment of documented AI risks is prioritized based on impact, likelihood, or available resources or methods.

About
Risk refers to the composite measure of an event’s probability of occurring and the magnitude (or degree) of the consequences of the corresponding events. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or risks.

Organizational risk tolerances are often informed by several internal and external factors, including existing industry practices, organizational values, and legal or regulatory requirements. Since risk management resources are often limited, organizations usually assign them based on risk tolerance. AI risks that are deemed more serious receive more oversight attention and risk management resources.

Suggested Actions
- Assign risk management resources relative to established risk tolerance. AI systems with lower risk tolerances receive greater oversight, mitigation and management resources.
- Document AI risk tolerance determination practices and resource decisions.
- Regularly review risk tolerances and re-calibrate, as needed, in accordance with information from AI system monitoring and assessment.

Transparency & Documentation
Organizations can document the following:
- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?
- What assessments has the entity conducted on data security and privacy impacts associated with the AI system?
- Does your organization have an existing governance structure that can be leveraged to oversee the organization’s use of AI?
AI Transparency Resources


References


Manage 1.3
Responses to the AI risks deemed high priority as identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.

About
Outcomes from GOVERN-1, MAP-5 and MEASURE-2, can be used to address and document identified risks based on established risk tolerances. Organizations can follow existing regulations and guidelines for risk criteria, tolerances and responses established by organizational, domain, discipline, sector, or professional requirements. In lieu of such guidance, organizations can develop risk response plans based on strategies such as accepted model risk management, enterprise risk management, and information sharing and disclosure practices.

Suggested Actions
- Observe regulatory and established organizational, sector, discipline, or professional standards and requirements for applying risk tolerances within the organization.
- Document procedures for acting on AI system risks related to trustworthiness characteristics.
- Prioritize risks involving physical safety, legal liabilities, regulatory compliance, and negative impacts on individuals, groups, or society.
- Identify risk response plans and resources and organizational teams for carrying out response functions.
- Store risk management and system documentation in an organized, secure repository that is accessible by relevant AI Actors and appropriate personnel.

Transparency & Documentation
Organizations can document the following:
- Has the system been reviewed to ensure the AI system complies with relevant laws, regulations, standards, and guidance?
- To what extent has the entity defined and documented the regulatory environment—including minimum requirements in laws and regulations?
- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?
AI Transparency Resources

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Datasheets for Datasets. URL

References


Manage 1.4
Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented.

About
Organizations may choose to accept or transfer some of the documented risks from MAP and MANAGE 1.3 and 2.1. Such risks, known as residual risk, may affect downstream AI actors such as those engaged in system procurement or use. Transparent monitoring and managing residual risks enable cost benefit analysis and the examination of potential values of AI systems versus its potential negative impacts.

Suggested Actions
- Document residual risks within risk response plans, denoting risks that have been accepted, transferred, or subject to minimal mitigation.
- Establish procedures for disclosing residual risks to relevant downstream AI actors.
- Inform relevant downstream AI actors of requirements for safe operation, known limitations, and suggested warning labels as identified in MAP 3.4.

Transparency & Documentation
Organizations can document the following:
- What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- How will updates/revisions be documented and communicated? How often and by whom?
- How easily accessible and current is the information available to external stakeholders?

AI Transparency Resources
- Artificial Intelligence Ethics Framework for the Intelligence Community.
- Datasheets for Datasets.
References


Manage 2
Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, and documented, and informed by input from relevant AI actors.

Manage 2.1
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.

About
Organizational risk response may entail identifying and analyzing alternative approaches, methods, processes or systems, and balancing tradeoffs between trustworthiness characteristics and how they relate to organizational principles and societal values. Analysis of these tradeoffs is informed by consulting with interdisciplinary organizational teams, independent domain experts, and engaging with individuals or community groups. These processes require sufficient resource allocation.

Suggested Actions
- Plan and implement risk management practices in accordance with established organizational risk tolerances.
- Verify risk management teams are resourced to carry out functions, including:
  - Establishing processes for considering methods that are not automated; semi-automated; or other procedural alternatives for AI functions.
  - Enhance AI system transparency mechanisms for AI teams.
  - Enable exploration of AI system limitations by AI teams.
  - Identify, assess, and catalog past failed designs and negative impacts or outcomes to avoid known failure modes.
- Identify resource allocation approaches for managing risks in systems:
  - deemed high-risk,
  - that self-update (adaptive, online, reinforcement self-supervised learning or similar),
• with high uncertainty or where risk management is insufficient.
• Regularly seek and integrate external expertise and perspectives to supplement organizational diversity (e.g., demographic, disciplinary), equity, inclusion, and accessibility where internal capacity is lacking.
• Enable and encourage regular, open communication and feedback among AI actors and internal or external stakeholders related to system design or deployment decisions.
• Prepare and document plans for continuous monitoring and feedback mechanisms.

**Transparency & Documentation**

Organizations can document the following:
• Are mechanisms in place to evaluate whether internal teams are empowered and resourced to effectively carry out risk management functions?
• How will user and other forms of stakeholder engagement be integrated into risk management processes?

**AI Transparency Resources**
• [Artificial Intelligence Ethics Framework for the Intelligence Community](#).
• [Datasheets for Datasets](#).
• [GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities](#).

**References**


Manage 2.2
Mechanisms are in place and applied to sustain the value of deployed AI systems.

About
System performance and trustworthiness may evolve and shift over time, once an AI system is deployed and put into operation. This phenomenon, generally known as drift, can degrade the value of the AI system to the organization and increase the likelihood of negative impacts. Regular monitoring of AI systems’ performance and trustworthiness enhances organizations’ ability to detect and respond to drift, and thus sustain an AI system’s value once deployed. Processes and mechanisms for regular monitoring address system functionality and behavior - as well as impacts and alignment with the values and norms within the specific context of use. For example, considerations regarding impacts on personal or public safety or privacy may include limiting high speeds when operating autonomous vehicles or restricting illicit content recommendations for minors.

Regular monitoring activities can enable organizations to systematically and proactively identify emergent risks and respond according to established protocols and metrics. Options for organizational responses include 1) avoiding the risk, 2) accepting the risk, 3) mitigating the risk, or 4) transferring the risk. Each of these actions require planning and resources. Organizations are encouraged to establish risk management protocols with consideration of the trustworthiness characteristics, the deployment context, and real-world impacts.

Suggested Actions
- Establish risk controls considering trustworthiness characteristics, including:
  - Data management, quality, and privacy (e.g., minimization, rectification or deletion requests) controls as part of organizational data governance policies.
  - Machine learning and end-point security countermeasures (e.g., robust models, differential privacy, authentication, throttling).
  - Business rules that augment, limit or restrict AI system outputs within certain contexts.
Establish risk controls considering trustworthiness characteristics, including:

- Data management, quality, and privacy (e.g., minimization, rectification or deletion requests) controls as part of organizational data governance policies.
- Machine learning and end-point security countermeasures (e.g., robust models, differential privacy, authentication, throttling).
- Business rules that augment, limit or restrict AI system outputs within certain contexts.
- Utilizing domain expertise related to deployment context for continuous improvement and TEVV across the AI lifecycle.
- Development and regular tracking of human-AI teaming configurations.
- Model assessment and test, evaluation, validation and verification (TEVV) protocols.
- Use of standardized documentation and transparency mechanisms.
- Software quality assurance practices across AI lifecycle.
- Mechanisms to explore system limitations and avoid past failed designs or deployments.
- Establish mechanisms to capture feedback from system end users and potentially impacted groups.
- Review insurance policies, warranties, or contracts for legal or oversight requirements for risk transfer procedures.
- Document risk tolerance decisions and risk acceptance procedures.

**Transparency & Documentation**

Organizations can document the following:

- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- Could the AI system expose people to harm or negative impacts? What was done to mitigate or reduce the potential for harm?
- How will the accountable human(s) address changes in accuracy and precision due to either an adversary’s attempts to disrupt the AI or unrelated changes in the operational or business environment?
AI Transparency Resources

- Artificial Intelligence Ethics Framework for the Intelligence Community.

References

Safety, Validity and Reliability Risk Management Approaches and Resources

AI Incident Database. 2022. AI Incident Database.

AI AIC Repository. 2022. AI, algorithmic and automation incidents collected, dissected, examined, and divulged.


Netflix, Chaos Monkey.


Managing Risk Bias

Bias Testing and Remediation Approaches


Security and Resilience Resources
FTC Start With Security Guidelines. 2015.


Interpretability and Explainability Approaches


Privacy Resources

Data Governance

Software Resources
- PiML (explainable models, performance assessment)
- Interpret (explainable models)
- Iml (explainable models)
- Drifter library (performance assessment)
- Manifold library (performance assessment)
- SALib library (performance assessment)
- What-If Tool (performance assessment)
- MLextend (performance assessment)
- AI Fairness 360:
  - Python (bias testing and mitigation)
  - R (bias testing and mitigation)
- Adversarial-robustness-toolbox (ML security)
- Robustness (ML security)
- tensorflow/privacy (ML security)
- NIST De-identification Tools (Privacy and ML security)
- Dvc (MLops, deployment)
- Gigantum (MLops, deployment)
- Mlflow (MLops, deployment)
- Mlmd (MLops, deployment)
- Modeldb (MLops, deployment)
Manage 2.3
Procedures are followed to respond to and recover from a previously unknown risk when it is identified.

About
AI systems – like any technology – can demonstrate non-functionality or failure or unexpected and unusual behavior. They also can be subject to attacks, incidents, or other misuse or abuse – which their sources are not always known a-priori. Organizations can establish, document, communicate and maintain treatment procedures to recognize and counter, mitigate and manage risks that were not previously identified.

Suggested Actions
- Protocols, resources, and metrics are in place for continual monitoring of AI systems’ performance, trustworthiness, and alignment with contextual norms and values.
- Establish and regularly review treatment and response plans for incidents, negative impacts, or outcomes.
- Establish and maintain procedures to regularly monitor system components for drift, decontextualization, or other AI system behavior factors,
- Establish and maintain procedures for capturing feedback about negative impacts.
- Verify contingency processes to handle any negative impacts associated with mission-critical AI systems, and to deactivate systems.
- Enable preventive and post-hoc exploration of AI system limitations by relevant AI actor groups.
- Decommission systems that exceed risk tolerances.

Transparency & Documentation
Organizations can document the following:
- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- Are the responsibilities of the personnel involved in the various AI governance processes clearly defined? (Including responsibilities to decommission the AI system.)
What processes exist for data generation, acquisition/collection, ingestion, staging/storage, transformations, security, maintenance, and dissemination?

How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?

**AI Transparency Resources**

- [GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.](#)
- [Artificial Intelligence Ethics Framework for the Intelligence Community.](#)

**References**

[AI Incident Database.](#) 2022. AI Incident Database.

[AIIncIC Repository.](#) 2022. AI, algorithmic and automation incidents collected, dissected, examined, and divulged.


[National Institute for Standards and Technology (NIST).](#) 2022. Cybersecurity Framework. [URL](#)


Manage 2.4
Mechanisms are in place and applied, responsibilities are assigned and understood to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.

About
Performance inconsistent with intended use does not always increase risk or lead to negative impacts. Rigorous TEVV practices are useful for protecting against negative impacts regardless of intended use. When negative impacts do arise, superseding (bypassing), disengaging, or deactivating/decommissioning a model, AI system component(s), or the entire AI system may be necessary, such as when:
- a system reaches the end of its lifetime
- detected or identified risks exceed tolerance thresholds
- adequate system mitigation actions are beyond the organization’s capacity
- feasible system mitigation actions do not meet regulatory, legal, norms or standards
- impending risk is detected during continual monitoring, for which feasible mitigation cannot be identified or implemented in a timely fashion

Safely removing AI systems from operation, either temporarily or permanently, under these scenarios requires standard protocols that minimize operational disruption and downstream negative impacts. Protocols can involve redundant or backup systems that are developed in alignment with established system governance policies (see GOVERN 1.7), regulatory compliance, legal frameworks, business requirements and norms and l standards within the application context of use. Decision thresholds and metrics for actions to bypass or deactivate system components are part of continual monitoring procedures. Incidents that result in a bypass/deactivate decision require documentation and review to understand root causes, impacts, and potential opportunities for mitigation and redeployment. Organizations are encouraged to develop risk and change management protocols that consider and anticipate upstream and downstream consequences of both temporary and/or permanent decommissioning, and provide contingency options.
**Suggested Actions**

- Regularly review established procedures for AI system bypass actions, including plans for redundant or backup systems to ensure continuity of operational and/or business functionality.
- Regularly review Identify system incident thresholds for activating bypass or deactivation responses.
- Apply change management processes to understand the upstream and downstream consequences of bypassing or deactivating an AI system or AI system components.
- Apply protocols, resources and metrics for decisions to supersede, bypass or deactivate AI systems or AI system components.
- Preserve materials for forensic, regulatory, and legal review.
- Conduct internal root cause analysis and process reviews of bypass or deactivation events.
- Decommission and preserve system components that cannot be updated to meet criteria for redeployment.
- Establish criteria for redeploying updated system components, in consideration of trustworthy characteristics.

**Transparency & Documentation**

Organizations can document the following:

- What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g., personnel risk or changes to commercial objectives)?
- What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e., adversarial or stress testing)?
- To what extent does the entity have established procedures for retiring the AI system, if it is no longer needed?
- How did the entity use assessments and/or evaluations to determine if the system can be scaled up, continue, or be decommissioned?

**AI Transparency Resources**

References

Deacommissioning Template, Application Lifecycle And Supporting Docs, Cloud and Infrastructure Community of Practice. URL

Develop a Decommission Plan, M3 Playbook, Office of Shared Services and Solutions and Performance Improvement, General Services Administration. URL
Manage 3
AI risks and benefits from third-party entities are managed.

Manage 3.1
AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.

About
AI systems may depend on external resources and associated processes, including third-party data, software or hardware systems. Third parties’ supplying organizations with components and services, including tools, software, and expertise for AI system design, development, deployment or use can improve efficiency and scalability. It can also increase complexity and opacity, and, in-turn, risk. Documenting third-party technologies, personnel, and resources that were employed can help manage risks. Focusing first and foremost on risks involving physical safety, legal liabilities, regulatory compliance, and negative impacts on individuals, groups, or society is recommended.

Suggested Actions
- Have legal requirements been addressed?
- Apply organizational risk tolerance to third-party AI systems.
- Apply and document organizational risk management plans and practices to third-party AI technology, personnel, or other resources.
- Identify and maintain documentation for third-party AI systems and components.
- Establish testing, evaluation, validation and verification processes for third-party AI systems which address the needs for transparency without exposing proprietary algorithms.
- Establish processes to identify beneficial use and risk indicators in third-party systems or components, such as inconsistent software release schedule, sparse documentation, and incomplete software change management (e.g., lack of forward or backward compatibility).
- Organizations can establish processes for third parties to report known and potential vulnerabilities, risks or biases in supplied resources.
Suggested Actions

- Have legal requirements been addressed?
- Apply organizational risk tolerance to third-party AI systems.
- Apply and document organizational risk management plans and practices to third-party AI technology, personnel, or other resources.
- Identify and maintain documentation for third-party AI systems and components.
- Establish testing, evaluation, validation and verification processes for third-party AI systems which address the needs for transparency without exposing proprietary algorithms.
- Establish processes to identify beneficial use and risk indicators in third-party systems or components, such as inconsistent software release schedule, sparse documentation, and incomplete software change management (e.g., lack of forward or backward compatibility).
- Organizations can establish processes for third parties to report known and potential vulnerabilities, risks or biases in supplied resources.
- Verify contingency processes for handling negative impacts associated with mission-critical third-party AI systems.
- Monitor third-party AI systems for potential negative impacts and risks associated with trustworthiness characteristics.
- Decommission third-party systems that exceed risk tolerances.

Transparency & Documentation

Organizations can document the following:

- If a third party created the AI system or some of its components, how will you ensure a level of explainability or interpretability? Is there documentation?
- If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?
- Did you establish a process for third parties (e.g., suppliers, end users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?
- Have legal requirements been addressed?
AI Transparency Resources
- Artificial Intelligence Ethics Framework for the Intelligence Community.
- Datasheets for Datasets. URL

References
Manage 3.2
Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.

About
A common approach in AI development is transfer learning, whereby an existing pre-trained model is adapted for use in a different, but related application. AI actors in development tasks often use pre-trained models from third-party entities for tasks such as image classification, language prediction, and entity recognition, because the resources to build such models may not be readily available to most organizations. Pre-trained models are typically trained to address various classification or prediction problems, using exceedingly large datasets and computationally intensive resources. The use of pre-trained models can make it difficult to anticipate negative system outcomes or impacts. Lack of documentation or transparency tools increases the difficulty and general complexity when deploying pre-trained models and hinders root cause analyses.

Suggested Actions
- Identify pre-trained models within AI system inventory for risk tracking.
- Establish processes to independently and continually monitor performance and trustworthiness of pre-trained models, and as part of third-party risk tracking.
- Monitor performance and trustworthiness of AI system components connected to pre-trained models, and as part of third-party risk tracking.
- Identify, document and remediate risks arising from AI system components and pre-trained models per organizational risk management procedures, and as part of third-party risk tracking.
- Decommission AI system components and pre-trained models which exceed risk tolerances, and as part of third-party risk tracking.
Transparency & Documentation

Organizations can document the following:

- How has the entity documented the AI system’s data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata?
- Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?
- How does the entity ensure that the data collected are adequate, relevant, and not excessive in relation to the intended purpose?
- If the dataset becomes obsolete how will this be communicated?

AI Transparency Resources

- [Artificial Intelligence Ethics Framework For The Intelligence Community.](#)
- [WEF - Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations.](#)
- [Datasheets for Datasets.](#)

References

Manage 4
Risk treatments including response and recovery, and communication plans to the identified and measured AI risks are documented and monitored regularly.

Manage 4.1
AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.

About
AI system performance and trustworthiness can change due to a variety of factors. Regular AI system monitoring can help deployers identify performance degradations, adversarial attacks, unexpected and unusual behavior, near-misses, and impacts. Including pre- and post-deployment external feedback about AI system performance can enhance organizational awareness about positive and negative impacts, and reduce the time to respond to risks and harms.

Suggested Actions
- Establish and maintain procedures to monitor AI system performance for risks and negative and positive impacts associated with trustworthiness characteristics.
- Perform post-deployment TEVV tasks to evaluate AI system validity and reliability, bias and fairness, privacy, and security and resilience.
- Evaluate AI system trustworthiness in conditions similar to deployment context of use, and prior to deployment.
- Establish and implement red-teaming exercises at a prescribed cadence and evaluate their efficacy.
- Establish procedures for tracking dataset modifications such as data deletion or rectification requests.
- Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders to capture information about system performance, trustworthiness and impact.
• Share information about errors, near-misses, and attack patterns with incident databases, other organizations with similar systems, and system users and stakeholders.
• Respond to and document detected or reported negative impacts or issues in AI system performance and trustworthiness.
• Decommission systems that exceed establish risk tolerances.

Transparency & Documentation
Organizations can document the following:
• To what extent has the entity documented the post-deployment AI system’s testing methodology, metrics, and performance outcomes?
• How easily accessible and current is the information available to external stakeholders?

AI Transparency Resources
• GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities.
• Datasheets for Datasets.

References
Manage 4.2
Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors.

About
Regular monitoring processes enable system updates to enhance performance and functionality in accordance with regulatory and legal frameworks, and organizational and contextual values and norms. These processes also facilitate analyses of root causes, system degradation, drift, near-misses, and failures, and incident response and documentation.

AI actors across the lifecycle have many opportunities to capture and incorporate external feedback about system performance, limitations, and impacts, and implement continuous improvements. Improvements may not always be to model pipeline or system processes, and may instead be based on metrics beyond accuracy or other quality performance measures. In these cases, improvements may entail adaptations to business or organizational procedures or practices. Organizations are encouraged to develop improvements that will maintain traceability and transparency for developers, end users, auditors, and relevant AI actors.

Suggested Actions
- Integrate trustworthiness characteristics into protocols and metrics used for continual improvement.
- Establish processes for evaluating and integrating feedback into AI system improvements.
- Assess and evaluate alignment of proposed improvements with relevant regulatory and legal frameworks.
- Assess and evaluate alignment of proposed improvements connected to the values and norms within the context of use.
- Document the basis for decisions made relative to tradeoffs between trustworthy characteristics, system risks, and system opportunities.
Transparency & Documentation

Organizations can document the following:

- How will user and other forms of stakeholder engagement be integrated into the model development process and regular performance review once deployed?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- To what extent has the entity defined and documented the regulatory environment—including minimum requirements in laws and regulations?

AI Transparency Resources

- Artificial Intelligence Ethics Framework for the Intelligence Community.

References


Manage 4.3
Incidents and errors are communicated to relevant AI actors including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.

About
Regularly documenting an accurate and transparent account of identified and reported errors can enhance AI risk management activities. Examples include:
- how errors were identified,
- incidents related to the error,
- whether the error has been repaired, and
- how repairs can be distributed to all impacted stakeholders and users.

Suggested Actions
- Establish procedures to regularly share information about errors, incidents and negative impacts with relevant stakeholders, operators, practitioners and users, and impacted parties.
- Maintain a database of reported errors, near-misses, incidents and negative impacts including date reported, number of reports, assessment of impact and severity, and responses.
- Maintain a database of system changes, reason for change, and details of how the change was made, tested and deployed.
- Maintain version history information and metadata to enable continuous improvement processes.
- Verify that relevant AI actors responsible for identifying complex or emergent risks are properly resourced and empowered.

Transparency & Documentation
Organizations can document the following:
- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders? How easily accessible and current is the information available to external stakeholders?
What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

**AI Transparency Resources**

**References**