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.