NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NIST
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-1

January 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

This publication is available free of charge from: https://doi.org/10.6028/NIST.AI.100-1

Update Schedule and Versions

The Artificial Intelligence Risk Management Framework (AI RMF) is intended to be a living document.

NIST will review the content and usefulness of the Framework regularly to determine if an update is appropriate; a review with formal input from the AI community is expected to take place no later than 2028. The Framework will employ a two-number versioning system to track and identify major and minor changes. The first number will represent the generation of the AI RMF and its companion documents (e.g., 1.0) and will change only with major revisions. Minor revisions will be tracked using ".n" after the generation number (e.g., 1.1). All changes will be tracked using a Version Control Table which identifies the history, including version number, date of change, and description of change. NIST plans to update the AI RMF Playbook frequently. Comments on the AI RMF Playbook may be sent via email to AIframework@nist.gov at any time and will be reviewed and integrated on a semi-annual basis.

Table of Contents

List of Tables

List of Figures

Executive Summary

Artificial intelligence (AI) technologies have significant potential to transform society and people's lives – from commerce and health to transportation and cybersecurity to the environment and our planet. AI technologies can drive inclusive economic growth and support scientific advancements that improve the conditions of our world. AI technologies, however, also pose risks that can negatively impact individuals, groups, organizations, communities, society, the environment, and the planet. Like risks for other types of technology, AI risks can emerge in a variety of ways and can be characterized as long- or short-term, high- or low-probability, systemic or localized, and high- or low-impact.

> The AI RMF refers to an *AI system* as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (Adapted from: OECD Recommendation on AI:2019; ISO/IEC 22989:2022).

While there are myriad standards and best practices to help organizations mitigate the risks of traditional software or information-based systems, the risks posed by AI systems are in many ways unique (See Appendix B). AI systems, for example, may be trained on data that can change over time, sometimes significantly and unexpectedly, affecting system functionality and trustworthiness in ways that are hard to understand. AI systems and the contexts in which they are deployed are frequently complex, making it difficult to detect and respond to failures when they occur. 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.

These risks make AI a uniquely challenging technology to deploy and utilize both for organizations and within society. Without proper controls, AI systems can amplify, perpetuate, or exacerbate inequitable or undesirable outcomes for individuals and communities. With proper controls, AI systems can mitigate and manage inequitable outcomes.

AI risk management is a key component of responsible development and use of AI systems. Responsible AI practices can help align the decisions about AI system design, development, and uses with intended aim and values. Core concepts in responsible AI emphasize human centricity, social responsibility, and sustainability. AI risk management can drive responsible uses and practices by prompting organizations and their internal teams who design, develop, and deploy AI to think more critically about context and potential or unexpected negative and positive impacts. Understanding and managing the risks of AI systems will help to enhance trustworthiness, and in turn, cultivate public trust.

Social responsibility can refer to the organization's responsibility "for the impacts of its decisions and activities on society and the environment through transparent and ethical behavior" (ISO 26000:2010). *Sustainability* refers to the "state of the global system, including environmental, social, and economic aspects, in which the needs of the present are met without compromising the ability of future generations to meet their own needs" (ISO/IEC TR 24368:2022). Responsible AI is meant to result in technology that is also equitable and accountable. The expectation is that organizational practices are carried out in accord with "*professional responsibility*," defined by ISO as an approach that "aims to ensure that professionals who design, develop, or deploy AI systems and applications or AI-based products or systems, recognize their unique position to exert influence on people, society, and the future of AI" (ISO/IEC TR 24368:2022).

As directed by the National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283), the goal of the AI RMF is to offer a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. The Framework is intended to be ***voluntary***, rights-preserving, non-sector-specific, and use-case agnostic, providing flexibility to organizations of all sizes and in all sectors and throughout society to implement the approaches in the Framework.

The Framework is designed to equip organizations and individuals – referred to here as *AI actors* – with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time. AI actors are defined by the Organisation for Economic Co-operation and Development (OECD) as "those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI" [OECD (2019) Artificial Intelligence in Society—OECD iLibrary] (See Appendix A).

The AI RMF is intended to be practical, to adapt to the AI landscape as AI technologies continue to develop, and to be operationalized by organizations in varying degrees and capacities so society can benefit from AI while also being protected from its potential harms.

The Framework and supporting resources will be updated, expanded, and improved based on evolving technology, the standards landscape around the world, and AI community experience and feedback. NIST will continue to align the AI RMF and related guidance with applicable international standards, guidelines, and practices. As the AI RMF is put into use, additional lessons will be learned to inform future updates and additional resources.

The Framework is divided into two parts. Part 1 discusses how organizations can frame the risks related to AI and describes the intended audience. Next, AI risks and trustworthiness are analyzed, outlining the characteristics of trustworthy AI systems, which include

valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with their harmful biases managed.

Part 2 comprises the "Core" of the Framework. It describes four specific functions to help organizations address the risks of AI systems in practice. These functions – **GOVERN**, **MAP**, **MEASURE**, and **MANAGE** – are broken down further into categories and subcategories. While **GOVERN** applies to all stages of organizations' AI risk management processes and procedures, the **MAP**, **MEASURE**, and **MANAGE** functions can be applied in AI system-specific contexts and at specific stages of the AI lifecycle.

Additional resources related to the Framework are included in the AI RMF Playbook, which is available via the NIST AI RMF website:
https://www.nist.gov/itl/ai-risk-management-framework.

Development of the AI RMF by NIST in collaboration with the private and public sectors is directed and consistent with its broader AI efforts called for by the National AI Initiative Act of 2020, the National Security Commission on Artificial Intelligence recommendations, and the Plan for Federal Engagement in Developing Technical Standards and Related Tools. Engagement with the AI community during this Framework's development – via responses to a formal Request for Information, three widely attended workshops, public comments on a concept paper and two drafts of the Framework, discussions at multiple public forums, and many small group meetings – has informed development of the AI RMF 1.0 as well as AI research and development and evaluation conducted by NIST and others. Priority research and additional guidance that will enhance this Framework will be captured in an associated AI Risk Management Framework Roadmap to which NIST and the broader community can contribute.

Part 1: Foundational Information

1. Framing Risk

AI risk management offers a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights, while also providing opportunities to maximize positive impacts. Addressing, documenting, and managing AI risks and potential negative impacts effectively can lead to more trustworthy AI systems.

1.1 Understanding and Addressing Risks, Impacts, and Harms

In the context of the AI RMF, *risk* refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018). When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence (Adapted from: OMB Circular A-130:2016). Negative impact or harm can be experienced by individuals, groups, communities, organizations, society, the environment, and the planet.

> "Risk management refers to coordinated activities to direct and control an organization with regard to risk" (Source: ISO 31000:2018).

While risk management processes generally address negative impacts, this Framework offers approaches to minimize anticipated negative impacts of AI systems *and* identify opportunities to maximize positive impacts. Effectively managing the risk of potential harms could lead to more trustworthy AI systems and unleash potential benefits to people (individuals, communities, and society), organizations, and systems/ecosystems. Risk management can enable AI developers and users to understand impacts and account for the inherent limitations and uncertainties in their models and systems, which in turn can improve overall system performance and trustworthiness and the likelihood that AI technologies will be used in ways that are beneficial.

The AI RMF is designed to address new risks as they emerge. This flexibility is particularly important where impacts are not easily foreseeable and applications are evolving. While some AI risks and benefits are well-known, it can be challenging to assess negative impacts and the degree of harms. Figure 1 provides examples of potential harms that can be related to AI systems.

AI risk management efforts should consider that humans may assume that AI systems work – and work well – in *all* settings. For example, whether correct or not, AI systems are often perceived as being more objective than humans or as offering greater capabilities than general software.

Fig. 1. Examples of potential harms related to AI systems. Trustworthy AI systems and their responsible use can mitigate negative risks and contribute to benefits for people, organizations, and ecosystems.

1.2 Challenges for AI Risk Management

Several challenges are described below. They should be taken into account when managing risks in pursuit of AI trustworthiness.

1.2.1 Risk Measurement

AI risks or failures that are not well-defined or adequately understood are difficult to measure quantitatively or qualitatively. The inability to appropriately measure AI risks does not imply that an AI system necessarily poses either a high or low risk. Some risk measurement challenges include:

Risks related to third-party software, hardware, and data: Third-party data or systems can accelerate research and development and facilitate technology transition. They also may complicate risk measurement. Risk can emerge both from third-party data, software or hardware itself and how it is used. Risk metrics or methodologies used by the organization developing the AI system may not align with the risk metrics or methodologies uses by the organization *deploying or operating* the system. Also, the organization developing the AI system may not be transparent about the risk metrics or methodologies it used. 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. Regardless, all parties and AI actors should manage risk in the AI systems they develop, deploy, or use as standalone or integrated components.

Tracking emergent risks: Organizations' risk management efforts will be enhanced by identifying and tracking emergent risks and considering techniques for measuring them.

AI system impact assessment approaches can help AI actors understand potential impacts or harms within specific contexts.

Availability of reliable metrics: The current lack of consensus on robust and verifiable measurement methods for risk and trustworthiness, and applicability to different AI use cases, is an AI risk measurement challenge. Potential pitfalls when seeking to measure negative risk or harms include the reality that development of metrics is often an institutional endeavor and may inadvertently reflect factors unrelated to the underlying impact. In addition, measurement approaches can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts.

Approaches for measuring impacts on a population work best if they recognize that contexts matter, that harms may affect varied groups or sub-groups differently, and that communities or other sub-groups who may be harmed are not always direct users of a system.

Risk at different stages of the AI lifecycle: Measuring risk at an earlier stage in the AI lifecycle may yield different results than measuring risk at a later stage; some risks may be latent at a given point in time and may increase as AI systems adapt and evolve. Furthermore, different AI actors across the AI lifecycle can have different risk perspectives. For example, an AI developer who makes AI software available, such as pre-trained models, can have a different risk perspective than an AI actor who is responsible for deploying that pre-trained model in a specific use case. Such deployers may not recognize that their particular uses could entail risks which differ from those perceived by the initial developer. All involved AI actors share responsibilities for designing, developing, and deploying a trustworthy AI system that is fit for purpose.

Risk in real-world settings: While measuring AI risks in a laboratory or a controlled environment may yield important insights pre-deployment, these measurements may differ from risks that emerge in operational, real-world settings.

Inscrutability: Inscrutable AI systems can complicate risk measurement. Inscrutability can be a result of the opaque nature of AI systems (limited explainability or interpretability), lack of transparency or documentation in AI system development or deployment, or inherent uncertainties in AI systems.

Human baseline: Risk management of AI systems that are intended to augment or replace human activity, for example decision making, requires some form of baseline metrics for comparison. This is difficult to systematize since AI systems carry out different tasks – and perform tasks differently – than humans.

1.2.2 Risk Tolerance

While the AI RMF can be used to prioritize risk, it does not prescribe risk tolerance. *Risk tolerance* refers to the organization's or AI actor's (see Appendix A) readiness to bear the risk in order to achieve its objectives. Risk tolerance can be influenced by legal or regulatory requirements (Adapted from: ISO GUIDE 73). Risk tolerance and the level of risk that is acceptable to organizations or society are highly contextual and application and use-case specific. Risk tolerances can be influenced by policies and norms established by AI system owners, organizations, industries, communities, or policy makers. Risk tolerances are likely to change over time as AI systems, policies, and norms evolve. Different organizations may have varied risk tolerances due to their particular organizational priorities and resource considerations.

Emerging knowledge and methods to better inform harm/cost-benefit tradeoffs will continue to be developed and debated by businesses, governments, academia, and civil society. To the extent that challenges for specifying AI risk tolerances remain unresolved, there may be contexts where a risk management framework is not yet readily applicable for mitigating negative AI risks.

The Framework is intended to be flexible and to augment existing risk practices which should align with applicable laws, regulations, and norms. Organizations should 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 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 should define reasonable risk tolerance. Once tolerance is defined, this AI RMF can be used to manage risks and to document risk management processes.

1.2.3 Risk Prioritization

Attempting to eliminate negative risk entirely can be counterproductive in practice because not all incidents and failures can be eliminated. Unrealistic expectations about risk may lead organizations to allocate resources in a manner that makes risk triage inefficient or impractical or wastes scarce resources. A risk management culture can help organizations recognize that not all AI risks are the same, and resources can be allocated purposefully. Actionable risk management efforts lay out clear guidelines for assessing trustworthiness of each AI system an organization develops or deploys. Policies and resources should be prioritized based on the assessed risk level and potential impact of an AI system. The extent to which an AI system may be customized or tailored to the specific context of use by the AI deployer can be a contributing factor.

When applying the AI RMF, risks which the organization determines to be highest for the AI systems within a given context of use call for the most urgent prioritization and most thorough risk management process. In cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed. If an AI system's development, deployment, and use cases are found to be low-risk in a specific context, that may suggest potentially lower prioritization.

Risk prioritization may differ between AI systems that are designed or deployed to directly interact with humans as compared to AI systems that are not. Higher initial prioritization may be called for in settings where the AI system is trained on large datasets comprised of sensitive or protected data such as personally identifiable information, or where the outputs of the AI systems have direct or indirect impact on humans. AI systems designed to interact only with computational systems and trained on non-sensitive datasets (for example, data collected from the physical environment) may call for lower initial prioritization. Nonetheless, regularly assessing and prioritizing risk based on context remains important because non-human-facing AI systems can have downstream safety or social implications.

Residual risk – defined as risk remaining after risk treatment (Source: ISO GUIDE 73) – directly impacts end users or affected individuals and communities. Documenting residual risks will call for the system provider to fully consider the risks of deploying the AI product and will inform end users about potential negative impacts of interacting with the system.

1.2.4 Organizational Integration and Management of Risk

AI risks should not be considered in isolation. Different AI actors have different responsibilities and awareness depending on their roles in the lifecycle. For example, organizations developing an AI system often will not have information about how the system may be used. AI risk management should be integrated and incorporated into broader enterprise risk management strategies and processes. Treating AI risks along with other critical risks, such as cybersecurity and privacy, will yield a more integrated outcome and organizational efficiencies.

The AI RMF may be utilized along with related guidance and frameworks for managing AI system risks or broader enterprise risks. Some risks related to AI systems are common across other types of software development and deployment. Examples of overlapping risks include: privacy concerns related to the use of underlying data to train AI systems; the energy and environmental implications associated with resource-heavy computing demands; security concerns related to the confidentiality, integrity, and availability of the system and its training and output data; and general security of the underlying software and hardware for AI systems.

Organizations need to establish and maintain the appropriate accountability mechanisms, roles and responsibilities, culture, and incentive structures for risk management to be effective. Use of the AI RMF alone will not lead to these changes or provide the appropriate incentives. Effective risk management is realized through organizational commitment at senior levels and may require cultural change within an organization or industry. In addition, small to medium-sized organizations managing AI risks or implementing the AI RMF may face different challenges than large organizations, depending on their capabilities and resources.

2. Audience

Identifying and managing AI risks and potential impacts – both positive and negative – requires a broad set of perspectives and actors across the AI lifecycle. Ideally, AI actors will represent a diversity of experience, expertise, and backgrounds and comprise demographically and disciplinarily diverse teams. The AI RMF is intended to be used by AI actors across the AI lifecycle and dimensions.

The OECD has developed a framework for classifying AI lifecycle activities according to five key socio-technical dimensions, each with properties relevant for AI policy and governance, including risk management [OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers]. Figure 2 shows these dimensions, slightly modified by NIST for purposes of this framework. The NIST modification highlights the importance of test, evaluation, verification, and validation (TEVV) processes throughout an AI lifecycle and generalizes the operational context of an AI system.

AI dimensions displayed in Figure 2 are the Application Context, Data and Input, AI Model, and Task and Output. AI actors involved in these dimensions who perform or manage the design, development, deployment, evaluation, and use of AI systems and drive AI risk management efforts are the *primary* AI RMF audience.

Representative AI actors across the lifecycle dimensions are listed in Figure 3 and described in detail in Appendix A. Within the AI RMF, all AI actors work together to manage risks and achieve the goals of trustworthy and responsible AI. AI actors with TEVV-specific expertise are integrated throughout the AI lifecycle and are especially likely to benefit from the Framework. Performed regularly, TEVV tasks can provide insights relative to technical, societal, legal, and ethical standards or norms, and can assist with anticipating impacts and assessing and tracking emergent risks. As a regular process within an AI lifecycle, TEVV allows for both mid-course remediation and post-hoc risk management.

The People & Planet dimension at the center of Figure 2 represents human rights and the broader well-being of society and the planet. The AI actors in this dimension comprise a separate AI RMF audience who *informs* the primary audience. These AI actors may include trade associations, standards developing organizations, researchers, advocacy groups,

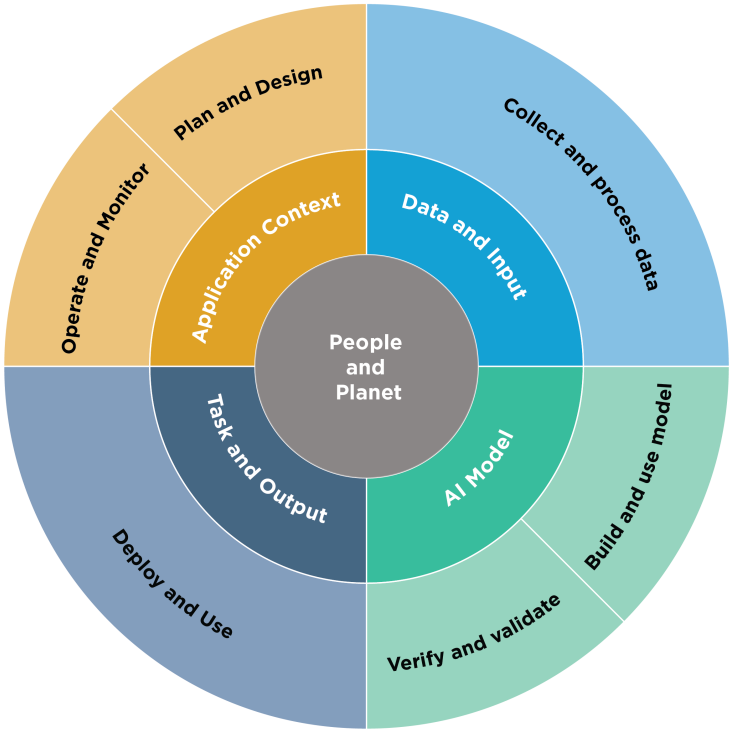

Fig. 2. Lifecycle and Key Dimensions of an AI System. Modified from OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers. The two inner circles show AI systems' key dimensions and the outer circle shows AI lifecycle stages. Ideally, risk management efforts start with the Plan and Design function in the application context and are performed throughout the AI system lifecycle. See Figure 3 for representative AI actors.

environmental groups, civil society organizations, end users, and potentially impacted individuals and communities. These actors can:

- assist in providing context and understanding potential and actual impacts;
- be a source of formal or quasi-formal norms and guidance for AI risk management;
- designate boundaries for AI operation (technical, societal, legal, and ethical); and
- promote discussion of the tradeoffs needed to balance societal values and priorities related to civil liberties and rights, equity, the environment and the planet, and the economy.

Successful risk management depends upon a sense of collective responsibility among AI actors shown in Figure 3. The AI RMF functions, described in Section 5, require diverse perspectives, disciplines, professions, and experiences. Diverse teams contribute to more open sharing of ideas and assumptions about the purposes and functions of technology – making these implicit aspects more explicit. This broader collective perspective creates opportunities for surfacing problems and identifying existing and emergent risks.

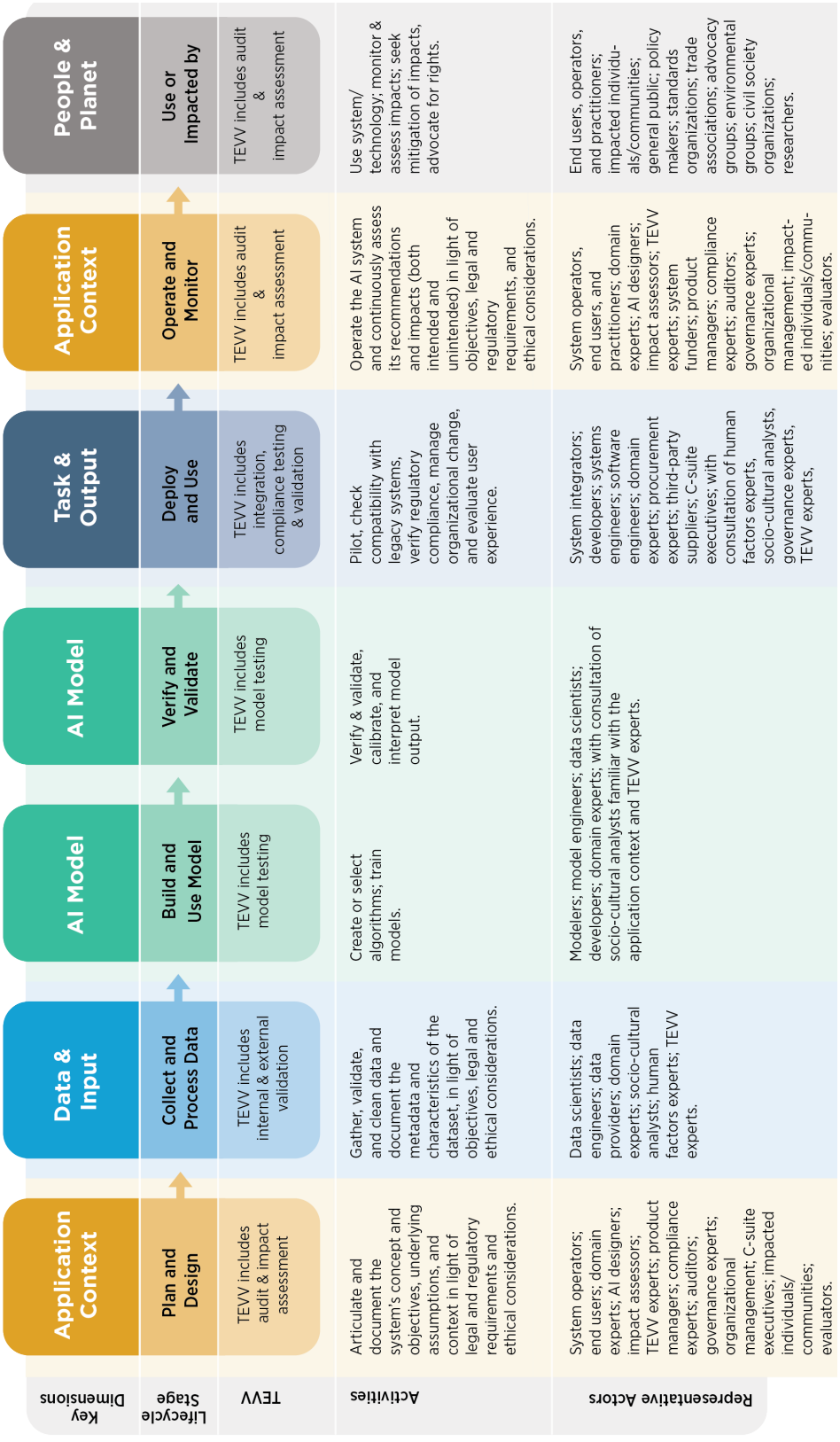

Fig. 3. AI actors across AI lifecycle stages. See Appendix A for detailed descriptions of AI actor tasks, including details about testing, evaluation, verification, and validation tasks. Note that AI actors in the AI Model dimension (Figure 2) are separated as a best practice, with those building and using the models separated from those verifying and validating the models.

3. AI Risks and Trustworthiness

For AI systems to be trustworthy, they often need to be responsive to a multiplicity of criteria that are of value to interested parties. Approaches which enhance AI trustworthiness can reduce negative AI risks. This Framework articulates the following **characteristics** of trustworthy AI and offers guidance for addressing them. Characteristics of trustworthy AI systems include: **valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.** Creating trustworthy AI requires balancing each of these characteristics based on the AI system's context of use. While all characteristics are socio-technical system attributes, accountability and transparency also relate to the processes and activities internal to an AI system and its external setting. Neglecting these characteristics can increase the probability and magnitude of negative consequences.

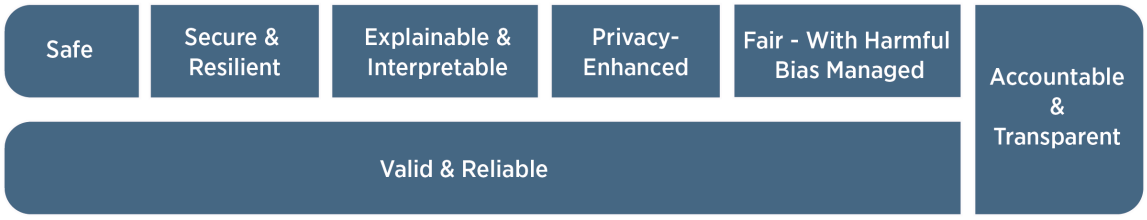

Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Trustworthiness characteristics (shown in Figure 4) are inextricably tied to social and organizational behavior, the datasets used by AI systems, selection of AI models and algorithms and the decisions made by those who build them, and the interactions with the humans who provide insight from and oversight of such systems. Human judgment should be employed when deciding on the specific metrics related to AI trustworthiness characteristics and the precise threshold values for those metrics.

Addressing AI trustworthiness characteristics individually will not ensure AI system trustworthiness; tradeoffs are usually involved, rarely do all characteristics apply in every setting, and some will be more or less important in any given situation. Ultimately, trustworthiness is a social concept that ranges across a spectrum and is only as strong as its weakest characteristics.

When managing AI risks, organizations can face difficult decisions in balancing these characteristics. For example, in certain scenarios tradeoffs may emerge between optimizing for interpretability and achieving privacy. In other cases, organizations might face a tradeoff between predictive accuracy and interpretability. Or, under certain conditions such as data sparsity, privacy-enhancing techniques can result in a loss in accuracy, affecting decisions

about fairness and other values in certain domains. Dealing with tradeoffs requires taking into account the decision-making context. These analyses can highlight the existence and extent of tradeoffs between different measures, but they do not answer questions about how to navigate the tradeoff. Those depend on the values at play in the relevant *context* and should be resolved in a manner that is both transparent and appropriately justifiable.

There are multiple approaches for enhancing contextual awareness in the AI lifecycle. For example, subject matter experts can assist in the evaluation of TEVV findings and work with product and deployment teams to align TEVV parameters to requirements and deployment conditions. When properly resourced, increasing the breadth and diversity of input from interested parties and relevant AI actors throughout the AI lifecycle can enhance opportunities for informing contextually sensitive evaluations, and for identifying AI system benefits and positive impacts. These practices can increase the likelihood that risks arising in social contexts are managed appropriately.

Understanding and treatment of trustworthiness characteristics depends on an AI actor's particular role within the AI lifecycle. For any given AI system, an AI designer or developer may have a different perception of the characteristics than the deployer.

Trustworthiness characteristics explained in this document influence each other. Highly secure but unfair systems, accurate but opaque and uninterpretable systems, and inaccurate but secure, privacy-enhanced, and transparent systems are all undesirable. A comprehensive approach to risk management calls for balancing tradeoffs among the trustworthiness characteristics. It is the joint responsibility of all AI actors to determine whether AI technology is an appropriate or necessary tool for a given context or purpose, and how to use it responsibly. The decision to commission or deploy an AI system should be based on a contextual assessment of trustworthiness characteristics and the relative risks, impacts, costs, and benefits, and informed by a broad set of interested parties.

3.1 Valid and Reliable

Validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015). Deployment of AI systems which are inaccurate, unreliable, or poorly generalized to data and settings beyond their training creates and increases negative AI risks and reduces trustworthiness.

Reliability is defined in the same standard as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022). Reliability is a goal for overall correctness of AI system operation under the conditions of expected use and over a given period of time, including the entire lifetime of the system.

Accuracy and robustness contribute to the validity and trustworthiness of AI systems, and can be in tension with one another in AI systems.

Accuracy is defined by ISO/IEC TS 5723:2022 as "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." Measures of accuracy should consider computational-centric measures (e.g., false positive and false negative rates), human-AI teaming, and demonstrate external validity (generalizable beyond the training conditions). Accuracy measurements should always be paired with clearly defined and realistic test sets – that are representative of conditions of expected use – and details about test methodology; these should be included in associated documentation. Accuracy measurements may include disaggregation of results for different data segments.

Robustness or *generalizability* is defined as the "ability of a system to maintain its level of performance under a variety of circumstances" (Source: ISO/IEC TS 5723:2022). Robustness is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated. Robustness requires not only that the system perform exactly as it does under expected uses, but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected setting.

Validity and reliability for deployed AI systems are often assessed by ongoing testing or monitoring that confirms a system is performing as intended. Measurement of validity, accuracy, robustness, and reliability contribute to trustworthiness and should take into consideration that certain types of failures can cause greater harm. AI risk management efforts should prioritize the minimization of potential negative impacts, and may need to include human intervention in cases where the AI system cannot detect or correct errors.

3.2 Safe

AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723:2022). Safe operation of AI systems is improved through:

- responsible design, development, and deployment practices;
- clear information to deployers on responsible use of the system;
- responsible decision-making by deployers and end users; and
- explanations and documentation of risks based on empirical evidence of incidents.

Different types of safety risks may require tailored AI risk management approaches based on context and the severity of potential risks presented. Safety risks that pose a potential risk of serious injury or death call for the most urgent prioritization and most thorough risk management process.

Employing safety considerations during the lifecycle and starting as early as possible with planning and design can prevent failures or conditions that can render a system dangerous. Other practical approaches for AI safety often relate to rigorous simulation and in-domain testing, real-time monitoring, and the ability to shut down, modify, or have human intervention into systems that deviate from intended or expected functionality.

AI safety risk management approaches should take cues from efforts and guidelines for safety in fields such as transportation and healthcare, and align with existing sector- or application-specific guidelines or standards.

3.3 Secure and Resilient

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 (Adapted from: ISO/IEC TS 5723:2022). 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*. Guidelines in the NIST Cybersecurity Framework and Risk Management Framework are among those which are applicable here.

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 goes beyond the provenance of the data to encompass unexpected or adversarial use (or abuse or misuse) of the model or data.

3.4 Accountable and Transparent

Trustworthy AI depends upon accountability. Accountability presupposes transparency. *Transparency* reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so. Meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system. By promoting higher levels of understanding, transparency increases confidence in the AI system.

This characteristic's scope spans from design decisions and training data to model training, the structure of the model, its intended use cases, and how and when deployment, post-deployment, or end user decisions were made and by whom. Transparency is often necessary for actionable redress related to AI system outputs that are incorrect or otherwise lead to negative impacts. Transparency should consider human-AI interaction: for exam-

ple, how a human operator or user is notified when a potential or actual adverse outcome caused by an AI system is detected. A transparent system is not necessarily an accurate, privacy-enhanced, secure, or fair system. However, it is difficult to determine whether an opaque system possesses such characteristics, and to do so over time as complex systems evolve.

The role of AI actors should be considered when seeking accountability for the outcomes of AI systems. The relationship between risk and accountability associated with AI and technological systems more broadly differs across cultural, legal, sectoral, and societal contexts. When consequences are severe, such as when life and liberty are at stake, AI developers and deployers should consider proportionally and proactively adjusting their transparency and accountability practices. Maintaining organizational practices and governing structures for harm reduction, like risk management, can help lead to more accountable systems.

Measures to enhance transparency and accountability should also consider the impact of these efforts on the implementing entity, including the level of necessary resources and the need to safeguard proprietary information.

Maintaining the provenance of training data and supporting attribution of the AI system's decisions to subsets of training data can assist with both transparency and accountability. Training data may also be subject to copyright and should follow applicable intellectual property rights laws.

As transparency tools for AI systems and related documentation continue to evolve, developers of AI systems are encouraged to test different types of transparency tools in cooperation with AI deployers to ensure that AI systems are used as intended.

3.5 Explainable and Interpretable

Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas *interpretability* refers to the meaning of AI systems' output in the context of their designed functional purposes. Together, 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. The underlying assumption is that perceptions of negative risk stem from a lack of ability to make sense of, or contextualize, system output appropriately. Explainable and interpretable AI systems offer information that will 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. (See "Four Principles of Explainable Artificial Intelligence" and "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence" found here.)

Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of "what happened" in the system. 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.

3.6 Privacy-Enhanced

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). (See The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management.)

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, affecting decisions about fairness and other values in certain domains.

3.7 Fair – with Harmful Bias Managed

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. Bias is tightly associated with the concepts of transparency as well as fairness in society. (For more information about bias, including the three categories, see NIST Special Publication 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.)

4. Effectiveness of the AI RMF

Evaluations of AI RMF effectiveness – including ways to measure bottom-line improvements in the trustworthiness of AI systems – will be part of future NIST activities, in conjunction with the AI community.

Organizations and other users of the Framework are encouraged to periodically evaluate whether the AI RMF has improved their ability to manage AI risks, including but not limited to their policies, processes, practices, implementation plans, indicators, measurements, and expected outcomes. NIST intends to work collaboratively with others to develop metrics, methodologies, and goals for evaluating the AI RMF's effectiveness, and to broadly share results and supporting information. Framework users are expected to benefit from:

- enhanced processes for governing, mapping, measuring, and managing AI risk, and clearly documenting outcomes;
- improved awareness of the relationships and tradeoffs among trustworthiness characteristics, socio-technical approaches, and AI risks;
- explicit processes for making go/no-go system commissioning and deployment decisions;
- established policies, processes, practices, and procedures for improving organizational accountability efforts related to AI system risks;
- enhanced organizational culture which prioritizes the identification and management of AI system risks and potential impacts to individuals, communities, organizations, and society;
- better information sharing within and across organizations about risks, decision-making processes, responsibilities, common pitfalls, TEVV practices, and approaches for continuous improvement;
- greater contextual knowledge for increased awareness of downstream risks;
- strengthened engagement with interested parties and relevant AI actors; and
- augmented capacity for TEVV of AI systems and associated risks.

Part 2: Core and Profiles

5. AI RMF Core

The AI RMF Core provides outcomes and actions that enable dialogue, understanding, and activities to manage AI risks and responsibly develop trustworthy AI systems. As illustrated in Figure 5, the Core is composed of four functions: **GOVERN, MAP, MEASURE,** and **MANAGE**. Each of these high-level functions is broken down into categories and subcategories. Categories and subcategories are subdivided into specific actions and outcomes. Actions do not constitute a checklist, nor are they necessarily an ordered set of steps.

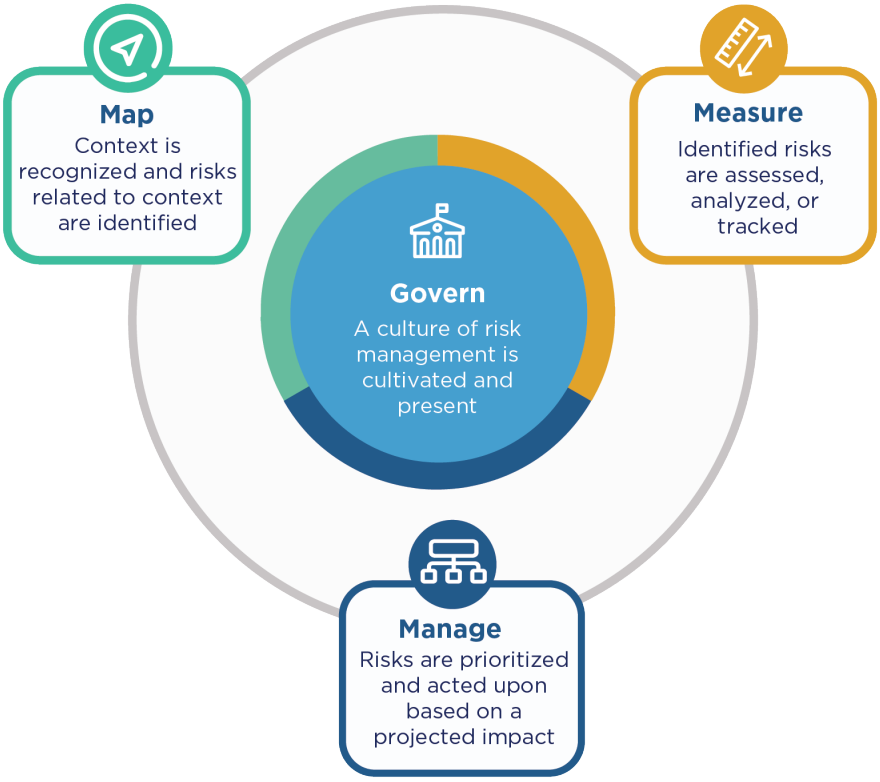

Fig. 5. Functions organize AI risk management activities at their highest level to govern, map, measure, and manage AI risks. Governance is designed to be a cross-cutting function to inform and be infused throughout the other three functions.

Risk management should be continuous, timely, and performed throughout the AI system lifecycle dimensions. AI RMF Core functions should be carried out in a way that reflects diverse and multidisciplinary perspectives, potentially including the views of AI actors outside the organization. Having a diverse team contributes to more open sharing of ideas and assumptions about purposes and functions of the technology being designed, developed,

deployed, or evaluated – which can create opportunities to surface problems and identify existing and emergent risks.

An online companion resource to the AI RMF, the NIST AI RMF Playbook, is available to help organizations navigate the AI RMF and achieve its outcomes through suggested tactical actions they can apply within their own contexts. Like the AI RMF, the Playbook is voluntary and organizations can utilize the suggestions according to their needs and interests. Playbook users can create tailored guidance selected from suggested material for their own use and contribute their suggestions for sharing with the broader community. Along with the AI RMF, the Playbook is part of the NIST Trustworthy and Responsible AI Resource Center.

Framework users may apply these functions as best suits their needs for managing AI risks based on their resources and capabilities. Some organizations may choose to select from among the categories and subcategories; others may choose and have the capacity to apply all categories and subcategories. Assuming a governance structure is in place, functions may be performed in any order across the AI lifecycle as deemed to add value by a user of the framework. After instituting the outcomes in **GOVERN**, most users of the AI RMF would start with the **MAP** function and continue to **MEASURE** or **MANAGE**. However users integrate the functions, the process should be iterative, with cross-referencing between functions as necessary. Similarly, there are categories and subcategories with elements that apply to multiple functions, or that logically should take place before certain subcategory decisions.

5.1 Govern

The **GOVERN** function:

- cultivates and implements a culture of risk management within organizations designing, developing, deploying, evaluating, or acquiring AI systems;
- outlines processes, documents, and organizational schemes that anticipate, identify, and manage the risks a system can pose, including to users and others across society – and procedures to achieve those outcomes;
- incorporates processes to assess potential impacts;
- provides a structure by which AI risk management functions can align with organizational principles, policies, and strategic priorities;
- connects technical aspects of AI system design and development to organizational values and principles, and enables organizational practices and competencies for the individuals involved in acquiring, training, deploying, and monitoring such systems; and
- addresses full product lifecycle and associated processes, including legal and other issues concerning use of third-party software or hardware systems and data.

GOVERN is a cross-cutting function that is infused throughout AI risk management and enables the other functions of the process. Aspects of **GOVERN**, especially those related to compliance or evaluation, should be integrated into each of the other functions. Attention to governance is a continual and intrinsic requirement for effective AI risk management over an AI system's lifespan and the organization's hierarchy.

Strong governance can drive and enhance internal practices and norms to facilitate organizational risk culture. Governing authorities can determine the overarching policies that direct an organization's mission, goals, values, culture, and risk tolerance. Senior leadership sets the tone for risk management within an organization, and with it, organizational culture. Management aligns the technical aspects of AI risk management to policies and operations. Documentation can enhance transparency, improve human review processes, and bolster accountability in AI system teams.

After putting in place the structures, systems, processes, and teams described in the **GOVERN** function, organizations should benefit from a purpose-driven culture focused on risk understanding and management. It is incumbent on Framework users to continue to execute the **GOVERN** function as knowledge, cultures, and needs or expectations from AI actors evolve over time.

Practices related to governing AI risks are described in the NIST AI RMF Playbook. Table 1 lists the **GOVERN** function's categories and subcategories.

Table 1: Categories and subcategories for the **GOVERN** function.

Categories	Subcategories
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. **GOVERN 1.2:** The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices. **GOVERN 1.3:** Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance. **GOVERN 1.4:** The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.

Continued on next page

Table 1: Categories and subcategories for the **GOVERN** function. (Continued)

Categories	Subcategories
	GOVERN 1.5: Ongoing monitoring and periodic review of the risk management process and its outcomes are planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review. **GOVERN 1.6:** Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities. **GOVERN 1.7:** Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.
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. **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. **GOVERN 2.3:** Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.
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-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds). **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.
GOVERN 4: Organizational teams are committed to a culture	**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 potential negative impacts.

Continued on next page

Table 1: Categories and subcategories for the **GOVERN** function. (Continued)

Categories	Subcategories
that considers and communicates AI risk.	**GOVERN 4.2:** Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.
	GOVERN 4.3: Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.
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.
	GOVERN 5.2: Mechanisms are established to enable the team that developed or deployed AI systems to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation.
GOVERN 6: Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.	**GOVERN 6.1:** 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.
	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.

5.2 Map

The **MAP** function establishes the context to frame risks related to an AI system. The AI lifecycle consists of many interdependent activities involving a diverse set of actors (See Figure 3). In practice, AI actors in charge of one part of the process often do not have full visibility or control over other parts and their associated contexts. The interdependencies between these activities, and among the relevant AI actors, can make it difficult to reliably anticipate impacts of AI systems. For example, early decisions in identifying purposes and objectives 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 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 into risk management practices. Anticipating, assessing, and otherwise addressing potential sources of negative risk can mitigate this uncertainty and enhance the integrity of the decision process.

The information gathered while carrying out the **MAP** function enables negative risk prevention and informs decisions for processes such as model management, as well as an initial decision about appropriateness or the need for an AI solution. Outcomes in the **MAP** function are the basis for the **MEASURE** and **MANAGE** functions. Without contextual knowledge, and awareness of risks within the identified contexts, risk management is difficult to perform. The **MAP** function is intended to enhance an organization's ability to identify risks and broader contributing factors.

Implementation of this function is enhanced by incorporating perspectives from a diverse internal team and engagement with those external to the team that developed or deployed the AI system. Engagement with external collaborators, end users, potentially impacted communities, and others may vary based on the risk level of a particular AI system, the makeup of the internal team, and organizational policies. Gathering such broad perspectives can help organizations proactively prevent negative risks and develop more trustworthy AI systems by:

- improving their capacity for understanding contexts;
- checking their assumptions about context of use;
- enabling recognition of when systems are not functional within or out of their intended context;
- identifying positive and beneficial uses of their existing AI systems;
- improving understanding of limitations in AI and ML processes;
- identifying constraints in real-world applications that may lead to negative impacts;
- identifying known and foreseeable negative impacts related to intended use of AI systems; and
- anticipating risks of the use of AI systems beyond intended use.

After completing the **MAP** function, Framework users should have sufficient contextual knowledge about AI system impacts to inform an initial go/no-go decision about whether to design, develop, or deploy an AI system. If a decision is made to proceed, organizations should utilize the **MEASURE** and **MANAGE** functions along with policies and procedures put into place in the **GOVERN** function to assist in AI risk management efforts. It is incumbent on Framework users to continue applying the **MAP** function to AI systems as context, capabilities, risks, benefits, and potential impacts evolve over time.

Practices related to mapping AI risks are described in the NIST AI RMF Playbook. Table 2 lists the **MAP** function's categories and subcategories.

Table 2: Categories and subcategories for the **MAP** function.

Categories	Subcategories
MAP 1: Context is established and understood.	**MAP 1.1:** Intended purposes, 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: the 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; and related TEVV and system metrics. **MAP 1.2:** Interdisciplinary 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. **MAP 1.3:** The organization's mission and relevant goals for AI technology are understood and documented. **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. **MAP 1.5:** Organizational risk tolerances are determined and documented. **MAP 1.6:** 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.
MAP 2: Categorization of the AI system is performed.	**MAP 2.1:** The specific tasks and methods used to implement the tasks that the AI system will support are defined (e.g., classifiers, generative models, recommenders). **MAP 2.2:** 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 decisions and taking subsequent actions.

Continued on next page

Table 2: Categories and subcategories for the **MAP** function. (Continued)

Categories	Subcategories
	MAP 2.3: 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.
MAP 3: AI 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. **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. **MAP 3.3:** Targeted application scope is specified and documented based on the system's capability, established context, and AI system categorization. **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. **MAP 3.5:** Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the **GOVERN** function.
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. **MAP 4.2:** Internal risk controls for components of the AI system, including third-party AI technologies, are identified and documented.
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.

Continued on next page

Table 2: Categories and subcategories for the MAP function. (Continued)

Categories	Subcategories
	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.

5.3 Measure

The MEASURE function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts. It uses knowledge relevant to AI risks identified in the MAP function and informs the MANAGE function. AI systems should be tested before their deployment and regularly while in operation. AI risk measurements include documenting aspects of systems' functionality and trustworthiness.

Measuring AI risks includes tracking metrics for trustworthy characteristics, social impact, and human-AI configurations. Processes developed or adopted in the MEASURE function should include rigorous software testing and performance assessment methodologies with associated measures of uncertainty, comparisons to performance benchmarks, and formalized reporting and documentation of results. Processes for independent review can improve the effectiveness of testing and can mitigate internal biases and potential conflicts of interest.

Where tradeoffs among the trustworthy characteristics arise, measurement provides a traceable basis to inform management decisions. Options may include recalibration, impact mitigation, or removal of the system from design, development, production, or use, as well as a range of compensating, detective, deterrent, directive, and recovery controls.

After completing the MEASURE function, objective, repeatable, or scalable test, evaluation, verification, and validation (TEVV) processes including metrics, methods, and methodologies are in place, followed, and documented. Metrics and measurement methodologies should adhere to scientific, legal, and ethical norms and be carried out in an open and transparent process. New types of measurement, qualitative and quantitative, may need to be developed. The degree to which each measurement type provides unique and meaningful information to the assessment of AI risks should be considered. Framework users will enhance their capacity to comprehensively evaluate system trustworthiness, identify and track existing and emergent risks, and verify efficacy of the metrics. Measurement outcomes will be utilized in the MANAGE function to assist risk monitoring and response efforts. It is incumbent on Framework users to continue applying the MEASURE function to AI systems as knowledge, methodologies, risks, and impacts evolve over time.

Practices related to measuring AI risks are described in the NIST AI RMF Playbook. Table 3 lists the **MEASURE** function's categories and subcategories.

Table 3: Categories and subcategories for the **MEASURE** function.

Categories	Subcategories
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.
	MEASURE 1.2: Appropriateness of AI metrics and effectiveness of existing controls are regularly assessed and updated, including reports of errors and potential impacts on affected communities.
	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.
MEASURE 2: AI systems are evaluated for trustworthy characteristics.	**MEASURE 2.1:** Test sets, metrics, and details about the tools used during TEVV are documented.
	MEASURE 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.
	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.
	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.
	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.

Continued on next page

Table 3: Categories and subcategories for the **MEASURE** function. (Continued)

Categories	Subcategories
	MEASURE 2.6: The 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 it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.
	MEASURE 2.7: AI system security and resilience – as identified in the **MAP** function – are evaluated and documented.
	MEASURE 2.8: Risks associated with transparency and accountability – as identified in the **MAP** function – are examined and documented.
	MEASURE 2.9: The AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the **MAP** function – to inform responsible use and governance.
	MEASURE 2.10: Privacy risk of the AI system – as identified in the **MAP** function – is examined and documented.
	MEASURE 2.11: Fairness and bias – as identified in the **MAP** function – are evaluated and results are documented.
	MEASURE 2.12: Environmental impact and sustainability of AI model training and management activities – as identified in the **MAP** function – are assessed and documented.
	MEASURE 2.13: Effectiveness of the employed TEVV metrics and processes in the **MEASURE** function are evaluated and documented.
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.
	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.

Continued on next page

Table 3: Categories and subcategories for the **MEASURE** function. (Continued)

Categories	Subcategories
	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.
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.
	MEASURE 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.
	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.

5.4 Manage

The **MANAGE** function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the **GOVERN** function. Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events.

Contextual information gleaned from expert consultation and input from relevant AI actors – established in **GOVERN** and carried out in **MAP** – is utilized in this function to decrease the likelihood of system failures and negative impacts. Systematic documentation practices established in **GOVERN** and utilized in **MAP** and **MEASURE** bolster AI risk management efforts and increase transparency and accountability. Processes for assessing emergent risks are in place, along with mechanisms for continual improvement.

After completing the **MANAGE** function, plans for prioritizing risk and regular monitoring and improvement will be in place. Framework users will have enhanced capacity to manage the risks of deployed AI systems and to allocate risk management resources based on assessed and prioritized risks. It is incumbent on Framework users to continue to apply the **MANAGE** function to deployed AI systems as methods, contexts, risks, and needs or expectations from relevant AI actors evolve over time.

Practices related to managing AI risks are described in the NIST AI RMF Playbook. Table 4 lists the **MANAGE** function's categories and subcategories.

Table 4: Categories and subcategories for the **MANAGE** function.

Categories	Subcategories
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 purposes and stated objectives and whether its development or deployment should proceed. **MANAGE 1.2:** Treatment of documented AI risks is prioritized based on impact, likelihood, and available resources or methods. **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. **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.
MANAGE 2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, 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. **MANAGE 2.2:** Mechanisms are in place and applied to sustain the value of deployed AI systems. **MANAGE 2.3:** Procedures are followed to respond to and recover from a previously unknown risk when it is identified. **MANAGE 2.4:** Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.
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. **MANAGE 3.2:** Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.

Continued on next page

Table 4: Categories and subcategories for the **MANAGE** function. (Continued)

Categories	Subcategories
MANAGE 4: Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.	**MANAGE 4.1:** Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management. **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. **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.

6. AI RMF Profiles

AI RMF *use-case profiles* are implementations of the AI RMF functions, categories, and subcategories for a specific setting or application based on the requirements, risk tolerance, and resources of the Framework user: for example, an AI RMF *hiring profile* or an AI RMF *fair housing profile*. Profiles may illustrate and offer insights into how risk can be managed at various stages of the AI lifecycle or in specific sector, technology, or end-use applications. AI RMF profiles assist organizations in deciding how they might best manage AI risk that is well-aligned with their goals, considers legal/regulatory requirements and best practices, and reflects risk management priorities.

AI RMF *temporal profiles* are descriptions of either the current state or the desired, target state of specific AI risk management activities within a given sector, industry, organization, or application context. An AI RMF Current Profile indicates how AI is currently being managed and the related risks in terms of current outcomes. A Target Profile indicates the outcomes needed to achieve the desired or target AI risk management goals.

Comparing Current and Target Profiles likely reveals gaps to be addressed to meet AI risk management objectives. Action plans can be developed to address these gaps to fulfill outcomes in a given category or subcategory. Prioritization of gap mitigation is driven by the user's needs and risk management processes. This risk-based approach also enables Framework users to compare their approaches with other approaches and to gauge the resources needed (e.g., staffing, funding) to achieve AI risk management goals in a cost-effective, prioritized manner.

AI RMF *cross-sectoral profiles* cover risks of models or applications that can be used across use cases or sectors. Cross-sectoral profiles can also cover how to govern, map, measure, and manage risks for activities or business processes common across sectors such as the use of large language models, cloud-based services or acquisition.

This Framework does not prescribe profile templates, allowing for flexibility in implementation.

Appendix A:
Descriptions of AI Actor Tasks from Figures 2 and 3

AI Design tasks are performed during the Application Context and Data and Input phases of the AI lifecycle in Figure 2. AI Design actors create the concept and objectives of AI systems and are responsible for the planning, design, and data collection and processing tasks of the AI system so that the AI system is lawful and fit-for-purpose. Tasks include articulating and documenting the system's concept and objectives, underlying assumptions, context, and requirements; gathering and cleaning data; and documenting the metadata and characteristics of the dataset. AI actors in this category include data scientists, domain experts, socio-cultural analysts, experts in the field of diversity, equity, inclusion, and accessibility, members of impacted communities, human factors experts (e.g., UX/UI design), governance experts, data engineers, data providers, system funders, product managers, third-party entities, evaluators, and legal and privacy governance.

AI Development tasks are performed during the AI Model phase of the lifecycle in Figure 2. AI Development actors provide the initial infrastructure of AI systems and are responsible for model building and interpretation tasks, which involve the creation, selection, calibration, training, and/or testing of models or algorithms. AI actors in this category include machine learning experts, data scientists, developers, third-party entities, legal and privacy governance experts, and experts in the socio-cultural and contextual factors associated with the deployment setting.

AI Deployment tasks are performed during the Task and Output phase of the lifecycle in Figure 2. AI Deployment actors are responsible for contextual decisions relating to how the AI system is used to assure deployment of the system into production. Related tasks include piloting the system, checking compatibility with legacy systems, ensuring regulatory compliance, managing organizational change, and evaluating user experience. AI actors in this category include system integrators, software developers, end users, operators and practitioners, evaluators, and domain experts with expertise in human factors, socio-cultural analysis, and governance.

Operation and Monitoring tasks are performed in the Application Context/Operate and Monitor phase of the lifecycle in Figure 2. These tasks are carried out by AI actors who are responsible for operating the AI system and working with others to regularly assess system output and impacts. AI actors in this category include system operators, domain experts, AI designers, users who interpret or incorporate the output of AI systems, product developers, evaluators and auditors, compliance experts, organizational management, and members of the research community.

Test, Evaluation, Verification, and Validation (TEVV) tasks are performed throughout the AI lifecycle. They are carried out by AI actors who examine the AI system or its components, or detect and remediate problems. Ideally, AI actors carrying out verification

and validation tasks are distinct from those who perform test and evaluation actions. Tasks can be incorporated into a phase as early as design, where tests are planned in accordance with the design requirement.

- TEVV tasks for design, planning, and data may center on internal and external validation of assumptions for system design, data collection, and measurements relative to the intended context of deployment or application.
- TEVV tasks for development (i.e., model building) include model validation and assessment.
- TEVV tasks for deployment include system validation and integration in production, with testing, and recalibration for systems and process integration, user experience, and compliance with existing legal, regulatory, and ethical specifications.
- TEVV tasks for operations involve ongoing monitoring for periodic updates, testing, and subject matter expert (SME) recalibration of models, the tracking of incidents or errors reported and their management, the detection of emergent properties and related impacts, and processes for redress and response.

Human Factors tasks and activities are found throughout the dimensions of the AI lifecycle. They include human-centered design practices and methodologies, promoting the active involvement of end users and other interested parties and relevant AI actors, incorporating context-specific norms and values in system design, evaluating and adapting end user experiences, and broad integration of humans and human dynamics in all phases of the AI lifecycle. Human factors professionals provide multidisciplinary skills and perspectives to understand context of use, inform interdisciplinary and demographic diversity, engage in consultative processes, design and evaluate user experience, perform human-centered evaluation and testing, and inform impact assessments.

Domain Expert tasks involve input from multidisciplinary practitioners or scholars who provide knowledge or expertise in – and about – an industry sector, economic sector, context, or application area where an AI system is being used. AI actors who are domain experts can provide essential guidance for AI system design and development, and interpret outputs in support of work performed by TEVV and AI impact assessment teams.

AI Impact Assessment tasks include assessing and evaluating requirements for AI system accountability, combating harmful bias, examining impacts of AI systems, product safety, liability, and security, among others. AI actors such as impact assessors and evaluators provide technical, human factor, socio-cultural, and legal expertise.

Procurement tasks are conducted by AI actors with financial, legal, or policy management authority for acquisition of AI models, products, or services from a third-party developer, vendor, or contractor.

Governance and Oversight tasks are assumed by AI actors with management, fiduciary, and legal authority and responsibility for the organization in which an AI system is de-

signed, developed, and/or deployed. Key AI actors responsible for AI governance include organizational management, senior leadership, and the Board of Directors. These actors are parties that are concerned with the impact and sustainability of the organization as a whole.

Additional AI Actors

Third-party entities include providers, developers, vendors, and evaluators of data, algorithms, models, and/or systems and related services for another organization or the organization's customers or clients. Third-party entities are responsible for AI design and development tasks, in whole or in part. By definition, they are external to the design, development, or deployment team of the organization that acquires its technologies or services. The technologies acquired from third-party entities may be complex or opaque, and risk tolerances may not align with the deploying or operating organization.

End users of an AI system are the individuals or groups that use the system for specific purposes. These individuals or groups interact with an AI system in a specific context. End users can range in competency from AI experts to first-time technology end users.

Affected individuals/communities encompass all individuals, groups, communities, or organizations directly or indirectly affected by AI systems or decisions based on the output of AI systems. These individuals do not necessarily interact with the deployed system or application.

Other AI actors may provide formal or quasi-formal norms or guidance for specifying and managing AI risks. They can include **trade associations, standards developing organizations, advocacy groups, researchers, environmental groups, and civil society organizations**.

The general public is most likely to directly experience positive and negative impacts of AI technologies. They may provide the motivation for actions taken by the AI actors. This group can include individuals, communities, and consumers associated with the context in which an AI system is developed or deployed.

Appendix B:
How AI Risks Differ from Traditional Software Risks

As with traditional software, risks from AI-based technology can be bigger than an enterprise, span organizations, and lead to societal impacts. AI systems also bring a set of risks that are not comprehensively addressed by current risk frameworks and approaches. Some AI system features that present risks also can be beneficial. For example, pre-trained models and transfer learning can advance research and increase accuracy and resilience when compared to other models and approaches. Identifying contextual factors in the **MAP** function will assist AI actors in determining the level of risk and potential management efforts.

Compared to traditional software, AI-specific risks that are new or increased include the following:

- The data used for building an AI system may not be a true or appropriate representation of the context or intended use of the AI system, and the ground truth may either not exist or not be available. Additionally, harmful bias and other data quality issues can affect AI system trustworthiness, which could lead to negative impacts.
- AI system dependency and reliance on data for training tasks, combined with increased volume and complexity typically associated with such data.
- Intentional or unintentional changes during training may fundamentally alter AI system performance.
- Datasets used to train AI systems may become detached from their original and intended context or may become stale or outdated relative to deployment context.
- AI system scale and complexity (many systems contain billions or even trillions of decision points) housed within more traditional software applications.
- Use of pre-trained models that can advance research and improve performance can also increase levels of statistical uncertainty and cause issues with bias management, scientific validity, and reproducibility.
- Higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models.
- Privacy risk due to enhanced data aggregation capability for AI systems.
- AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to data, model, or concept drift.
- Increased opacity and concerns about reproducibility.
- Underdeveloped software testing standards and inability to document AI-based practices to the standard expected of traditionally engineered software for all but the simplest of cases.
- Difficulty in performing regular AI-based software testing, or determining what to test, since AI systems are not subject to the same controls as traditional code development.

- Computational costs for developing AI systems and their impact on the environment and planet.
- Inability to predict or detect the side effects of AI-based systems beyond statistical measures.

Privacy and cybersecurity risk management considerations and approaches are applicable in the design, development, deployment, evaluation, and use of AI systems. Privacy and cybersecurity risks are also considered as part of broader enterprise risk management considerations, which may incorporate AI risks. As part of the effort to address AI trustworthiness characteristics such as "Secure and Resilient" and "Privacy-Enhanced," organizations may consider leveraging available standards and guidance that provide broad guidance to organizations to reduce security and privacy risks, such as, but not limited to, the NIST Cybersecurity Framework, the NIST Privacy Framework, the NIST Risk Management Framework, and the Secure Software Development Framework. These frameworks have some features in common with the AI RMF. Like most risk management approaches, they are outcome-based rather than prescriptive and are often structured around a Core set of functions, categories, and subcategories. While there are significant differences between these frameworks based on the domain addressed – and because AI risk management calls for addressing many other types of risks – frameworks like those mentioned above may inform security and privacy considerations in the **MAP**, **MEASURE**, and **MANAGE** functions of the AI RMF.

At the same time, guidance available before publication of this AI RMF does not comprehensively address many AI system risks. For example, existing frameworks and guidance are unable to:

- adequately manage the problem of harmful bias in AI systems;
- confront the challenging risks related to generative AI;
- comprehensively address security concerns related to evasion, model extraction, membership inference, availability, or other machine learning attacks;
- account for the complex attack surface of AI systems or other security abuses enabled by AI systems; and
- consider risks associated with third-party AI technologies, transfer learning, and off-label use where AI systems may be trained for decision-making outside an organization's security controls or trained in one domain and then "fine-tuned" for another.

Both AI and traditional software technologies and systems are subject to rapid innovation. Technology advances should be monitored and deployed to take advantage of those developments and work towards a future of AI that is both trustworthy and responsible.

Appendix C:
AI Risk Management and Human-AI Interaction

Organizations that design, develop, or deploy AI systems for use in operational settings may enhance their AI risk management by understanding current limitations of human-AI interaction. The AI RMF provides opportunities to clearly define and differentiate the various human roles and responsibilities when using, interacting with, or managing AI systems.

Many of the data-driven approaches that AI systems rely on attempt to convert or represent individual and social observational and decision-making practices into measurable quantities. Representing complex human phenomena with mathematical models can come at the cost of removing necessary context. This loss of context may in turn make it difficult to understand individual and societal impacts that are key to AI risk management efforts.

Issues that merit further consideration and research include:

1. **Human roles and responsibilities in decision making and overseeing AI systems need to be clearly defined and differentiated.** 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. Some AI systems may not require human oversight, such as models used to improve video compression. Other systems may specifically require human oversight.

2. **Decisions that go into the design, development, deployment, evaluation, and use of AI systems reflect systemic and human cognitive biases.** AI actors bring their cognitive biases, both individual and group, into the process. Biases can stem from end-user decision-making tasks and be introduced across the AI lifecycle via human assumptions, expectations, and decisions during design and modeling tasks. These biases, which are not necessarily always harmful, may be exacerbated by AI system opacity and the resulting lack of transparency. Systemic biases at the organizational level can influence how teams are structured and who controls the decision-making processes throughout the AI lifecycle. These biases can also influence downstream decisions by end users, decision makers, and policy makers and may lead to negative impacts.

3. **Human-AI interaction results vary.** Under certain conditions – for example, in perceptual-based judgment tasks – the AI part of the human-AI interaction can amplify human biases, leading to more biased decisions than the AI or human alone. When these variations are judiciously taken into account in organizing human-AI teams, however, they can result in complementarity and improved overall performance.

4. **Presenting AI system information to humans is complex.** Humans perceive and derive meaning from AI system output and explanations in different ways, reflecting different individual preferences, traits, and skills.

The **GOVERN** function provides organizations with the opportunity to clarify and define the roles and responsibilities for the humans in the Human-AI team configurations and those who are overseeing the AI system performance. The **GOVERN** function also creates mechanisms for organizations to make their decision-making processes more explicit, to help counter systemic biases.

The **MAP** function suggests opportunities to define and document processes for operator and practitioner proficiency with AI system performance and trustworthiness concepts, and to define relevant technical standards and certifications. Implementing **MAP** function categories and subcategories may help organizations improve their internal competency for analyzing context, identifying procedural and system limitations, exploring and examining impacts of AI-based systems in the real world, and evaluating decision-making processes throughout the AI lifecycle.

The **GOVERN** and **MAP** functions describe the importance of interdisciplinarity and demographically diverse teams and utilizing feedback from potentially impacted individuals and communities. AI actors called out in the AI RMF who perform human factors tasks and activities can assist technical teams by anchoring in design and development practices to user intentions and representatives of the broader AI community, and societal values. These actors further help to incorporate context-specific norms and values in system design and evaluate end user experiences – in conjunction with AI systems.

AI risk management approaches for human-AI configurations will be augmented by ongoing research and evaluation. For example, the degree to which humans are empowered and incentivized to challenge AI system output requires further studies. Data about the frequency and rationale with which humans overrule AI system output in deployed systems may be useful to collect and analyze.

Appendix D:
Attributes of the AI RMF

NIST described several key attributes of the AI RMF when work on the Framework first began. These attributes have remained intact and were used to guide the AI RMF's development. They are provided here as a reference.

The AI RMF strives to:

1. Be risk-based, resource-efficient, pro-innovation, and voluntary.
2. Be consensus-driven and developed and regularly updated through an open, transparent process. All stakeholders should have the opportunity to contribute to the AI RMF's development.
3. Use clear and plain language that is understandable by a broad audience, including senior executives, government officials, non-governmental organization leadership, and those who are not AI professionals – while still of sufficient technical depth to be useful to practitioners. The AI RMF should allow for communication of AI risks across an organization, between organizations, with customers, and to the public at large.
4. Provide common language and understanding to manage AI risks. The AI RMF should offer taxonomy, terminology, definitions, metrics, and characterizations for AI risk.
5. Be easily usable and fit well with other aspects of risk management. Use of the Framework should be intuitive and readily adaptable as part of an organization's broader risk management strategy and processes. It should be consistent or aligned with other approaches to managing AI risks.
6. Be useful to a wide range of perspectives, sectors, and technology domains. The AI RMF should be universally applicable to any AI technology and to context-specific use cases.
7. Be outcome-focused and non-prescriptive. The Framework should provide a catalog of outcomes and approaches rather than prescribe one-size-fits-all requirements.
8. Take advantage of and foster greater awareness of existing standards, guidelines, best practices, methodologies, and tools for managing AI risks – as well as illustrate the need for additional, improved resources.
9. Be law- and regulation-agnostic. The Framework should support organizations' abilities to operate under applicable domestic and international legal or regulatory regimes.
10. Be a living document. The AI RMF should be readily updated as technology, understanding, and approaches to AI trustworthiness and uses of AI change and as stakeholders learn from implementing AI risk management generally and this framework in particular.

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-1

NIST AIRC - Playbook

Type	Title	AI Actors	Topics	Description
Govern	Govern 1.1	Governance and Oversight	Legal and Regulatory, Governance	Legal and regulatory requirements involving AI are understood, managed, and documented.
Govern	Govern 1.2	Governance and Oversight	Trustworthy Characteristics, Governance, Validity and Reliability, Safety, Secure and Resilient, Accountability and Transparency, Explainability and Interpretability, Privacy, Fairness and Bias	The characteristics of trustworthy AI are integrated into organizational policies, processes, and procedures.
Govern	Govern 1.3	Governance and Oversight	Risk Tolerance, Governance	Processes and procedures are in place to determine the needed level of risk management activities based on the organization's risk tolerance.
Govern	Govern 1.4	Governance and Oversight	Risk Management, Governance, Documentation	The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.
Govern	Govern 1.5	Governance and Oversight, Operation and Monitoring	Continuous monitoring, Governance	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.
Govern	Govern 1.6	Governance and Oversight	Risk Management, Governance, Data, Documentation	Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.
Govern	Govern 1.7	AI Deployment, Operation and Monitoring	Decommission, Governance	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.
Govern	Govern 2.1	Governance and Oversight	Governance, Risk Culture	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.
Govern	Govern 2.2	Governance and Oversight	Governance, Training	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.
Govern	Govern 2.3	Governance and Oversight	Governance, Risk Tolerance	Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.
Govern	Govern 3.1	Governance and Oversight, AI Design	Diversity, Interdisciplinarity, Governance	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).
Govern	Govern 3.2	AI Design	Human-AI teaming, Human oversight, Governance	Policies and procedures are in place to define and differentiate roles and

Type	Title	AI Actors	Topics	Description
				responsibilities for human-AI configurations and oversight of AI systems.
Govern	Govern 4.1	AI Design, AI Development, AI Deployment, Operation and Monitoring	Risk Culture, Governance	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.
Govern	Govern 4.2	AI Design, AI Development, AI Deployment, Operation and Monitoring	Risk Culture, Governance, Impact Assessment	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.
Govern	Govern 4.3	TEVV, Operation and Monitoring, Governance and Oversight, Fairness and Bias	Risk Culture, Governance, AI Incidents, Impact Assessment, Drift, Fairness and Bias	Organizational practices are in place to enable AI testing, identification of incidents, and information sharing.
Govern	Govern 5.1	AI Design, Governance and Oversight, AI Impact Assessment, Affected Individuals and Communities	Participation, Governance, Impact Assessment	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.
Govern	Govern 5.2	AI Impact Assessment, Governance and Oversight, Operation and Monitoring	Participation, Governance, Impact Assessment	Mechanisms are established to enable AI actors to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation.
Govern	Govern 6.1	Third-party entities, Operation and Monitoring, Procurement	Third-party, Legal and Regulatory, Procurement, Supply Chain, Governance	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.
Govern	Govern 6.2	AI Deployment, TEVV, Operation and Monitoring, Third-party entities	Third-party, Governance, Risk Management, Supply Chain	Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.
Manage	Manage 1.1	AI Deployment, Operation and Monitoring, AI Impact Assessment	AI Deployment, Risk Assessment	A determination is as to whether the AI system achieves its intended purpose and stated objectives and whether its development or deployment should proceed.
Manage	Manage 1.2	AI Deployment, Operation and Monitoring, AI Impact Assessment	Risk Tolerance	Treatment of documented AI risks is prioritized based on impact, likelihood, or available resources or methods.
Manage	Manage 1.3	AI Deployment, Operation and Monitoring, AI Impact Assessment	Legal and Regulatory, Risk Tolerance	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.
Manage	Manage 1.4	AI Deployment, Operation and Monitoring, AI Impact Assessment	Risk Response	Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented.
Manage	Manage	AI Deployment,	Risk Tolerance, Trade-offs	Resources required to manage AI risks are

Type	Title	AI Actors	Topics	Description
	2.1	Operation and Monitoring, AI Impact Assessment, Governance and Oversight		taken into account, along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts.
Manage	Manage 2.2	AI Deployment, Operation and Monitoring, AI Impact Assessment, Governance and Oversight	AI Deployment, Drift, Societal Values	Mechanisms are in place and applied to sustain the value of deployed AI systems.
Manage	Manage 2.3	AI Deployment, Operation and Monitoring	Risk Response	Procedures are followed to respond to and recover from a previously unknown risk when it is identified.
Manage	Manage 2.4	AI Deployment, Operation and Monitoring, Governance and Oversight	Risk Response, Decommission	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.
Manage	Manage 3.1	Third-party entities, Operation and Monitoring, AI Deployment	Third-party, Supply Chain	AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.
Manage	Manage 3.2	Third-party entities, Operation and Monitoring, AI Deployment	Pre-trained models, Monitoring	Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.
Manage	Manage 4.1	AI Deployment, Operation and Monitoring, End-Users, Human Factors, Domain Experts, Affected Individuals and Communities	Monitoring, Participation, AI Deployment, AI Incidents, Risk Response	Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.
Manage	Manage 4.2	TEVV, AI Design, AI Development, AI Deployment, Operation and Monitoring, End-Users, Affected Individuals and Communities	Monitoring, Impact Assessment, Risk Assessment	Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors.
Manage	Manage 4.3	AI Deployment, Operation and Monitoring, End-Users, Human Factors, Domain Experts, Affected Individuals and Communities	AI Incidents, Monitoring	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.
Map	Map 1.1		Socio-technical systems, Societal Values, Context of Use, Impact Assessment, TEVV, Trustworthy Characteristics, Validity and Reliability, Safety, Secure	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

Type	Title	AI Actors	Topics	Description
			and Resilient, Accountability and Transparency, Explainability and Interpretability, Privacy, Fairness and Bias	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.
Map	Map 1.2		Diversity, Interdisciplinarity, Socio-technical systems	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.
Map	Map 1.3		Socio-technical systems, Societal Values	The organization's mission and relevant goals for the AI technology are understood and documented.
Map	Map 1.4		Context of Use	The business value or context of business use has been clearly defined or – in the case of assessing existing AI systems – re-evaluated.
Map	Map 1.5		Risk Tolerance	Organizational risk tolerances are determined and documented.
Map	Map 1.6		Socio-technical systems, Impact Assessment, Documentation	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.
Map	Map 2.1		Socio-technical systems	The specific task, and methods used to implement the task, that the AI system will support is defined (e.g., classifiers, generative models, recommenders).
Map	Map 2.2		Limitations, Human oversight, Impact Assessment, Documentation	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.
Map	Map 2.3	AI Development, TEVV, Domain Experts	TEVV, Data, Impact Assessment, Limitations	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.
Map	Map 3.1	AI Development, AI Deployment, AI Impact Assessment	Socio-technical systems, Documentation	Potential benefits of intended AI system functionality and performance are examined and documented.
Map	Map 3.2	AI Design, AI Development, Operation and Monitoring, AI Design, AI Impact Assessment	Impact Assessment, Trustworthy Characteristics, Validity and Reliability, Safety, Secure and Resilient, Accountability and Transparency, Explainability	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.

Type	Title	AI Actors	Topics	Description
			and Interpretability, Privacy, Fairness and Bias	
Map	Map 3.3	AI Design, AI Development, Human Factors	Context of Use, Documentation	Targeted application scope is specified and documented based on the system's capability, established context, and AI system categorization.
Map	Map 3.4	AI Design, AI Development, Human Factors, End-Users, Domain Experts, Operation and Monitoring	Human-AI teaming	Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed and documented.
Map	Map 3.5	Human Factors, End-Users, Domain Experts, Operation and Monitoring, AI Design	Human oversight	Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from GOVERN function.
Map	Map 4.1	Third-party entities, Procurement, Operation and Monitoring, Governance and Oversight	Legal and Regulatory, Third-party, Pre-trained models, Supply Chain, Risk Tolerance	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.
Map	Map 4.2	AI Deployment, TEVV, Operation and Monitoring, Third-party entities	Third-party, Pre-trained models	Internal risk controls for components of the AI system including third-party AI technologies are identified and documented.
Map	Map 5.1	AI Design, AI Development, AI Deployment, AI Impact Assessment, Operation and Monitoring, Affected Individuals and Communities, End-Users	Participation, Impact Assessment	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.
Map	Map 5.2	AI Design, Human Factors, AI Deployment, AI Impact Assessment, Operation and Monitoring, Domain Experts, Affected Individuals and Communities, End-Users	Participation, Impact Assessment	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.
Measure	Measure 1.1	AI Development, TEVV, Domain Experts	Trustworthy Characteristics, Risk Assessment, TEVV, Validity and Reliability, Safety, Secure and Resilient, Accountability and Transparency, Explainability and Interpretability, Privacy, Fairness and Bias	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.
Measure	Measure 1.2	TEVV, AI Impact Assessment, AI Development, AI	Impact Assessment, TEVV, Context of Use	Appropriateness of AI metrics and effectiveness of existing controls is regularly assessed and updated including

Type	Title	AI Actors	Topics	Description
		Deployment, Affected Individuals and Communities		reports of errors and impacts on affected communities.
Measure	Measure 1.3	TEVV, AI Impact Assessment, AI Development, AI Deployment, Affected Individuals and Communities, Domain Experts, End-Users, Operation and Monitoring	Participation, Impact Assessment, Context of Use	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.
Measure	Measure 2.1	TEVV	TEVV, Documentation, Validity and Reliability	Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.
Measure	Measure 2.2	TEVV, Human Factors, AI Development	Data, Human Subjects Protection	Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.
Measure	Measure 2.3	TEVV, AI Deployment	TEVV, Impact Assessment	AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.
Measure	Measure 2.4	AI Deployment, TEVV	TEVV, Monitoring, Drift	The functionality and behavior of the AI system and its components – as identified in the MAP function – are monitored when in production.
Measure	Measure 2.5	TEVV, Domain Experts	TEVV, Validity and Reliability, Trustworthy Characteristics, Data	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.
Measure	Measure 2.6	TEVV, Domain Experts, Operation and Monitoring, AI Impact Assessment, AI Deployment	TEVV, Safety, Trustworthy Characteristics, Context of Use	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.
Measure	Measure 2.7	TEVV, Domain Experts, Operation and Monitoring, AI Impact Assessment, AI Deployment	TEVV, Secure and Resilient, Trustworthy Characteristics	AI system security and resilience – as identified in the MAP function – are evaluated and documented.
Measure	Measure 2.8	TEVV, Domain Experts, Operation and Monitoring, AI Impact Assessment, AI Deployment	TEVV, Transparency and Accountability, Trustworthy Characteristics	Risks associated with transparency and accountability – as identified in the MAP function – are examined and documented.
Measure	Measure	TEVV, Domain	TEVV, Explainability and	The AI model is explained, validated, and

Type	Title	AI Actors	Topics	Description
	2.9	Experts, Operation and Monitoring, AI Impact Assessment, AI Deployment, End-Users	Interpretability, Trustworthy Characteristics	documented, and AI system output is interpreted within its context – as identified in the MAP function – and to inform responsible use and governance.
Measure	Measure 2.10	TEVV, Domain Experts, Operation and Monitoring, AI Impact Assessment, AI Deployment, End-Users	TEVV, Privacy, Trustworthy Characteristics	Privacy risk of the AI system – as identified in the MAP function – is examined and documented.
Measure	Measure 2.11	TEVV, Domain Experts, Operation and Monitoring, AI Impact Assessment, AI Deployment, End-Users, Affected Individuals and Communities	TEVV, Fairness and Bias, Trustworthy Characteristics	Fairness and bias – as identified in the MAP function – is evaluated and results are documented.
Measure	Measure 2.12	TEVV, Domain Experts, Operation and Monitoring, AI Impact Assessment, AI Deployment	TEVV, Environmental Impact	Environmental impact and sustainability of AI model training and management activities – as identified in the MAP function – are assessed and documented.
Measure	Measure 2.13	TEVV, AI Deployment, Operation and Monitoring	TEVV, Effectiveness	Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated and documented.
Measure	Measure 3.1	TEVV, AI Impact Assessment, Operation and Monitoring	TEVV, Monitoring, Continual Improvement	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.
Measure	Measure 3.2	TEVV, Domain Experts, AI Impact Assessment, Operation and Monitoring	Monitoring	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.
Measure	Measure 3.3	TEVV, AI Deployment, Operation and Monitoring, End-Users, Affected Individuals and Communities	Participation, Contestability, TEVV, Impact Assessment	Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.
Measure	Measure 4.1	TEVV, AI Deployment, Operation and Monitoring, End-Users, Affected Individuals and Communities	TEVV, Participation, Context of Use	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.
Measure	Measure 4.2	TEVV, AI Deployment, Domain Experts,	TEVV, Participation, Trustworthy Characteristics, Validity and Reliability,	Measurement results regarding AI system trustworthiness in deployment context(s) and across AI lifecycle are informed by

Type	Title	AI Actors	Topics	Description
		Operation and Monitoring, End-Users	Safety, Secure and Resilient, Accountability and Transparency, Explainability and Interpretability, Privacy, Fairness and Bias	input from domain experts and other relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.
Measure	Measure 4.3	TEVV, AI Deployment, Operation and Monitoring, End-Users, Affected Individuals and Communities	TEVV, Participation, Trustworthy Characteristics, Validity and Reliability, Safety, Secure and Resilient, Accountability and Transparency, Explainability and Interpretability, Privacy, Fairness and Bias	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.

Govern

A culture of risk management is cultivated and present.

Expand All Collapse All

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 and 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. URL

References

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

Rebecca Kelly Slaughter, "Algorithms and Economic Justice," ISP Digital Future Whitepaper & YJoLT Special Publication (2021). URL

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). URL

AI Hiring Tools and the Law, Partnership on Employment & Accessible Technology (PEAT, peatworks.org). URL

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. Polices, 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 and 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. URL

References

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021). URL

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

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

Lipton, Zachary and McAuley, Julian and Chouldechova, Alexandra, Does mitigating ML's impact disparity require treatment disparity? Advances in Neural Information Processing Systems, 2018. URL

Jessica Newman (2023) "A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Management and the AI Lifecycle," UC Berkeley Center for Long-Term Cybersecurity. URL

Emily Hadley (2022). Prioritizing Policies for Furthering Responsible Artificial Intelligence in the United States. 2022 IEEE International Conference on Big Data (Big Data), 5029-5038. URL

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

ISO, "Information technology — Reference Model of Data Management, " ISO/IEC TR 10032:200. URL

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

"National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity. URL

Kaitlin R. Boeckl and Naomi B. Lefkovitz. "NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0." National Institute of Standards and Technology (NIST), January 16, 2020. URL

"plainlanguage.gov – Home," The U.S. Government. URL

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 (risk ≈ impact x 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 ≈ impact x 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 and 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

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement. (Nov. 20, 2019). URL

Brenda Boultwood, How to Develop an Enterprise Risk-Rating Approach (Aug. 26, 2021). Global Association of Risk Professionals (garp.org). Accessed Jan. 4, 2023. URL

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 informations
  - 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 and 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
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL

References

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011).

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021). URL

Margaret Mitchell et al., "Model Cards for Model Reporting." Proceedings of 2019 FATML Conference. URL

Timnit Gebru et al., "Datasheets for Datasets," Communications of the ACM 64, No. 12, 2021. URL

Emily M. Bender, Batya Friedman, Angelina McMillan-Major (2022). A Guide for Writing Data Statements for Natural Language Processing. University of Washington. Accessed July 14, 2022. URL

M. Arnold, R. K. E. Bellamy, M. Hind, et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development 63, 4/5 (July-September 2019), 6:1-6:13. URL

Navdeep Gill, Abhishek Mathur, Marcos V. Conde (2022). A Brief Overview of AI Governance for Responsible Machine Learning Systems. ArXiv, abs/2211.13130. URL

John Richards, David Piorkowski, Michael Hind, et al. A Human-Centered Methodology for Creating AI FactSheets. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. URL

Christoph Molnar, Interpretable Machine Learning, lulu.com. URL

David A. Broniatowski. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8367. National Institute of Standards and Technology, Gaithersburg, MD. URL

OECD (2022), "OECD Framework for the Classification of AI systems", OECD Digital Economy Papers, No. 323, OECD Publishing, Paris. URL

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 and 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

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- WEF Model AI Governance Framework Assessment 2020. URL

References

National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity. URL

National Institute of Standards and Technology. (2012). Computer Security Incident Handling Guide. NIST Special Publication 800-61 Revision 2. URL

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 and 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?
- What processes exist for data generation, acquisition/collection, ingestion, staging/storage, transformations, security, maintenance, and dissemination?

AI Transparency Resources

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL

References

"A risk-based integrity level schema", in IEEE 1012, IEEE Standard for System, Software, and Hardware Verification and Validation. See Annex B. URL

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021). See "Model Inventory," pg. 26. URL

VertaAI, "ModelDB: An open-source system for Machine Learning model versioning, metadata, and experiment management." Accessed Jan. 5, 2023. URL

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 and 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

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL
- Datasheets for Datasets. URL

References

Michelle De Mooy, Joseph Jerome and Vijay Kasschau, “Should It Stay or Should It Go? The Legal, Policy and Technical Landscape Around Data Deletion,” Center for Democracy and Technology, 2017. URL

Burcu Baykurt, "Algorithmic accountability in US cities: Transparency, impact, and political economy." Big Data & Society 9, no. 2 (2022): 20539517221115426. URL

“Information System Decommissioning Guide,” Bureau of Land Management, 2011. URL

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 and 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

- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL

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

References

Andrew Smith, “Using Artificial Intelligence and Algorithms,” FTC Business Blog (Apr. 8, 2020). URL

Off. Superintendent Fin. Inst. Canada, Enterprise-Wide Model Risk Management for Deposit-Taking Institutions, E-23 (Sept. 2017).

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011).

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021). URL

ISO, “Information Technology — Artificial Intelligence — Guidelines for AI applications,” ISO/IEC CD 5339. See Section 6, “Stakeholders’ perspectives and AI application framework.” URL

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 and 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

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

References

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021). URL

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

References

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

Off. Superintendent Fin. Inst. Canada, Enterprise-Wide Model Risk Management for Deposit-Taking Institutions, E-23 (Sept. 2017).

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 and 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

- WEF Model AI Governance Framework Assessment 2020. URL
- Datasheets for Datasets. URL

References

Dylan Walsh, “How can human-centered AI fight bias in machines and people?” MIT Sloan Mgmt. Rev., 2021. URL

Michael Li, “To Build Less-Biased AI, Hire a More Diverse Team,” Harvard Bus. Rev., 2020. URL

Bo Cowgill et al., “Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics,” 2020. URL

Naomi Ellemers, Floortje Rink, “Diversity in work groups,” Current opinion in psychology, vol. 11, pp. 49–53, 2016.

Katrin Talke, Søren Salomo, Alexander Kock, “Top management team diversity and strategic innovation orientation: The relationship and consequences for innovativeness and performance,” Journal of Product Innovation Management, vol. 28, pp. 819–832, 2011.

Sarah Myers West, Meredith Whittaker, and Kate Crawford,, “Discriminating Systems: Gender, Race, and Power in AI,” AI Now Institute, Tech. Rep., 2019. URL

Sina Fazelpour, Maria De-Arteaga, Diversity in sociotechnical machine learning systems. Big Data & Society. January 2022. doi:10.1177/20539517221082027

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

“Staffing for Equitable AI: Roles & Responsibilities,” Partnership on Employment & Accessible Technology (PEAT, peatworks.org). Accessed Jan. 6, 2023. URL

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 and 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

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL

References

Madeleine Clare Elish, "Moral Crumple Zones: Cautionary tales in human-robot interaction," Engaging Science, Technology, and Society, Vol. 5, 2019. URL

“Human-AI Teaming: State-Of-The-Art and Research Needs,” National Academies of Sciences, Engineering, and Medicine, 2022. URL

Ben Green, "The Flaws Of Policies Requiring Human Oversight Of Government Algorithms," Computer Law & Security Review 45 (2022). URL

David A. Broniatowski. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8367. National Institute of Standards and Technology, Gaithersburg, MD. URL

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021). URL

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 and 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. URL
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- WEF Model AI Governance Framework Assessment 2020. URL

References

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

Patrick Hall, Navdeep Gill, and Benjamin Cox, “Responsible Machine Learning,” O’Reilly Media, 2020. URL

Off. Superintendent Fin. Inst. Canada, Enterprise-Wide Model Risk Management for Deposit-Taking Institutions, E-23 (Sept. 2017).

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

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

Kathy Baxter, AI Ethics Maturity Model, Salesforce. URL

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 and 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

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

References

Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability," AI Now Institute, 2018. URL

H.R. 2231, 116th Cong. (2019). URL

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

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

David Wright, "Making Privacy Impact Assessments More Effective." The Information Society 29, 2013. URL

Konstantinia Charitoudi and Andrew Blyth. A Socio-Technical Approach to Cyber Risk Management and Impact Assessment. Journal of Information Security 4, 1 (2013), 33-41. URL

Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, & Jacob Metcalf. 2021. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest". URL

Microsoft. Responsible AI Impact Assessment Template. 2022. URL

Microsoft. Responsible AI Impact Assessment Guide. 2022. URL

Microsoft. Foundations of assessing harm. 2022. URL

Mauritz Kop, “AI Impact Assessment & Code of Conduct,” Futurium, May 2019. URL

Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, “Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability,” AI Now, Apr. 2018. URL

Andrew D. Selbst, “An Institutional View Of Algorithmic Impact Assessments,” Harvard Journal of Law & Technology, vol. 35, no. 1, 2021

Ada Lovelace Institute. 2022. Algorithmic Impact Assessment: A Case Study in Healthcare. Accessed July 14, 2022. URL

Kathy Baxter, AI Ethics Maturity Model, Salesforce URL

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 underspecification 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 and 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. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL

References

Sean McGregor, “Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database,” arXiv:2011.08512 [cs], Nov. 2020, arXiv:2011.08512. URL

Christopher Johnson, Mark Badger, David Waltermire, Julie Snyder, and Clem Skorupka, “Guide to cyber threat information sharing,” National Institute of Standards and Technology, NIST Special Publication 800-150, Nov 2016. URL

Mengyi Wei, Zhixuan Zhou (2022). AI Ethics Issues in Real World: Evidence from AI Incident Database. ArXiv, abs/2206.07635. URL

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

“Using Combined Expertise to Evaluate Web Accessibility,” W3C Web Accessibility Initiative. URL

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 and 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. URL
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019. URL
- Stakeholders in Explainable AI, Sep. 2018. URL

References

ISO, "Ergonomics of human-system interaction — Part 210: Human-centered design for interactive systems," ISO 9241-210:2019 (2nd ed.), July 2019. URL

Rumman Chowdhury and Jutta Williams, "Introducing Twitter's first algorithmic bias bounty challenge," URL

Leonard Haas and Sebastian Gießler, "In the realm of paper tigers – exploring the failings of AI ethics guidelines," AlgorithmWatch, 2020. URL

Josh Kenway, Camille Francois, Dr. Sasha Costanza-Chock, Inioluwa Deborah Raji, & Dr. Joy Buolamwini. 2022. Bug Bounties for Algorithmic Harms? Algorithmic Justice League. Accessed July 14, 2022. URL

Microsoft Community Jury , Azure Application Architecture Guide. URL

“Definition of independent verification and validation (IV&V)”, in IEEE 1012, IEEE Standard for System, Software, and Hardware Verification and Validation. Annex C, URL

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.
  - Establishment of policies for allocating risk management resources along established risk tolerance levels, with higher-risk systems receiving more risk management resources and oversight.
  - Establishment of policies for approval, conditional approval, and disapproval of the design, implementation, and deployment of AI systems.
- Establish policies facilitating the early decommissioning of AI systems that surpass an organization's ability to reasonably mitigate risks.

Transparency and 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

- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL
- Stakeholders in Explainable AI, Sep. 2018. URL
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019. URL

References

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021). URL

The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement. (Nov. 20, 2019). Retrieved on July 12, 2022. URL

GOVERN 6.1

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 and 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

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL
- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL

- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019. URL
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019. URL

References

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

“Proposed Interagency Guidance on Third-Party Relationships: Risk Management,” 2021. URL

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021). URL

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 and 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. URL
- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019. URL

References

Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011)

“Proposed Interagency Guidance on Third-Party Relationships: Risk Management,” 2021. URL

Off. Comptroller Currency, Comptroller’s Handbook: Model Risk Management (Aug. 2021). URL

100 Bureau Drive
Gaithersburg, MD 20899
301-975-2000

Map

Context is recognized and risks related to context are identified.

Expand All Collapse All

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 and Documentation

Organizations can document the following

- To what extent is the output of each component appropriate for the operational context?
- Which AI actors are responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?
- Which AI actors are responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- Who is the person(s) accountable for the ethical considerations across the AI lifecycle?

AI Transparency Resources

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL
- “Stakeholders in Explainable AI,” Sep. 2018. URL
- "Microsoft Responsible AI Standard, v2". URL

References

Socio-technical systems

Andrew D. Selbst, danah boyd, Sorelle A. Friedler, et al. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'19). Association for Computing Machinery, New York, NY, USA, 59–68. URL

Problem formulation

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-3702. URL

Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'19). Association for Computing Machinery, New York, NY, USA, 39–48. URL

Context mapping

Emilio Gómez-González and Emilia Gómez. 2020. Artificial intelligence in medicine and healthcare. Joint Research Centre (European Commission). URL

Sarah Spiekermann and Till Winkler. 2020. Value-based Engineering for Ethics by Design. arXiv:2004.13676. URL

Social Impact Lab. 2017. Framework for Context Analysis of Technologies in Social Change Projects (Draft v2.0). URL

Solon Barocas, Asia J. Biega, Margarita Boyarskaya, et al. 2021. Responsible computing during COVID-19 and beyond. Commun. ACM 64, 7 (July 2021), 30–32. URL

Identification of harms

Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:1901.10002. URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416. URL

Microsoft. Foundations of assessing harm. 2022. URL

Understanding and documenting limitations in ML

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395. URL

Arvind Narayanan. "How to Recognize AI Snake Oil." Arthur Miller Lecture on Science and Ethics (2019). URL

Jessie J. Smith, Saleema Amershi, Solon Barocas, et al. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. arXiv:2205.08363. URL

Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. URL

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, et al. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv:1808.07261. URL

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul et al. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." Proceedings of the National Academy of Sciences 117, No. 15 (2020): 8398-8403. URL

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. URL

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. 2021. Datasheets for Datasets. arXiv:1803.09010. URL

Bender, E. M., Friedman, B. & McMillan-Major, A., (2022). A Guide for Writing Data Statements for Natural Language Processing. University of Washington. Accessed July 14, 2022. URL

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

When not to deploy

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

Statistical balance

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 6464 (25 Oct. 2019), 447-453. URL

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). URL

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 and 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. URL
- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019. URL

References

Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. Big Data & Society 9, 1 (Jan. 2022). URL

Microsoft Community Jury , Azure Application Architecture Guide. URL

Fernando Delgado, Stephen Yang, Michael Madaio, Qian Yang. (2021). Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir". URL

Kush Varshney, Tina Park, Inioluwa Deborah Raji, Gaurush Hiranandani, Narasimhan Harikrishna, Oluwasanmi Koyejo, Brianna Richardson, and Min Kyung Lee. Participatory specification of trustworthy machine learning, 2021.

Donald Martin, Vinodkumar Prabhakaran, Jill A. Kuhlberg, Andrew Smart and William S. Isaac. “Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics”, ArXiv abs/2005.07572 (2020). URL

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 and 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?

AI Transparency Resources

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

M.S. Ackerman (2000). The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. Human–Computer Interaction, 15, 179 - 203. URL

McKane Andrus, Sarah Dean, Thomas Gilbert, Nathan Lambert, Tom Zick (2021). AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks. URL

Abeba Birhane, Pratyusha Kalluri, Dallas Card, et al. 2022. The Values Encoded in Machine Learning Research. arXiv:2106.15590. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

Iason Gabriel, Artificial Intelligence, Values, and Alignment. Minds & Machines 30, 411–437 (2020). URL

PEAT “Business Case for Equitable AI”. URL

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 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 and 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

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

References

Algorithm Watch. AI Ethics Guidelines Global Inventory. URL

Ethical OS toolkit. URL

Emanuel Moss and Jacob Metcalf. 2020. Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies. Data & Society Research Institute. URL

Future of Life Institute. Asilomar AI Principles. URL

Leonard Haas, Sebastian Gießler, and Veronika Thiel. 2020. In the realm of paper tigers – exploring the failings of AI ethics guidelines. (April 28, 2020). URL

MAP 1.5

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 and 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. URL
- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL

References

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement. (Nov. 20, 2019). URL

Brenda Boultwood, How to Develop an Enterprise Risk-Rating Approach (Aug. 26, 2021). Global Association of Risk Professionals (garp.org). Accessed Jan. 4, 2023. URL

Virginia Eubanks, 1972-, Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor. New York, NY, St. Martin's Press, 2018.

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

NIST Risk Management Framework. URL

MAP 1.6

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.
- 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 and 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. URL
- Stakeholders in Explainable AI, Sep. 2018. URL
- High-Level Expert Group on Artificial Intelligence set up by the European Commission, Ethics Guidelines for Trustworthy AI. URL, PDF

References

National Academies of Sciences, Engineering, and Medicine 2022. Fostering Responsible Computing Research: Foundations and Practices. Washington, DC: The National Academies Press. URL

Abeba Birhane, William S. Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel and Shakir Mohamed. “Power to the People? Opportunities and Challenges for Participatory AI.” Equity and Access in Algorithms, Mechanisms, and Optimization (2022). URL

Amit K. Chopra, Fabiano Dalpiaz, F. Başak Aydemir, et al. 2014. Protos: Foundations for engineering innovative sociotechnical systems. In 2014 IEEE 22nd International Requirements Engineering Conference (RE) (2014), 53-62. URL

Andrew D. Selbst, danah boyd, Sorelle A. Friedler, et al. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 59–68. URL

Gordon Baxter and Ian Sommerville. 2011. Socio-technical systems: From design methods to systems engineering. Interacting with Computers, 23, 1 (Jan. 2011), 4–17. URL

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-

3702. URL

Yilin Huang, Giacomo Poderi, Sanja Šćepanović, et al. 2019. Embedding Internet-of-Things in Large-Scale Socio-technical Systems: A Community-Oriented Design in Future Smart Grids. In The Internet of Things for Smart Urban Ecosystems (2019), 125-150. Springer, Cham. URL

Victor Udoewa, (2022). An introduction to radical participatory design: decolonising participatory design processes. Design Science. 8. 10.1017/dsj.2022.24. URL

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 and 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. URL
- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL
- ATARC Model Transparency Assessment (WD) – 2020. URL
- Transparency in Artificial Intelligence - S. Larsson and F. Heintz – 2020. URL

References

Leong, Brenda (2020). The Spectrum of Artificial Intelligence - An Infographic Tool. Future of Privacy Forum. URL

Brownlee, Jason (2020). A Tour of Machine Learning Algorithms. Machine Learning Mastery. URL

MAP 2.2

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 and 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. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL
- ATARC Model Transparency Assessment (WD) – 2020. URL
- Transparency in Artificial Intelligence - S. Larsson and F. Heintz – 2020. URL

References

Context of use

International Standards Organization (ISO). 2019. ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. URL

National Institute of Standards and Technology (NIST), Mary Theofanos, Yee-Yin Choong, et al. 2017. NIST Handbook 161 Usability Handbook for Public Safety Communications: Ensuring Successful Systems for First Responders. URL

Human-AI interaction

Committee on Human-System Integration Research Topics for the 711th Human Performance Wing of the Air Force Research Laboratory and the National Academies of Sciences, Engineering, and Medicine. 2022. Human-AI Teaming: State-of-the-Art and Research Needs. Washington, D.C. National Academies Press. URL

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. URL

Saar Alon-Barkat, Madalina Busuioc, Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice, Journal of Public Administration Research and Theory, 2022;, muac007. URL

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 188 (April 2021), 21 pages. URL

Mary L. Cummings. 2006 Automation and accountability in decision support system interface design.The Journal of Technology Studies 32(1): 23–31. URL

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M. F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. NYU School of Law, Public Law Research Paper, (20-54). URL

Susanne Gaube, Harini Suresh, Martina Raue, et al. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digital Medicine 4, Article 31 (2021). URL

Ben Green. 2021. The Flaws of Policies Requiring Human Oversight of Government Algorithms. Computer Law & Security Review 45 (26 Apr. 2021). URL

Ben Green and Amba Kak. 2021. The False Comfort of Human Oversight as an Antidote to A.I. Harm. (June 15, 2021). URL

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-25. URL

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, et al. 2021. Manipulating and Measuring Model Interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 237, 1–52. URL

C. J. Smith (2019). Designing trustworthy AI: A human-machine teaming framework to guide development. arXiv preprint arXiv:1910.03515. URL

T. Warden, P. Carayon, EM et al. The National Academies Board on Human System Integration (BOHSI) Panel: Explainable AI, System Transparency, and Human Machine

Teaming. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2019;63(1):631-635. doi:10.1177/1071181319631100. URL

MAP 2.3

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 and Documentation

Organizations can document the following

- Are there any known errors, sources of noise, or redundancies in the data?
- Over what time-frame was the data collected? Does the collection time-frame 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. URL
- WEF Model AI Governance Framework Assessment 2020. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- ATARC Model Transparency Assessment (WD) – 2020. URL
- Transparency in Artificial Intelligence - S. Larsson and F. Heintz – 2020. URL

References

Challenges with dataset selection

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Front. Big Data 2, 13 (11 July 2019). URL

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, et al. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. arXiv:2012.05345. URL

Catherine D'Ignazio and Lauren F. Klein. 2020. Data Feminism. The MIT Press, Cambridge, MA. URL

Miceli, M., & Posada, J. (2022). The Data-Production Dispositif. ArXiv, abs/2205.11963.

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

National Institute of Standards and Technology (NIST), Reva Schwartz, Apostol Vassilev, et al. 2022. NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. URL

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, et al. 2021. AI and the Everything in the Whole Wide World Benchmark. arXiv:2111.15366. URL

Statistical balance

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations.

Science 366, 6464 (25 Oct. 2019), 447-453. URL

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, et al. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. arXiv:2012.05345. URL

Solon Barocas, Anhong Guo, Ece Kamar, et al. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, 368–378. URL

Measurement and evaluation

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 375–385. URL

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, et al. 2022. Evaluation Gaps in Machine Learning Practice. arXiv:2205.05256. URL

Laura Freeman, "Test and evaluation for artificial intelligence." Insight 23.1 (2020): 27-30. URL

Existing frameworks

National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity. URL

Kaitlin R. Boeckl and Naomi B. Lefkovitz. "NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0." National Institute of Standards and Technology (NIST), January 16, 2020. URL

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 and 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. URL
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019. LINK, URL

References

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-3702. URL

Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 39–48. URL

Vincent T. Covello. 2021. Stakeholder Engagement and Empowerment. In Communicating in Risk, Crisis, and High Stress Situations (Vincent T. Covello, ed.), 87-109. URL

Yilin Huang, Giacomo Poderi, Sanja Šćepanović, et al. 2019. Embedding Internet-of-Things in Large-Scale Socio-technical Systems: A Community-Oriented Design in Future Smart Grids. In The Internet of Things for Smart Urban Ecosystems (2019), 125-150. Springer, Cham. URL

Eloise Taysom and Nathan Crilly. 2017. Resilience in Sociotechnical Systems: The Perspectives of Multiple Stakeholders. She Ji: The Journal of Design, Economics, and Innovation, 3, 3 (2017), 165-182, ISSN 2405-8726. URL

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 and 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

- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019. LINK, URL

References

Abagayle Lee Blank. 2019. Computer vision machine learning and future-oriented ethics. Honors Project. Seattle Pacific University (SPU), Seattle, WA. URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416. URL

Jeff Patton. 2014. User Story Mapping. O'Reilly, Sebastopol, CA. URL

Margarita Boenig-Liptsin, Anissa Tanweer & Ari Edmundson (2022) Data Science Ethos Lifecycle: Interplay of ethical thinking and data science practice, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2022.2089411

J. Cohen, D. S. Katz, M. Barker, N. Chue Hong, R. Haines and C. Jay, "The Four Pillars of Research Software Engineering," in IEEE Software, vol. 38, no. 1, pp. 97-105, Jan.-Feb. 2021, doi: 10.1109/MS.2020.2973362.

National Academies of Sciences, Engineering, and Medicine 2022. Fostering Responsible Computing Research: Foundations and Practices. Washington, DC: The National Academies Press. URL

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

References

Mark J. Van der Laan and Sherri Rose (2018). Targeted Learning in Data Science. Cham: Springer International Publishing, 2018.

Alice Zheng. 2015. Evaluating Machine Learning Models (2015). O'Reilly. URL

Brenda Leong and Patrick Hall (2021). 5 things lawyers should know about artificial intelligence. ABA Journal. URL

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

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 and 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

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- WEF Companion to the Model AI Governance Framework- 2020. URL

References

National Academies of Sciences, Engineering, and Medicine. 2022. Human-AI Teaming: State-of-the-Art and Research Needs. Washington, DC: The National Academies Press. URL

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

Human-Machine Teaming Systems Engineering Guide. P McDermott, C Dominguez, N Kasdaglis, M Ryan, I Trahan, A Nelson. MITRE Corporation, 2018.

Saar Alon-Barkat, Madalina Busuioc, Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice, Journal of Public Administration Research and Theory, 2022;, muac007. URL

Breana M. Carter-Browne, Susannah B. F. Paletz, Susan G. Campbell , Melissa J. Carraway, Sarah H. Vahlkamp, Jana Schwartz , Polly O’Rourke, “There is No “AI” in Teams: A Multidisciplinary Framework for AIs to Work in Human Teams; Applied Research Laboratory for Intelligence and Security (ARLIS) Report, June 2021. URL

R Crootof, ME Kaminski, and WN Price II. Humans in the Loop (March 25, 2022). Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011. URL

S Mo Jones-Jang, Yong Jin Park, How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability, Journal of Computer-Mediated Communication, Volume 28, Issue 1, January 2023, zmac029. URL

A Knack, R Carter and A Babuta, "Human-Machine Teaming in Intelligence Analysis: Requirements for developing trust in machine learning systems," CETaS Research Reports (December 2022). URL

SD Ramchurn, S Stein , NR Jennings. Trustworthy human-AI partnerships. iScience. 2021;24(8):102891. Published 2021 Jul 24. doi:10.1016/j.isci.2021.102891. URL

M. Veale, M. Van Kleek, and R. Binns, “Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making,” in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18. Montreal QC, Canada: ACM Press, 2018, pp. 1–14. URL

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 and 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

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

References

Ben Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms," SSRN Journal, 2021. URL

Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn Jonker, Jeroen van den Hoven, Deborah Forster, & Reginald Lagendijk (2021). Meaningful human control: actionable properties for AI system development. AI and Ethics. URL

Mary Cummings, (2014). Automation and Accountability in Decision Support System Interface Design. The Journal of Technology Studies. 32. 10.21061/jots.v32i1.a.4. URL

Madeleine Elish, M. (2016). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (WeRobot 2016). SSRN Electronic Journal. 10.2139/ssrn.2757236. URL

R Crootof, ME Kaminski, and WN Price II. Humans in the Loop (March 25, 2022). Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011. LINK, URL

Bogdana Rakova, Jingying Yang, Henriette Cramer, & Rumman Chowdhury (2020). Where Responsible AI meets Reality. Proceedings of the ACM on Human-Computer Interaction, 5, 1 - 23. URL

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 and 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. URL
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL
- WEF Model AI Governance Framework Assessment 2020. URL

References

Language models

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. URL

Julia Kreutzer, Isaac Caswell, Lisa Wang, et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Transactions of the Association for Computational Linguistics 10 (2022), 50–72. URL

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. 2022. Taxonomy of Risks posed by Language Models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 214–229. URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. 2021. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258. URL

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus. “Emergent Abilities of Large Language Models.” ArXiv abs/2206.07682 (2022). URL

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 and 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

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities. URL
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020. URL
- WEF Model AI Governance Framework Assessment 2020. URL
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019. LINK, URL.

References

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. Retrieved on July 7, 2022. URL

Proposed Interagency Guidance on Third-Party Relationships: Risk Management, 2021. URL

Kang, D., Raghavan, D., Bailis, P.D., & Zaharia, M.A. (2020). Model Assertions for Monitoring and Improving ML Models. ArXiv, abs/2003.01668. URL

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

References

Emilio Gómez-González and Emilia Gómez. 2020. Artificial intelligence in medicine and healthcare. Joint Research Centre (European Commission). URL

Artificial Intelligence Incident Database. 2022. URL

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. “Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks". ArXiv abs/2206.08966 (2022) URL

Ganguli, D., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv. https://arxiv.org/abs/2209.07858

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

References

Susanne Vernim, Harald Bauer, Erwin Rauch, et al. 2022. A value sensitive design approach for designing AI-based worker assistance systems in manufacturing. Procedia Comput. Sci. 200, C (2022), 505–516. URL

Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:1901.10002. Retrieved from URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416. URL

Konstantinia Charitoudi and Andrew Blyth. A Socio-Technical Approach to Cyber Risk Management and Impact Assessment. Journal of Information Security 4, 1 (2013), 33-41. URL

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, & Jacob Metcalf. 2021. Assemlbing Accountability: Algorithmic Impact Assessment for the Public Interest. Data & Society. Accessed 7/14/2022 at URL

Shari Trewin (2018). AI Fairness for People with Disabilities: Point of View. ArXiv, abs/1811.10670. URL

Ada Lovelace Institute. 2022. Algorithmic Impact Assessment: A Case Study in Healthcare. Accessed July 14, 2022. URL

Microsoft Responsible AI Impact Assessment Template. 2022. Accessed July 14, 2022. URL

Microsoft Responsible AI Impact Assessment Guide. 2022. Accessed July 14, 2022. URL

Microsoft Responsible AI Standard, v2. URL

Microsoft Research AI Fairness Checklist. URL

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

Measure

Identified risks are assessed, analyzed or tracked.

Expand All Collapse All

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 and 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

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

References

Sara R. Jordan. “Designing Artificial Intelligence Review Boards: Creating Risk Metrics for Review of AI.” 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019. URL

IEEE. “IEEE-1012-2016: IEEE Standard for System, Software, and Hardware Verification and Validation.” IEEE Standards Association. URL

ACM Technology Policy Council. “Statement on Principles for Responsible Algorithmic Systems.” Association for Computing Machinery (ACM), October 26, 2022. URL

Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv. https://arxiv.org/abs/2212.09251

Ganguli, D., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv. https://arxiv.org/abs/2209.07858

David Piorkowski, Michael Hind, and John Richards. "Quantitative AI Risk Assessments: Opportunities and Challenges." arXiv preprint, submitted January 11, 2023. URL

Daniel Schiff, Aladdin Ayesh, Laura Musikanski, and John C. Havens. “IEEE 7010: A New Standard for Assessing the Well-Being Implications of Artificial Intelligence.”

2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020.
URL

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 and 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. URL

References

ACM Technology Policy Council. “Statement on Principles for Responsible Algorithmic Systems.” Association for Computing Machinery (ACM), October 26, 2022. URL

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

Harini Suresh and John Guttag. “A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle.” Equity and Access in Algorithms, Mechanisms, and Optimization, October 2021. URL

Christopher M. Bishop. Pattern Recognition and Machine Learning. New York: Springer, 2006. URL

Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. “Designing

Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, July 2021, 368–78. URL

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 assessors 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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Board of Governors of the Federal Reserve System. “SR 11-7: Guidance on Model Risk Management.” April 4, 2011. URL

“Definition of independent verification and validation (IV&V)”, in IEEE 1012, IEEE Standard for System, Software, and Hardware Verification and Validation. Annex C, URL

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. “Participation Is Not a Design Fix for Machine Learning.” Equity and Access in Algorithms, Mechanisms, and Optimization, October 2022. URL

Rediet Abebe and Kira Goldner. “Mechanism Design for Social Good.” AI Matters 4, no. 3 (October 2018): 27–34. URL

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 and 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

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL

References

Emily M. Bender and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." Transactions of the Association for Computational Linguistics 6 (2018): 587–604. URL

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. URL

IEEE Computer Society. "Software Engineering Body of Knowledge Version 3: IEEE Computer Society." IEEE Computer Society. URL

IEEE. "IEEE-1012-2016: IEEE Standard for System, Software, and Hardware Verification and Validation." IEEE Standards Association. URL

Board of Governors of the Federal Reserve System. "SR 11-7: Guidance on Model Risk Management." April 4, 2011. URL

Abigail Z. Jacobs and Hanna Wallach. "Measurement and Fairness." FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and

Transparency, March 2021, 375–85. URL

Jeanna Matthews, Bruce Hedin, Marc Canellas. Trustworthy Evidence for Trustworthy Technology: An Overview of Evidence for Assessing the Trustworthiness of Autonomous and Intelligent Systems. IEEE-USA, September 29 2022. URL

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. “Hard Choices in Artificial Intelligence.” Artificial Intelligence 300 (November 2021). URL

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 and 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

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL
- Datasheets for Datasets. URL

References

United States Department of Health, Education, and Welfare's National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Volume II. United States Department of Health and Human Services Office for Human Research Protections. April 18, 1979. URL

Office for Human Research Protections (OHRP). “45 CFR 46.” United States Department of Health and Human Services Office for Human Research Protections, March 10, 2021. URL Note: Federal Policy for Protection of Human Subjects (Common Rule). 45 CFR 46 (2018)

Office for Human Research Protections (OHRP). “Human Subject Regulations Decision Chart.” United States Department of Health and Human Services Office for Human Research Protections, June 30, 2020. URL

Jacob Metcalf and Kate Crawford. “Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide.” Big Data and Society 3, no. 1 (2016). URL

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. "Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing." arXiv preprint, submitted April 20, 2021. URL

Divyansh Kaushik, Zachary C. Lipton, and Alex John London. "Resolving the Human Subjects Status of Machine Learning's Crowdworkers." arXiv preprint, submitted June 8, 2022. URL

Office for Human Research Protections (OHRP). “International Compilation of Human Research Standards.” United States Department of Health and Human Services Office for Human Research Protections, February 7, 2022. URL

National Institutes of Health. “Definition of Human Subjects Research.” NIH Central Resource for Grants and Funding Information, January 13, 2020. URL

Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” Proceedings of the 1st Conference on Fairness, Accountability and Transparency in PMLR 81 (2018): 77–91. URL

Eun Seo Jo and Timnit Gebru. “Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, 306–16. URL

Marco Gerardi, Katarzyna Barud, Marie-Catherine Wagner, Nikolaus Forgo, Francesca Fallucchi, Noemi Scarpato, Fiorella Guadagni, and Fabio Massimo Zanzotto. "Active Informed Consent to Boost the Application of Machine Learning in Medicine." arXiv preprint, submitted September 27, 2022. URL

Shari Trewin. "AI Fairness for People with Disabilities: Point of View." arXiv preprint, submitted November 26, 2018. URL

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. URL

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 and 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 theAI system to identify errors and limitations (i.e.adversarial or stress testing)?

AI Transparency Resources

- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL
- Datasheets for Datasets. URL

References

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

Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman. "Model Selection's Disparate Impact in Real-World Deep Learning Applications." arXiv preprint, submitted September 7, 2021. URL

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. “The Fallacy of AI Functionality.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 2022, 959–72. URL

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. “Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research.” Patterns 2, no. 11 (2021): 100336. URL

Christopher M. Bishop. Pattern Recognition and Machine Learning. New York: Springer, 2006. URL

Md Johirul Islam, Giang Nguyen, Rangeet Pan, and Hridesh Rajan. "A Comprehensive Study on Deep Learning Bug Characteristics." arXiv preprint, submitted June 3, 2019. URL

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. "DQI: Measuring Data Quality in NLP." arXiv preprint, submitted May 2, 2020. URL

Doug Wielenga. "Paper 073-2007: Identifying and Overcoming Common Data Mining Mistakes." SAS Global Forum 2007: Data Mining and Predictive Modeling, SAS Institute, 2007. URL

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 and 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

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

References

Luca Piano, Fabio Garcea, Valentina Gatteschi, Fabrizio Lamberti, and Lia Morra. “Detecting Drift in Deep Learning: A Methodology Primer.” IT Professional 24, no. 5 (2022): 53–60. URL

Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub. URL

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 it's 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 and Documentation

Organizations can document the following

- What testing, if any, has the entity conducted on theAI 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

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

References

Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85. URL

Debugging Machine Learning Models. Proceedings of ICLR 2019 Workshop, May 6, 2019, New Orleans, Louisiana. URL

Patrick Hall. “Strategies for Model Debugging.” Towards Data Science, November 8, 2019. URL

Suchi Saria and Adarsh Subbaswamy. "Tutorial: Safe and Reliable Machine Learning." arXiv preprint, submitted April 15, 2019. URL

Google Developers. “Overview of Debugging ML Models.” Google Developers Machine Learning Foundational Courses, n.d. URL

R. Mohanani, I. Salman, B. Turhan, P. Rodríguez and P. Ralph, "Cognitive Biases in Software Engineering: A Systematic Mapping Study," in IEEE Transactions on Software Engineering, vol. 46, no. 12, pp. 1318-1339, Dec. 2020,

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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

AI Incident Database. 2022. URL

AIAAIC Repository. 2022. URL

Netflix. Chaos Monkey. URL

IBM. "IBM's Principles of Chaos Engineering." IBM, n.d. URL

Suchi Saria and Adarsh Subbaswamy. "Tutorial: Safe and Reliable Machine Learning." arXiv preprint, submitted April 15, 2019. URL

Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. "Model assertions for monitoring and improving ML models." Proceedings of Machine Learning and Systems 2 (2020): 481-496. URL

Larysa Visengeriyeva, et al. "Awesome MLOps." GitHub. URL

McGregor, S., Paeth, K., & Lam, K.T. (2022). Indexing AI Risks with Incidents, Issues, and Variants. ArXiv, abs/2211.10384.

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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Matthew P. Barrett. “Framework for Improving Critical Infrastructure Cybersecurity Version 1.1.” National Institute of Standards and Technology (NIST), April 16, 2018. URL

Nicolas Papernot. "A Marauder's Map of Security and Privacy in Machine Learning." arXiv preprint, submitted on November 3, 2018. URL

Gary McGraw, Harold Figueroa, Victor Shepardson, and Richie Bonett. “BIML Interactive Machine Learning Risk Framework.” Berryville Institute of Machine Learning (BIML), 2022. URL

Mitre Corporation. “Mitre/Advmlthreatmatrix: Adversarial Threat Landscape for AI Systems.” GitHub, 2023. URL

National Institute of Standards and Technology (NIST). “Cybersecurity Framework.” NIST, 2023. URL

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 blindspots 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 and 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

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

- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

National Academies of Sciences, Engineering, and Medicine. Human-AI Teaming: State-of-the-Art and Research Needs. 2022. URL

Inioluwa Deborah Raji and Jingying Yang. "ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles." arXiv preprint, submitted January 8, 2020. URL

Andrew Smith. "Using Artificial Intelligence and Algorithms." Federal Trade Commission Business Blog, April 8, 2020. URL

Board of Governors of the Federal Reserve System. “SR 11-7: Guidance on Model Risk Management.” April 4, 2011. URL

Joshua A. Kroll. “Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 1, 2021, 758–71. URL

Jennifer Cobbe, Michelle Seng Lee, and Jatinder Singh. “Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 1, 2021, 598–609. URL

MEASURE 2.9

The AI model is explained, validated, and documented, and 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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL

References

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. "This Looks Like That: Deep Learning for Interpretable Image Recognition." arXiv preprint, submitted December 28, 2019. URL

Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." arXiv preprint, submitted September 22, 2019. URL

David A. Broniatowski. "NISTIR 8367 Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST), 2021. URL

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges Toward Responsible AI.” Information Fusion 58 (June 2020): 82–115. URL

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. “Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems.” IUI '20: Proceedings of the 25th International Conference on Intelligent User Interfaces, March 17, 2020, 454–64. URL

P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, and Mark A. Przybocki. "NISTIR 8312 Four Principles of Explainable Artificial Intelligence." National Institute of Standards and Technology (NIST), September 2021. URL

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. URL

Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. “A Nutritional Label for Rankings.” SIGMOD '18: Proceedings of the 2018 International Conference on Management of Data, May 27, 2018, 1773–76. URL

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." arXiv preprint, submitted August 9, 2016. URL

Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions." NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4, 2017, 4768-4777. URL

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. “Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods.” AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 7, 2020, 180–86. URL

David Alvarez-Melis and Tommi S. Jaakkola. "Towards robust interpretability with self-explaining neural networks." NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, December 3, 2018, 7786-7795. URL

FinRegLab, Laura Biattner, and Jann Spiess. "Machine Learning Explainability & Fairness: Insights from Consumer Lending." FinRegLab, April 2022. URL

Miguel Ferreira, Muhammad Bilal Zafar, and Krishna P. Gummadi. "The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems." arXiv preprint, submitted October 31, 2016. URL

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. "Interpretable & Explorable Approximations of Black Box Models." arXiv preprint, July 4, 2017. URL

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 and 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

- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. (URL
- Datasheets for Datasets. URL

References

Kaitlin R. Boeckl and Naomi B. Lefkovitz. "NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0." National Institute of Standards and Technology (NIST), January 16, 2020. URL

Latanya Sweeney. “K-Anonymity: A Model for Protecting Privacy.” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 5 (2002): 557–70. URL

Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. “L-Diversity: Privacy beyond K-Anonymity.” 22nd International Conference on Data Engineering (ICDE'06), 2006. URL

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "CERIAS Tech Report 2007-78 t-Closeness: Privacy Beyond k-Anonymity and -Diversity." Center for Education and Research, Information Assurance and Security, Purdue University, 2001. URL

J. Domingo-Ferrer and J. Soria-Comas. "From t-closeness to differential privacy and vice versa in data anonymization." arXiv preprint, submitted December 21, 2015. URL

Joseph Near, David Darais, and Kaitlin Boeckly. "Differential Privacy for Privacy-Preserving Data Analysis: An Introduction to our Blog Series." National Institute of Standards and Technology (NIST), July 27, 2020. URL

Cynthia Dwork. “Differential Privacy.” Automata, Languages and Programming, 2006, 1–12. URL

Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. "Differential Privacy and Machine Learning: a Survey and Review." arXiv preprint, submitted December 24,2014. URL

Michael B. Hawes. "Implementing Differential Privacy: Seven Lessons From the 2020 United States Census." Harvard Data Science Review 2, no. 2 (2020). URL

Harvard University Privacy Tools Project. “Differential Privacy.” Harvard University, n.d. URL

John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Matthew Spence Sexton and Pavel Zhuravlev. "The 2020 Census Disclosure Avoidance System TopDown Algorithm." United States Census Bureau, April 7, 2022. URL

Nicolas Papernot and Abhradeep Guha Thakurta. "How to deploy machine learning with differential privacy." National Institute of Standards and Technology (NIST), December 21, 2021. URL

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

Helen Nissenbaum. "Contextual Integrity Up and Down the Data Food Chain." Theoretical Inquiries in Law 20, L. 221 (2019): 221-256. URL

Sebastian Benthall, Seda Gürses, and Helen Nissenbaum. “Contextual Integrity through the Lens of Computer Science.” Foundations and Trends in Privacy and Security 2, no. 1 (December 22, 2017): 1–69. URL

Jenifer Sunrise Winter and Elizabeth Davidson. “Big Data Governance of Personal Health Information and Challenges to Contextual Integrity.” The Information Society: An International Journal 35, no. 1 (2019): 36–51. [URL] (https://doi.org/10.1080/01972243.2018.1542648.

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 regards 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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL
- Datasheets for Datasets. URL

References

Ali Hasan, Shea Brown, Jovana Davidovic, Benjamin Lange, and Mitt Regan. “Algorithmic Bias and Risk Assessments: Lessons from Practice.” Digital Society 1 (2022). URL

Richard N. Landers and Tara S. Behrend. “Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models.” American Psychologist 78, no. 1 (2023): 36–49. URL

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “A Survey on Bias and Fairness in Machine Learning.” ACM Computing Surveys 54, no. 6 (July 2021): 1–35. URL

Michele Loi and Christoph Heitz. “Is Calibration a Fairness Requirement?” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 2022, 2026–34. URL

Shea Brown, Ryan Carrier, Merve Hickok, and Adam Leon Smith. “Bias Mitigation in Data Sets.” SocArXiv, July 8, 2021. URL

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. "NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." National Institute of Standards and Technology (NIST), 2022. URL

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

Samir Passi and Solon Barocas. “Problem Formulation and Fairness.” FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency,

January 2019, 39–48. URL

Jade S. Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P. Bennett, Jamie McCusker, and Deborah L. McGuinness. “An Ontology for Fairness Metrics.” AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, July 2022, 265–75. URL

Zhang, B., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. https://arxiv.org/pdf/1801.07593.pdf

Ganguli, D., et al. (2023). The Capacity for Moral Self-Correction in Large Language Models. arXiv. https://arxiv.org/abs/2302.07459

Arvind Narayanan. “Tl;DS - 21 Fairness Definition and Their Politics by Arvind Narayanan.” Dora's world, July 19, 2019. URL

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

Alexandra Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” Big Data 5, no. 2 (June 1, 2017): 153–63. URL

Sina Fazelpour and Zachary C. Lipton. “Algorithmic Fairness from a Non-Ideal Perspective.” AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 7, 2020, 57–63. URL

Hemank Lamba, Kit T. Rodolfa, and Rayid Ghani. “An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings.” ACM SIGKDD Explorations Newsletter 23, no. 1 (May 29, 2021): 69–85. URL

ISO. “ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making.” ISO Standards, November 2021. URL

Shari Trewin. "AI Fairness for People with Disabilities: Point of View." arXiv preprint, submitted November 26, 2018. URL

MathWorks. “Explore Fairness Metrics for Credit Scoring Model.” MATLAB & Simulink, 2023. URL

Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness.” FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85. URL

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Quantifying and Reducing Stereotypes in Word Embeddings." arXiv preprint, submitted June 20, 2016. URL

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.” Science 356, no. 6334 (April 14, 2017): 183–86. URL

Sina Fazelpour and Maria De-Arteaga. “Diversity in Sociotechnical Machine Learning Systems.” Big Data and Society 9, no. 1 (2022). URL

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

Safiya Umoja Noble. Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY: New York University Press, 2018. URL

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” Science 366, no. 6464 (October 25, 2019): 447–53. URL

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. "A Reductions Approach to Fair Classification." arXiv preprint, submitted July 16, 2018. URL

Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning." arXiv preprint, submitted October 7, 2016. URL

Alekh Agarwal, Miroslav Dudik, Zhiwei Steven Wu. "Fair Regression: Quantitative Definitions and Reduction-Based Algorithms." Proceedings of the 36th International Conference on Machine Learning, PMLR 97:120-129, 2019. URL

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. “Fairness and Abstraction in Sociotechnical Systems.” FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 29, 2019, 59–68. URL

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. “Unequal Representation and Gender Stereotypes in Image Search Results for Occupations.” CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, April 18, 2015, 3819–28. 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 and 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. URL
- Datasheets for Datasets. URL

References

Organisation for Economic Co-operation and Development (OECD). "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint." OECD Digital Economy Papers, No. 341, OECD Publishing, Paris. URL

Victor Schmidt, Alexandra Luccioni, Alexandre Lacoste, and Thomas Dandres. "Machine Learning CO2 Impact Calculator." ML CO2 Impact, n.d. URL

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. "Quantifying the Carbon Emissions of Machine Learning." arXiv preprint, submitted November 4, 2019. URL

Matthew Hutson. "Measuring AI's Carbon Footprint: New Tools Track and Reduce Emissions from Machine Learning." IEEE Spectrum, November 22, 2022. URL

Association for Computing Machinery (ACM). "TechBriefs: Computing and Climate Change." ACM Technology Policy Council, November 2021. URL

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green AI." Communications of the ACM 63, no. 12 (December 2020): 54–63. URL

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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Arvind Narayanan. "The limits of the quantitative approach to discrimination." 2022 James Baldwin lecture, Princeton University, October 11, 2022. URL

Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. “A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System.” CHI ‘20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, April 23, 2020, 1–15. URL

Rachel Thomas and David Uminsky. “Reliance on Metrics Is a Fundamental Challenge for AI.” Patterns 3, no. 5 (May 13, 2022): 100476. URL

Momin M. Malik. "A Hierarchy of Limitations in Machine Learning." arXiv preprint, submitted February 29, 2020. [URL](https://arxiv.org/abs/2002.05193

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 and 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

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL

References

ISO. "ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems." 2nd ed. ISO Standards, July 2019.

URL

Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub. URL

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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

ISO. "ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems." 2nd ed. ISO Standards, July 2019. URL

Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber. “Capability Maturity Model, Version 1.1.” IEEE Software 10, no. 4 (1993): 18–27. URL

Jeff Patton, Peter Economy, Martin Fowler, Alan Cooper, and Marty Cagan. User Story Mapping: Discover the Whole Story, Build the Right Product. O'Reilly, 2014. URL

Rumman Chowdhury and Jutta Williams. "Introducing Twitter’s first algorithmic bias bounty challenge." Twitter Engineering Blog, July 30, 2021. URL

HackerOne. "Twitter Algorithmic Bias." HackerOne, August 8, 2021. URL

Josh Kenway, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. "Bug Bounties for Algorithmic Harms?" Algorithmic Justice League, January 2022. URL

Microsoft. “Community Jury.” Microsoft Learn's Azure Application Architecture Guide, 2023. URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. "Overcoming Failures of Imagination in AI Infused System Development and Deployment." arXiv preprint, submitted December 10, 2020. URL

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 and 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

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL

References

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of Human Factors

and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

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

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125. URL

Batya Friedman, Peter H. Kahn, Jr., and Alan Borning. "Value Sensitive Design: Theory and Methods." University of Washington Department of Computer Science & Engineering Technical Report 02-12-01, December 2002. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL

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 and 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

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL

References

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019. URL

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125. URL

Steven Umbrello, and Ibo van de Poel. “Mapping Value Sensitive Design onto AI for Social Good Principles.” AI and Ethics 1, no. 3 (February 1, 2021): 283–96. URL

Karen Boyd. “Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82. URL

Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques.” In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11–40. URL

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

Shneiderman, Ben. “Human-Centered AI.” Issues in Science and Technology 37, no. 2 (2021): 56–61. URL

Shneiderman, Ben. “Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy.” IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

Caitlin Thompson. “Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides.” Coda, September 21, 2021. URL

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Algorithmic Decision-Making and the Control Problem.” Minds and Machines 29, no. 4 (December 11, 2019): 555–78. URL

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018. URL

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. “Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work.” Evaluation 15, no. 3 (2009): 285–306. URL

Asit Bhattacharyya and Lorne Cummings. “Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation.” Business Strategy and the Environment 24, no. 5 (2013): 309–25. URL

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. “A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation.” Healthcare 6, no. 3 (September 2018): 191–96. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL

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.

Transparency and Documentation

Organizations can document the following

- 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

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

References

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019. URL

Batya Friedman, David G. Hendry, and Alan Borning. "A Survey of Value Sensitive Design Methods." Foundations and Trends in Human-Computer Interaction 11, no. 2

(November 22, 2017): 63–125. URL

Steven Umbrello, and Ibo van de Poel. “Mapping Value Sensitive Design onto AI for Social Good Principles.” AI and Ethics 1, no. 3 (February 1, 2021): 283–96. URL

Karen Boyd. “Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82. URL

Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques.” In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11–40. URL

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

Shneiderman, Ben. “Human-Centered AI.” Issues in Science and Technology 37, no. 2 (2021): 56–61. URL

Shneiderman, Ben. “Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy.” IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

Caitlin Thompson. “Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides.” Coda, September 21, 2021. URL

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Algorithmic Decision-Making and the Control Problem.” Minds and Machines 29, no. 4 (December 11, 2019): 555–78. URL

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018. URL

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. “Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work.” Evaluation 15, no. 3 (2009): 285–306. URL

Asit Bhattacharyya and Lorne Cummings. “Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation.” Business Strategy and the Environment 24, no. 5 (2013): 309–25. URL

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. “A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation.” Healthcare 6, no. 3 (September 2018): 191–96. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL

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 and 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. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019. URL

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125. URL

Steven Umbrello, and Ibo van de Poel. “Mapping Value Sensitive Design onto AI for Social Good Principles.” AI and Ethics 1, no. 3 (February 1, 2021): 283–96. URL

Karen Boyd. “Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development.” FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82. URL

Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques.” In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11–40. URL

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

Shneiderman, Ben. “Human-Centered AI.” Issues in Science and Technology 37, no. 2 (2021): 56–61. URL

Shneiderman, Ben. “Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy.” IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

Caitlin Thompson. “Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides.” Coda, September 21, 2021. URL

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Algorithmic Decision-Making and the Control Problem.” Minds and Machines 29, no. 4 (December 11, 2019): 555–78. URL

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018. URL

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. “Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work.” Evaluation 15, no. 3 (2009): 285–306. URL

Asit Bhattacharyya and Lorne Cummings. “Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation.” Business Strategy and the Environment 24, no. 5 (2013): 309–25. URL

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. “A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation.” Healthcare 6, no. 3 (September 2018): 191–96. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'" arXiv preprint, submitted November 1, 2021. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL

HEADQUARTERS

100 Bureau Drive
Gaithersburg, MD 20899
301-975-2000

Manage

Risks are prioritized and acted upon based on a projected impact.

Expand All Collapse All

MANAGE 1.1

A determination is 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 and 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

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL

References

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

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 and 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

- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL
- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL

References

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and

Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

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 and 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

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and

Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

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 enables 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 and 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

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

References

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

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),
    - trained without access to ground truth (unsupervised, semi-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 and 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. URL
- Datasheets for Datasets. URL
- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL

References

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011). URL

David Wright. 2013. Making Privacy Impact Assessments More Effective. The Information Society, 29 (Oct 2013), 307-315. URL

Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. 2021. Datasheets for Datasets. arXiv:1803.09010. URL

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
  - 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 and 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

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

References

Safety, Validity and Reliability Risk Management Approaches and Resources

AI Incident Database. 2022. AI Incident Database. URL

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

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395. URL

Andrew L. Beam, Arjun K. Manrai, Marzyeh Ghassemi. 2020. Challenges to the Reproducibility of Machine Learning Models in Health Care. Jama 323, 4 (January 6, 2020), 305-306. URL

Anthony M. Barrett, Dan Hendrycks, Jessica Newman et al. 2022. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. arXiv:2206.08966. URL

Debugging Machine Learning Models, In Proceedings of ICLR 2019 Workshop, May 6, 2019, New Orleans, Louisiana. URL

Jessie J. Smith, Saleema Amershi, Solon Barocas, et al. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. arXiv:2205.08363. URL

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, et al. 2020. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program) arXiv:2003.12206. URL

Kirstie Whitaker. 2017. Showing your working: a how to guide to reproducible research. (August 2017). LINK, URL

Netflix. Chaos Monkey. URL

Peter Henderson, Riashat Islam, Philip Bachman, et al. 2018. Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence. 32, 1 (Apr. 2018). URL

Suchi Saria, Adarsh Subbaswamy. 2019. Tutorial: Safe and Reliable Machine Learning. arXiv:1904.07204. URL

Kang, Daniel, Deepti Raghavan, Peter Bailis, and Matei Zaharia. "Model assertions for monitoring and improving ML models." Proceedings of Machine Learning and Systems 2 (2020): 481-496. URL

Managing Risk Bias

National Institute of Standards and Technology (NIST), Reva Schwartz, Apostol Vassilev, et al. 2022. NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. URL

Bias Testing and Remediation Approaches

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, et al. 2018. A Reductions Approach to Fair Classification. arXiv:1803.02453. URL

Brian Hu Zhang, Blake Lemoine, Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. arXiv:1801.07593. URL

Drago Plečko, Nicolas Bennett, Nicolai Meinshausen. 2021. Fairadapt: Causal Reasoning for Fair Data Pre-processing. arXiv:2110.10200. URL

Faisal Kamiran, Toon Calders. 2012. Data Preprocessing Techniques for Classification without Discrimination. Knowledge and Information Systems 33 (2012), 1–33. URL

Faisal Kamiran; Asim Karim; Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, December 10-13, 2012, Brussels, Belgium. IEEE, 924-929. URL

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, et al. 2017. Optimized Data Pre-Processing for Discrimination Prevention. arXiv:1704.03354. URL

Geoff Pleiss, Manish Raghavan, Felix Wu, et al. 2017. On Fairness and Calibration. arXiv:1709.02012. URL

L. Elisa Celis, Lingxiao Huang, Vijay Keswani, et al. 2020. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. arXiv:1806.06055. URL

Michael Feldman, Sorelle Friedler, John Moeller, et al. 2014. Certifying and Removing Disparate Impact. arXiv:1412.3756. URL

Michael Kearns, Seth Neel, Aaron Roth, et al. 2017. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. arXiv:1711.05144. URL

Michael Kearns, Seth Neel, Aaron Roth, et al. 2018. An Empirical Study of Rich Subgroup Fairness for Machine Learning. arXiv:1808.08166. URL

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), 2016, Barcelona, Spain. URL

Rich Zemel, Yu Wu, Kevin Swersky, et al. 2013. Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning 2013, PMLR 28, 3, 325-333. URL

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh & Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Peter A. Flach, Tijl De Bie, Nello Cristianini (eds) Machine Learning and Knowledge Discovery in Databases. European Conference ECML PKDD 2012, Proceedings Part II, September 24-28, 2012, Bristol, UK. Lecture Notes in Computer Science 7524. Springer, Berlin, Heidelberg. URL

Security and Resilience Resources

FTC Start With Security Guidelines. 2015. URL

Gary McGraw et al. 2022. BIML Interactive Machine Learning Risk Framework. Berryville Institute for Machine Learning. URL

Ilia Shumailov, Yiren Zhao, Daniel Bates, et al. 2021. Sponge Examples: Energy-Latency Attacks on Neural Networks. arXiv:2006.03463. URL

Marco Barreno, Blaine Nelson, Anthony D. Joseph, et al. 2010. The Security of Machine Learning. Machine Learning 81 (2010), 121-148. URL

Matt Fredrikson, Somesh Jha, Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15), October 2015. Association for Computing Machinery, New York, NY, USA, 1322–1333. URL

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

Nicolas Papernot. 2018. A Marauder's Map of Security and Privacy in Machine Learning. arXiv:1811.01134. URL

Reza Shokri, Marco Stronati, Congzheng Song, et al. 2017. Membership Inference Attacks against Machine Learning Models. arXiv:1610.05820. URL

Adversarial Threat Matrix (MITRE). 2021. URL

Interpretability and Explainability Approaches

Chaofan Chen, Oscar Li, Chaofan Tao, et al. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. arXiv:1806.10574. URL

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. arXiv:1811.10154. URL

Daniel W. Apley, Jingyu Zhu. 2019. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv:1612.08468. URL

David A. Broniatowski. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8367. National Institute of Standards and Technology, Gaithersburg, MD. URL

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, et al. 2021. Manipulating and Measuring Model Interpretability. arXiv:1802.07810. URL

Hongyu Yang, Cynthia Rudin, Margo Seltzer. 2017. Scalable Bayesian Rule Lists. arXiv:1602.08610. URL

P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, et al. 2021. Four Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8312. National Institute of Standards and Technology, Gaithersburg, MD. URL

Scott Lundberg, Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874. URL

Susanne Gaube, Harini Suresh, Martina Raue, et al. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digital Medicine 4, Article 31 (2021). URL

Yin Lou, Rich Caruana, Johannes Gehrke, et al. 2013. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13), August 2013. Association for Computing Machinery, New York, NY, USA, 623–631. URL

Privacy Resources

National Institute for Standards and Technology (NIST). 2022. Privacy Framework. URL

Data Governance

Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, Tomasz Janowski, Data governance: Organizing data for trustworthy Artificial Intelligence, Government Information Quarterly, Volume 37, Issue 3, 2020, 101493, ISSN 0740-624X. URL

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 apriori. 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 and 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

- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

- WEF - Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations. URL
- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL

References

AI Incident Database. 2022. AI Incident Database. URL

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

Andrew Burt and Patrick Hall. 2018. What to Do When AI Fails. O'Reilly Media, Inc. (May 18, 2020). Retrieved October 17, 2022. URL

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

SANS Institute. 2022. Security Consensus Operational Readiness Evaluation (SCORE) Security Checklist [or Advanced Persistent Threat (APT) Handling Checklist]. URL

Suchi Saria, Adarsh Subbaswamy. 2019. Tutorial: Safe and Reliable Machine Learning. arXiv:1904.07204. 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 and 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

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

References

Decommissioning 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 3.1

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

References

Office of the Comptroller of the Currency. 2021. Proposed Interagency Guidance on Third-Party Relationships: Risk Management. July 12, 2021. URL

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 and 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. URL

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

References

Larysa Visengeriyeva et al. “Awesome MLOps,“ GitHub. Accessed January 9, 2023. URL

MANAGE 4.1

Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.

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 and 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, URL
- Datasheets for Datasets. URL

References

Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing." Information 11, no. 3 (2020): 137. URL

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 and 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

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

References

Yen, Po-Yin, et al. "Development and Evaluation of Socio-Technical Metrics to Inform HIT Adaptation." URL

Carayon, Pascale, and Megan E. Salwei. "Moving toward a sociotechnical systems approach to continuous health information technology design: the path forward for improving electronic health record usability and reducing clinician burnout." Journal of the American Medical Informatics Association 28.5 (2021): 1026-1028. URL

Mishra, Deepa, et al. "Organizational capabilities that enable big data and predictive analytics diffusion and organizational performance: A resource-based perspective." Management Decision (2018).

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 and 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

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

References

Wei, M., & Zhou, Z. (2022). AI Ethics Issues in Real World: Evidence from AI Incident Database. ArXiv, abs/2206.07635. URL

McGregor, Sean. "Preventing repeated real world AI failures by cataloging incidents: The AI incident database." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 17. 2021. URL

Macrae, Carl. "Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk." Risk analysis 42.9 (2022): 1999-2025. URL