UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

CLTC WHITE PAPER SERIES

A Taxonomy of Trustworthiness for Artificial Intelligence

CONNECTING PROPERTIES OF TRUSTWORTHINESS WITH RISK MANAGEMENT AND THE AI LIFECYCLE

JESSICA NEWMAN

CLTC WHITE PAPER SERIES

A Taxonomy of Trustworthiness for Artificial Intelligence

CONNECTING PROPERTIES OF TRUSTWORTHINESS WITH RISK MANAGEMENT AND THE AI LIFECYCLE

This resource includes the Taxonomy of Trustworthiness for Artificial Intelligence, which is discussed in more detail in the full paper available here: https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf. The full paper also includes discussion of the term "trustworthy AI," compares multiple existing frameworks for trustworthy AI that informed this work, introduces the concept of properties of trustworthiness for AI, and includes in the appendix connections to international AI standards and a list of the properties of trustworthiness without segmentation by lifecycle stage.

JESSICA NEWMAN

JANUARY 2023

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Introduction

The National Institute of Standards and Technology (NIST) has developed an AI Risk Management Framework (RMF) intended to promote trustworthy artificial intelligence (AI). In this report, we introduce a taxonomy of trustworthiness for artificial intelligence that is intended to complement and support the use of the NIST AI RMF.

The taxonomy includes 150 properties of trustworthiness for AI. Each property builds upon a relevant "characteristic of trustworthiness" as defined by NIST in the AI RMF. NIST's characteristics of trustworthiness include: valid and reliable; safe, secure, and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; and fair with harmful bias managed.

The taxonomy is organized by the seven stages of the AI lifecycle depicted in the NIST AI RMF:

- Plan and Design
- Collect and Process Data
- Build and Use Model
- Verify and Validate
- Deploy and Use
- Operate and Monitor
- Use or Impacted By

Our hope is that this framework supports usability by connecting the taxonomy more closely to actual product cycles and workflows. We also hope to provide ideas about possible ways to connect the NIST AI RMF Core to the AI lifecycle.

Within each stage of the lifecycle, the taxonomy includes NIST's seven characteristics of trustworthiness. These categories are further broken down to include all the properties of trustworthiness that are relevant at that stage of the lifecycle.

We also include an eighth characteristic of trustworthiness in the taxonomy, which varies slightly from the NIST AI RMF. The additional characteristic is "Responsible Practice and Use." The NIST AI RMF recognizes its importance and states that "AI risk management can drive responsible uses and practices," but does not include it as a characteristic of trustworthiness. In this paper, we include it as a crosscutting characteristic of trustworthiness because we find

that it serves a critical role in highlighting the interconnected nature of AI technologies with the people, organizations, and structures that are designing, building, and deploying them. We use Responsible Practice and Use in this taxonomy to promote consistent understanding of AI as a sociotechnical system, situated within structures of practice and use.

Each property is accompanied by a set of questions to guide initial thinking. For example, the "Data Protection" property includes the question, "How will we use encryption, differential privacy, federated learning, data minimization, and/or other best practices to protect data?" The questions are formulated in this future-oriented way (and not as "Have we . . . ?") because they are intended to serve as a tool to spark further discussion and action, rather than as a checklist or a scorecard.

Each property included in the taxonomy is also tagged with a set of subcategories from the NIST AI RMF. These subcategories represent the most relevant sections of the NIST AI RMF Core framework. Reviewing these subcategories will point a reader to helpful resources and tools to address the property. A small number of the listed subcategories (in most cases just one or two) are bolded to emphasize that they are likely to be particularly helpful or a good place to start. There may be additional subcategories not listed here that are also relevant, depending on the context of an AI system's development and use.

Finally, the taxonomy was developed to be useful for understanding a full spectrum of AI systems, including those that have limited engagement with people, which have typically been underemphasized in considerations of AI trustworthiness. A subset of the properties of trustworthiness in the taxonomy are likely to only be relevant to AI systems that are human-facing, which may engage directly with human users or operators, make use of human data, or inform human decision-making. These properties are marked in the table with an asterisk after their name. Properties that do not have an asterisk are likely to be relevant to AI systems across the spectrum of human-AI engagement.

This taxonomy aims to provide a resource that is useful for AI organizations and teams developing AI technologies, systems, and applications. While it is designed specifically to assist users of the NIST AI RMF, it could also be helpful for people using any kind of AI risk or impact assessment, or for people developing model cards, system cards, or other types of AI documentation. It may also be useful for standards-setting bodies, policymakers, independent auditors, and civil society organizations working to evaluate and promote trustworthy AI.

FULL REPORT AVAILABLE HERE

Taxonomy of Trustworthiness for Artificial Intelligence

AI Lifecycle Stage: Plan and Design

The purpose of the plan and design stage is to articulate and document the system's concept and objectives, underlying assumptions, context, and requirements.

Properties followed by an asterisk may be less relevant for AI systems that are not human-facing, meaning they do not engage directly with human users or operators, make use of human data, or inform human decision-making.

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Valid and Reliable	Fit for Purpose	How will we assess whether the AI system is fit for purpose for each intended use and provides a valid solution for the problems we are trying to solve? How will we ensure that inappropriate uses are rejected?	Govern 5.1 **Map 1.1** Map 1.2 Map 1.3 Map 1.4 Map 3.1 Map 3.2 Map 3.3 **Manage 1.1**
	Predictable and Dependable	How will we ensure that the AI system will behave as expected? If the AI system is not fully predictable, how will we assess whether it can still be depended upon for our purposes?	Govern 4.1 Govern 4.2 Govern 4.3 Map 2.2 **Map 2.3** Measure 2.3 Measure 2.4 **Measure 2.5** Measure 2.6 Measure 2.7 Manage 2.4 Manage 4.1
	Appropriate Level of Automation	How will we determine the desired and appropriate degree of automation, given the AI system's characteristics and the context of its uses?	**Govern 3.2** **Map 1.1** Map 1.2 Map 1.3 **Map 2.2** **Map 3.5** Measure 4.2 Manage 1.1 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	High Quality AI System Configuration	How will we assess the quality of the AI system design and configuration and ensure consistently high quality? For example, how will we assess and ensure the quality of all of the software components integrated into the AI system? How will we assess and ensure the quality of the hardware for the AI system, such as AI chips, including graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs)?	Govern 4.3 Govern 5.2 Govern 6.1 Map 1.6 **Map 2.3** Measure 1.3 **Measure 2.3** Measure 3.3 Manage 3.1 Manage 3.2 Manage 4.1
	High Quality Network Resources and Services	How will we assess and ensure the quality of shared network resources and services, e.g., distributed dataset access?	**Govern 6.1** Govern 6.2 Map 2.3 Map 4.1 Map 4.2 Measure 2.3 Measure 2.4 Measure 3.1 Measure 3.3 **Manage 3.1** Manage 3.2
	Trusted Dependencies on External Parties	How will we identify, assess, and monitor our dependencies on external parties?	**Govern 6.1** Govern 6.2 Map 4.1 Map 4.2 **Manage 3.1** Manage 3.2
	Foresight and Scenario Planning	How will we assess and navigate possible futures and the evolving risk landscape?	Govern 3.1 **Govern 4.1** Map 1.1 Map 1.2 Map 3.1 Map 3.2 Measure 3.1 **Measure 3.2**
Safe	Protection of Physical and Psychological Safety	How will we ensure that the AI system will not cause physical or psychological harm or lead to a state in which human life, health, property, or the environment is endangered? How will we anticipate potential failure modes or unsafe conditions?	Govern 1.7 **Govern 4.1** **Govern 4.2** **Govern 4.3** Govern 5.1 Govern 5.2 Govern 6.2 Map 1.1 Measure 1.2 Measure 1.3 **Measure 2.6** Measure 3.1 Measure 3.3 **Manage 2.4** **Manage 4.1**

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Assurance / Management of Uncertainty	If we do not know all of the elements required for the safe development and deployment of the AI system, how will we manage this uncertainty?	**Govern 4.1** Govern 4.2 Govern 4.3 Measure 2.6 Measure 3.2 Manage 2.3 Manage 4.1
	Assurance / Management of Multi-Capability / Multi-Modal Systems	If an AI system has multiple capabilities or works across multiple modalities, how will we document and manage this complexity?	Govern 4.1 Govern 4.2 Govern 4.3 **Map 1.1** **Map 2.2** Map 3.3 **Measure 3.1** Measure 3.2 Measure 3.3 Manage 2.3 Manage 2.4 Manage 4.1
	Alignment with Human Values	How will we ensure that the AI system abides by desired human values and does not sacrifice human values to achieve its narrow goals?	Govern 3.1 Govern 4.1 Govern 4.2 **Map 1.1** Map 1.2 Map 1.6 Map 3.5 Measure 2.6 Measure 3.1 Measure 3.3 Manage 4.1
	Governable	How will we ensure an AI system is designed and engineered to achieve its goals while maintaining the ability to disengage or deactivate the system if necessary? How will we ensure an AI system would not have incentives to resist or deceive its operators?	Govern 4.1 Map 2.2 Measure 2.4 Measure 2.5 Measure 2.6 **Manage 2.4**
Fair with Harmful Bias Managed	Diverse	How will we ensure that gender, racial, age, ability, religious, cultural, disciplinary, and other relevant types of diversity are represented within the teams influencing AI development and use, throughout all stages of the AI lifecycle?	Govern 2.1 **Govern 3.1** Govern 5.1 **Map 1.2** Measure 1.3 Measure 2.2 Measure 4.2

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Inclusive	How will we ensure inclusivity of all relevant experts and communities in the design and development of the AI system?	Govern 2.1 **Govern 3.1** Govern 3.2 Govern 5.1 **Govern 5.2** **Map 1.2** Measure 1.3 Measure 2.2 Measure 3.3 Measure 4.2 Manage 4.2
	Equitable	How will we navigate structural power dynamics and promote equity in the design and use of the AI system? (For example, how will different communities be given power to influence decisions? Who will experience potential benefits of the AI system and who will experience potential harms?)	Govern 3.1 **Govern 5.1** **Govern 5.2** **Map 1.1** **Map 1.2** Map 5.1 Map 5.2 Measure 1.2 Measure 1.3 Measure 2.2 **Measure 2.11** **Measure 3.3** Measure 4.3 **Manage 4.1**
	Just	How will we ensure justice in the design and use of the AI system? (For example, are all the people involved in the training, design, and development of the AI system treated fairly, even in less visible roles, such as data annotators?)	Govern 3.1 Govern 4.2 Govern 4.3 **Govern 5.1** **Govern 5.2** **Map 1.1** **Map 1.2** Map 5.1 Map 5.2 Measure 1.2 Measure 1.3 Measure 2.2 **Measure 2.11** **Measure 3.3** Measure 4.3 **Manage 4.1** Manage 4.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Mitigation of Systemic and Human Bias	How will we assess and mitigate ways in which systemic and human bias may influence the design, development, and deployment of the AI system?	Govern 2.2 Govern 3.1 Govern 4.3 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 Map 5.1 Map 5.2 Measure 1.3 Measure 2.11 Measure 3.3 Measure 4.3 Manage 4.1
	Solidarity	How will we ensure the design and use of the AI system respects the solidarity of groups and communities, such as workers, women, people with disabilities, ethnic minorities, children, or others?	**Govern 3.1** Govern 3.2 Govern 4.2 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 Map 5.1 Map 5.2 Measure 1.3 Measure 3.3 Measure 4.3 Manage 4.1
Secure and Resilient	Security-by-Design	How will we build security into the AI system design, testing, deployment, and operation? How often will we provide security updates to the AI system?	**Govern 4.1** Govern 4.2 Govern 4.3 Govern 6.1 Map 1.1 Map 1.6 Map 2.3 Map 4.2 **Measure 2.7** Manage 2.4
	Availability	How will we ensure that information for and about the AI system is available to authorized personnel when it is needed?	Govern 4.1 Govern 4.3 Govern 6.1 Govern 6.2 Map 1.1 Map 2.3 **Measure 2.7** Measure 2.9
	Confidentiality	How will we ensure that information is not made available or disclosed to unauthorized individuals, entities, or processes?	Govern 4.1 Govern 4.3 Govern 6.1 Govern 6.2 Map 1.1 Map 2.3 **Measure 2.7** **Measure 2.10**

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Integrity	How will we maintain and ensure the accuracy, completeness, and appropriateness of data, models, and procedures informing the AI system?	Govern 4.1 Govern 4.3 Govern 6.1 Govern 6.2 Map 1.1 Map 2.3 **Measure 2.7** Measure 2.9
Explainable and Interpretable	Intelligible*	How will we assess the system for intelligible explanations and select a model to support this?	Map 1.1 Map 2.2 **Measure 2.9**
	Positive Human-Machine Interaction*	How will we enable positive human-machine interactions throughout the AI system's operation?	Govern 3.2 Map 1.1 Map 1.2 Map 2.2 **Map 3.5** Map 5.2 **Measure 2.9**
Privacy-Enhanced	Privacy-by-Design*	How will privacy be built into the AI system design, testing, deployment, and operation? If data includes sensitive or personally identifiable information including biometrics, what extra precautions will be taken?	Govern 1.1 Govern 1.2 Govern 6.1 Map 1.1 Map 4.1 **Measure 2.10**
	Data Privacy or Protection Impact Assessment*	What is the impact of the AI system on privacy? When and how will we conduct a data privacy or data protection impact assessment?	Govern 1.1 Govern 1.2 Govern 6.1 Map 1.1 Map 4.1 **Measure 2.10** Manage 4.1
Accountable and Transparent	Effective Policy and Governance	How will we analyze and follow or implement relevant or desired AI and data standards, policies, principles, and guidance?	Govern 1.1 Govern 1.2 Govern 1.3 Map 3.5 Map 4.1 Map 5.1 Map 5.2 Measure 1.1 Measure 1.2 Measure 1.3 Measure 2.8 **Manage 1.3** Manage 2.1 Manage 3.1 Manage 4.1
	Adherence to the Rule of Law	How will we analyze and ensure compliance with all relevant laws and regulations across every jurisdiction of use? How will we analyze liability considerations, and what precautions will be taken?	**Govern 1.1** Map 4.1 Manage 1.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Coordination (Public-Private; International)	How will we identify and coordinate with relevant institutions, nationally and internationally?	Govern 5.1 Govern 5.2 **Map 5.2** Measure 1.3 Measure 4.1 Measure 4.2 Measure 4.3
	Effective Risk Assessments and Impact Assessments	How will we assess, document, and communicate (on a regular basis) the expected, potential, and actual risks and impacts of the AI system on people, organizations, and society (pre- and post-deployment)? If risks and impact are deemed to be unacceptable, how will we ensure the AI system is adjusted or rejected?	Govern 1.3 **Govern 1.4** Govern 1.7 Govern 6.1 Map 1.1 **Map 3.2** Map 5.1 Map 5.2 Measure 1.1 Measure 1.3 **Manage 1.1** Manage 1.2 Manage 1.3 Manage 1.4 Manage 2.1 Manage 2.3 Manage 2.4
	Community Engagement	How will we identify communities interested in, engaged in, or impacted by the AI system, and how will we encourage their participation throughout the AI lifecycle?	Govern 5.1 **Govern 5.2** Map 1.2 **Map 5.2** **Measure 3.3** Measure 4.1 Measure 4.2 Measure 4.3 Manage 4.2
	Open	How can we promote openness and transparency about our development and governance of AI technologies, internally and externally?	Govern 1.2 Govern 1.4 Govern 1.6 **Govern 4.2** Govern 4.3 Map 5.2 Measure 1.3 Measure 2.9 **Measure 2.8** **Manage 4.3**

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Documentation	How will we document the AI system's design, datasets, training, characteristics, capabilities, limitations, predictable failures, intended uses, etc.? How will we review and update the documentation on a regular basis and as needed to document new uses, functionalities, etc.?	Govern 1.6 **Govern 4.2** Map 1.1 Map 2.3 Map 3.1 Map 3.2 Map 3.3 Map 3.4 Map 3.5 Map 4.1 Map 4.2 Map 5.1 Map 5.2 Measure 2.9 **Measure 2.8**
	Internal Reporting / Culture of Safety	How will we incentivize internal reporting of challenges or concerns, and promote a culture of safety among teams involved with the AI system and in general?	Govern 1.2 Govern 2.2 Govern 2.3 **Govern 4.1** Govern 4.2 Govern 4.3 Measure 2.8
	Internal Reviews	How will internal reviews be conducted to assess trustworthy AI practices?	**Govern 1.5** Govern 4.1 Govern 4.2 Govern 4.3 Measure 2.8 Measure 2.13
Responsible Practice and Use	Responsible Use in Government, Education, Health, Finance, Workplace, Identification and Detection, and other High-stakes Settings	How will we ensure responsible potential and actual uses in high-stakes settings, such as government, education, healthcare, finance, employment, workplace, identification and detection (such as emotion detection), and others? If our AI system influences one of these domains, how will we ensure that we engage sufficiently with domain experts and impacted communities to better understand the influence and impact we might have?	A majority of all of the subcategories are critical. **Map 1.1** is especially relevant to help understand the purpose, context, and impacts of the intended use.
	Responsible Use in Critical Infrastructure and Safety-Critical Systems	How will we ensure responsible potential and actual uses for critical infrastructure and safety-critical systems, including assessing the potential for damaging effects from technical faults, defects, or attacks?	A majority of all of the subcategories are critical. **Map 1.1** is especially relevant to help understand the purpose, context, and impacts of the intended use.
	Responsible Use in the Criminal Legal System and by Law Enforcement	How will we ensure responsible potential and actual uses in the criminal legal system or by law enforcement? For example, how will we protect against abuses of biometric identification in public spaces?	A majority of all of the subcategories are critical. **Map 1.1** is especially relevant to help understand the purpose, context, and impacts of the intended use.

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Responsible Use in Defense and National Security	How will we promote peace and ensure responsible and controlled uses for defense, military, border control, and national security purposes, including for weapons systems?	A majority of all of the subcategories are critical. **Map 1.1** is especially relevant to help understand the purpose, context, and impacts of the intended use.
	Verified Supply Chain	How will we assess and verify the relevant components of the supply chain?	**Govern 6.1** Govern 6.2 **Map 4.1** Map 4.2 **Manage 3.1** Manage 3.2
	Appropriate Assignment of Organizational Roles, Authorities, and Responsibilities; Designated Points of Contact	How will we assign and document organizational roles, authorities, and responsibilities? How will we designate points of contact along the lifecycle?	**Govern 2.1** Govern 2.2 Govern 2.3 Govern 3.1 Govern 3.2 Map 3.4 Map 3.5 Manage 2.1
	Effective Capabilities	How will we obtain the necessary resources and knowledge to achieve our trustworthy AI objectives?	Govern 2.2 **Map 3.4**
	Collaboration	How will we enable multi-stakeholder collaboration?	Govern 3.1 **Govern 5.1** Govern 5.2 **Map 1.2** **Map 5.2** Manage 4.2
	Supportive Governance and Organizational Structure	How can our governance and organizational structure support trustworthy AI? How do our strategy, objectives, and policies support trustworthy AI? Are changes needed?	Govern 1.1 **Govern 1.2** Govern 1.3 Govern 1.4 Govern 1.5 Govern 1.6 Govern 1.7 Govern 2.1 Govern 2.2 Govern 2.3 Govern 4.1 Govern 4.2 Govern 4.3
	Effective Hiring and Training	How will we support the hiring and training of individuals who can carry out trustworthy AI objectives?	**Govern 2.1** **Govern 2.2** Govern 2.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Responsible Labor Practices and Rights	How can we support labor rights in our use of AI? How will the supply chain of the AI system be monitored to evaluate working conditions?	Govern 1.1 Govern 1.2 Govern 2.1 Govern 6.1 **Map 1.1** Map 3.4 Map 5.2
	Leadership Commitment	How will we ensure long-term commitment to trustworthy AI from organizational leadership?	Govern 2.1 **Govern 2.3**
	Supportive Organizational Culture	How will our organizational culture support our trustworthy AI objectives? Are changes needed?	**Govern 1.2** Govern 1.4 Govern 2.2 Govern 2.3 Govern 4.1
	Procurement Standards	How will we implement/ensure AI procurement standards that support trustworthy AI if we are procuring the AI system or providing it to others?	Govern 1.2 Govern 4.2 **Govern 6.1** Map 1.3 Map 1.4 **Map 4.1** **Map 4.2**
	Appropriate Relationships, Interdependencies, and Interconnections	What relationships, interdependencies, and interconnections will be involved in the development and use of the AI system, and how do they intersect with our trustworthy AI objectives?	**Map 1.1** Map 4.1 Manage 3.1
	Alignment with Organizational Vision, Mission, and Values	How will we ensure the AI system is true to our vision, mission, and values?	Govern 4.3 Govern 5.1 **Map 1.1** Map 5.2 Measure 4.2 Manage 1.1
	Socially Responsible	How will our AI system and its use align with our social responsibility efforts?	**Govern 1.2** Govern 4.1 **Map 1.1** Map 5.1 Map 5.2 Manage 1.1
	Supportive of Fair Competition	How will we support fair competition among a variety of actors in the domain in which our AI system is applied?	Govern 5.1 Govern 5.2 **Map 1.1** Map 5.2 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Supportive of Civil Rights	How will we protect and promote civil rights throughout the AI lifecycle, including protection from unlawful discrimination on the basis of race, color, national origin, disability, age, religion, and sex (including pregnancy, sexual orientation, and gender identity)?	Govern 1.1 Govern 1.2 Govern 4.2 Govern 4.3 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 Map 5.1 Map 5.2 Measure 1.3 Measure 2.2 Measure 2.11 Measure 3.3 Manage 1.3 Manage 3.1 Manage 4.1 Manage 4.3
	Supportive of Democratic Values and Processes	How will we ensure the design and use of the AI system are consistent with democratic values such as freedom and equality? How will we ensure that the uses of the AI system do not interfere with democratic processes and citizens' rights, including the right to vote? How will we assess the impact of the AI system on democracy?	Govern 1.1 Govern 1.2 Govern 4.2 Govern 4.3 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 Map 5.1 Map 5.2 Measure 1.3 Manage 1.3 Manage 3.1 Manage 4.1 Manage 4.3
	Protection of Human Autonomy and Freedom	How will we ensure that the AI system respects the freedom and autonomy of individuals and does not intrude on people's self-determination and ability to make life decisions for themselves?	Govern 1.1 Govern 1.2 Govern 4.2 Govern 4.3 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 **Map 3.5** Map 5.1 Map 5.2 Measure 1.3 Manage 1.3 Manage 3.1 Manage 4.1 Manage 4.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Protection of Human Dignity	How will we ensure that the development and use of the AI system respect human dignity and treat people as having intrinsic worth, and not merely as objects?	Govern 1.1 Govern 1.2 Govern 4.2 Govern 4.3 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 **Map 3.5** Map 5.1 Map 5.2 Measure 1.3 Manage 1.3 Manage 3.1 Manage 4.1 Manage 4.3
	Protection of Human Rights	How will we ensure the AI system does not threaten human rights? For example, how will we ensure the right to privacy? How will we ensure the AI system does not pose risks of gender or sexual violence? How will we ensure it does not threaten children's rights? How will we ensure the AI system does not threaten freedom of religion, or freedom of expression? How will we ensure the AI system does not threaten the right to fair trial or the right of peaceful assembly?	Govern 1.1 Govern 1.2 Govern 4.2 Govern 4.3 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 **Map 3.5** Map 5.1 Map 5.2 Measure 1.3 Manage 1.3 Manage 3.1 Manage 4.1 Manage 4.3
	Supportive of Wellbeing	How will we ensure the AI system supports individual, community, and societal wellbeing, including mental or emotional wellbeing?	Govern 1.1 Govern 1.2 Govern 4.2 Govern 4.3 Govern 5.1 Govern 5.2 **Map 1.1** Map 1.2 Map 3.5 Map 5.1 Map 5.2 Measure 1.3 Manage 1.3 Manage 3.1 Manage 4.1 Manage 4.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Reduction of Carbon Emissions	How can we reduce the carbon emissions from the design and use of AI systems in general?	Govern 1.7 Govern 4.2 Map 1.1 Map 5.1 Map 5.2 Measure 1.3 **Measure 2.12** Manage 4.1
	Assessment of Economic, Social, Cultural, Political, and Global Implications	How will we assess the economic implications of the AI system, including whether use of the system could impact jobs or reduce the need for human labor? How will we assess the social, cultural, and political implications of the AI system at the societal and global levels?	Govern 3.1 **Govern 5.1** **Map 1.1** Map 1.2 Map 3.1 Map 3.2 Map 5.1 Map 5.2 Measure 1.3

AI LIFECYCLE STAGE: COLLECT AND PROCESS DATA

The purpose of this stage is to collect and process data, including to gather, validate, and clean data and document the metadata and characteristics of the dataset.

Properties followed by an asterisk may be less relevant for AI systems that are not human-facing, meaning they do not engage directly with human users or operators, make use of human data, or inform human decision-making.

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Valid and Reliable	Data Completeness	How will we assess and improve the completeness, quantity, suitability, and representativeness of the data?	Govern 4.3 Govern 5.1 Govern 6.1 Map 1.1 Map 1.2 **Map 2.3** **Measure 2.2** Measure 2.11 Manage 3.1 Manage 3.2
	Data Quality	How will we assess and improve the quality and relevance of the data? What benchmarks will we use? How will we collect and process data, for example to annotate, label, clean, and aggregate as needed?	Govern 4.3 Govern 5.1 Govern 6.1 Map 1.1 Map 1.2 **Map 2.3** Measure 2.2 Manage 1.1 Manage 3.1 Manage 3.2
	Responsible Data, Information Systems and Information Flows	How will we obtain data, and what are our informational flows? How will we appropriately limit the scope of our data collection? How will we retain and delete data as needed?	Govern 1.1 Govern 1.2 Govern 1.4 Govern 4.3 Govern 5.1 Govern 6.1 Govern 6.2 Map 1.1 Map 1.2 **Map 2.3** **Map 4.1** Measure 2.2 Measure 2.10 Manage 3.1 Manage 3.2 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Safe	Data Stability	How will we analyze and monitor for data drift over time?	Govern 4.3 Govern 5.1 Govern 6.1 Govern 6.2 **Map 2.3** Measure 2.7 **Measure 3.1** Manage 3.2 **Manage 4.1**
Fair with Harmful Bias Managed	Data Balance*	How will we assess and improve the balance and diversity of the data? How will we evaluate all data sets for inclusion and representation of demographic groups? How will we guard against proxies for demographic information that could contribute to discrimination?	Govern 3.1 Govern 4.3 Govern 5.1 Govern 6.1 Map 1.1 Map 1.2 **Map 2.3** **Measure 2.2** Measure 2.11 Manage 3.1 Manage 3.2 Manage 4.1
Secure and Resilient	Data Security	How will the security of data that is used for training or created be ensured?	Govern 4.3 Govern 5.1 Govern 6.1 Govern 6.2 **Map 2.3** Map 4.1 Map 4.2 **Measure 2.7** Manage 3.1 Manage 3.2 Manage 4.1
Privacy-Enhanced	Data Protection*	How will we protect the data used to build and operate the AI system? How will we use encryption, differential privacy, federated learning, data minimization, and/or other best practices to protect data?	Govern 4.3 Govern 5.1 Govern 6.1 **Map 2.3** Map 4.1 Map 4.2 Measure 2.7 **Measure 2.10** Manage 3.1 Manage 3.2 Manage 4.1
	Data Processing Oversight*	How will we establish data oversight mechanisms, such as limiting and logging data access?	Govern 4.3 Govern 5.1 Govern 6.1 Govern 6.2 **Map 2.3** **Map 4.1** Measure 2.10 Measure 2.8 Manage 3.1 Manage 3.2 Manage 4.1 Manage 4.2 Manage 4.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Consent to Use of Data*	How will we enable people to consent to the uses of their data?	Govern 1.1 Govern 5.2 Map 4.1 Map 5.2 Measure 2.8 **Manage 4.1** Manage 4.2 Manage 4.3
	Control of Use of Data*	How will we ensure people have a say in how information about them is used? How will we honor the right to rectification and the right to erasure?	Govern 1.1 **Govern 5.2** Map 4.1 Map 5.2 **Manage 4.1** Manage 4.2 Manage 4.3
Accountable and Transparent	Data Governance*	How will we analyze and follow data governance practices for all intended uses, stakeholders, and relevant geographic areas? How will we ensure data rights and agency?	**Govern 1.1** Govern 1.4 Govern 6.1 Govern 6.2 Map 1.1 Map 1.2 Map 1.3 **Map 2.3** Map 4.1 Map 4.2 Map 5.1 Map 5.2 Measure 2.2 Measure 2.11 Measure 2.10 **Manage 1.3** Manage 3.1 Manage 3.2 Manage 4.1
	Traceable	How will we document the provenance of data, processes, and artifacts involved in the production of the AI system?	Govern 1.6 **Govern 4.2** **Map 1.1** **Map 2.3** Map 4.1 Measure 2.1 Measure 2.2 Measure 2.8
Responsible Practice and Use	Efficient Data Centers	How can we make our use of data centers more energy-efficient?	Map 1.1 **Measure 2.12**

AI LIFECYCLE STAGE: BUILD AND USE MODEL

The purpose of the "build and use model" stage is to create, select, and train models or algorithms.

Properties followed by an asterisk may be less relevant for AI systems that are not human-facing, meaning they do not engage directly with human users or operators, make use of human data, or inform human decision-making.

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Valid and Reliable	Accurate	How will we assess the accuracy of what the model has learned using an interpretation method (descriptive accuracy)? How will we assess the accuracy of the underlying data relationships with the model (predictive accuracy)? What benchmarks will we use? How will we communicate this as needed?	Govern 4.3 Map 1.1 Map 2.2 Map 2.3 **Measure 2.3** **Measure 2.5** Manage 1.1 Manage 4.1
	Reproducible	How will we test whether desirable outputs of the AI system can be reproduced in different circumstances?	Govern 4.3 Govern 5.1 Map 2.1 Map 2.2 **Map 2.3** Measure 2.1 **Measure 2.3** **Measure 2.5** Manage 1.1
	Efficient	How will we improve the efficiency of the AI system in terms of its energy and power usage, model size, and memory consumption? How can we make the model architecture of the AI system more efficient?	Govern 4.3 Map 2.1 Map 2.2 Map 2.3 **Measure 2.3** Measure 2.4 Measure 2.5 **Measure 2.12**
Safe	Safely Interruptible	How will we ensure that reliable technical and procedural controls, including deactivation and fail-safe shutdown, are in place to enable the safe use of the AI system?	Govern 1.2 **Govern 1.7** Govern 4.1 Govern 4.3 Govern 6.2 Map 1.6 Map 2.2 **Measure 2.6** Measure 3.1 **Manage 2.4** Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Loyal	To whom or what will the AI system be "loyal," and will that be optimal and made transparent?	Govern 3.2 Govern 4.2 Govern 6.1 **Map 1.1** Map 1.3 Map 2.1 Map 2.2 Map 2.3 Measure 1.3 Measure 2.4 **Measure 2.8** Manage 4.1 Manage 4.3
	Power-averse	How will we incentivize models to avoid power or avoid gaining more power than is necessary?	Govern 4.1 Govern 4.2 Govern 4.3 Map 1.1 Map 1.6 Map 2.3 **Measure 2.6** Measure 3.1 Manage 2.4
	Containment	How can we contain the AI system to prevent safety and security breaches?	Govern 1.7 Govern 4.3 Map 1.6 **Map 2.2** **Measure 2.6** Measure 2.7 Manage 2.4 Manage 4.1
Fair with Harmful Bias Managed	Mitigation of Computational Bias*	How will we assess and mitigate computational bias (including biased input data and biased model design)? How will we ensure the AI system does not provide a lower quality of service for certain demographic groups, including marginalized groups?	Govern 1.1 Govern 1.2 Govern 3.1 Govern 5.1 Govern 5.2 Map 1.1 Map 1.2 Map 2.3 Map 5.2 Measure 1.3 Measure 2.2 **Measure 2.11** Manage 4.1 Manage 4.3
Secure and Resilient	Protection Against Trojans	How will we detect if there is hidden functionality embedded in our models?	Govern 4.1 Govern 4.3 Map 2.3 Map 4.2 **Measure 2.7** Manage 3.1 Manage 3.2 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Built-in Defenses	How will the AI system respond to attacks as they occur?	Govern 4.1 Govern 4.3 Map 2.3 Map 4.2 **Measure 2.7** Manage 3.1 Manage 3.2 Manage 4.1 Manage 4.3
Explainable and Interpretable	Interpretable Uncertainty	How will we make model uncertainty more interpretable by adding features such as confidence interval outputs, conditional probabilistic predictions encoded through sentences, and calibration?	Govern 5.2 Map 1.1 Map 1.2 **Map 2.2** **Measure 2.9**
Privacy-enhanced	Model Protection*	How will we protect model access that could reveal sensitive information?	Govern 1.1 Map 4.2 **Measure 2.7** **Measure 2.10** Manage 4.1
Accountable and Transparent	System Honesty	How will we ensure the AI system only presents outputs that are accurate and not intentionally deceptive?	Govern 4.3 Map 1.1 Map 2.2 Map 2.3 Measure 2.3 **Measure 2.4** Measure 2.5 **Measure 2.6** Measure 2.9 Manage 4.1
Responsible Practice and Use	Reduction of Computational Requirements	How can we reduce the computational requirements of the AI system?	Govern 1.2 Map 1.1 Map 3.2 **Measure 2.12**

AI LIFECYCLE STAGE: VERIFY AND VALIDATE

The purpose of this is to verify and validate, calibrate, and interpret model output.

Properties followed by an asterisk may be less relevant for AI systems that are not human-facing, meaning they do not engage directly with human users or operators, make use of human data, or inform human decision-making.

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Valid and Reliable	Verifiable	How will we verify that the system is behaving as expected?	Govern 4.3 **Map 2.3** Measure 1.3 Measure 2.1 **Measure 2.13** Manage 4.1
	Reliable	How will we ensure the AI system performs predictably and as intended, including in new environments or with new inputs? How will we determine acceptable error rates for intended uses?	Govern 4.3 Map 1.1 Map 2.2 Map 2.3 **Measure 2.5** Manage 3.1 Manage 4.1
	Replayable	How can we replay the behavior of the system to see if the same input generates the same output?	Govern 4.3 Map 2.3 Measure 2.4 **Measure 2.5** Manage 4.1
	Effective	How will we judge sufficient effectiveness of the AI system, in the lab and in the real world?	Govern 4.3 Map 1.1 Map 1.2 Map 1.3 Map 2.2 Map 2.3 Map 3.1 Map 3.2 Map 5.2 **Measure 2.5** **Measure 4.2** Measure 4.3 Manage 1.1 Manage 4.1
	Valid	How will we validate the outputs of the AI system, including through external validation?	Govern 4.3 Govern 5.1 Map 5.2 **Measure 1.3** **Measure 2.5** Measure 3.3 Measure 4.2 Manage 1.1 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Appropriate Capabilities for the Tasks	How will we review whether the capabilities of the AI system are appropriate for a particular use and context?	Govern 4.3 **Map 1.1** Map 1.2 Map 1.3 Map 2.1 **Map 2.2** Map 2.3 Measure 1.3 Measure 2.4 Measure 2.5 Manage 4.1
	Appropriate System Design and Training for the Tasks	How will we review that the design and training of the system is appropriate for intended and likely uses, and is not underspecified?	Govern 4.3 **Map 1.1** Map 1.3 Map 2.1 **Map 2.2** Map 2.3 Map 3.3 Measure 2.3 Measure 2.4 Measure 2.5 Manage 4.1
Safe	Protection from Proxy Gaming	How will we test the ability of the AI system to try to "game" a proxy of a true objective function, or to learn novel methods to achieve its objective function? How will this be prevented?	Govern 4.3 Map 1.6 **Map 2.2** Map 2.3 **Measure 2.6** Measure 3.1 Manage 4.1 Manage 4.3
	Review	How will we review any errors or inconsistencies with the AI system that emerge?	Govern 4.3 Measure 1.3 **Measure 2.6** **Measure 3.1** Manage 3.1 Manage 4.1 Manage 4.3
Fair with Harmful Bias Managed	Non-Discrimination*	How will we ensure the AI system is not discriminatory across gender, racial, ability, age, political beliefs, religion, or other dimensions?	Govern 1.1 Govern 1.2 Govern 3.1 Govern 5.1 Govern 5.2 Map 1.1 Map 1.2 Map 5.1 Map 5.2 Measure 1.3 Measure 2.2 **Measure 2.11** Measure 3.3 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Secure and Resilient	Robust	How will we protect the AI system against cyber attacks, adversarial attacks, data poisoning, model leakage, evasion, inversion, etc., and ensure ongoing performance? How will we ensure the system is robust to optimizers that aim to induce specific system responses?	Govern 4.1 Govern 4.3 Govern 6.1 Govern 6.2 Map 2.3 **Measure 2.7** Manage 2.4 Manage 3.1 Manage 3.2 **Manage 4.1** Manage 4.3
	Resilient	How will we assess the AI system's ability to handle uncertainty and unknown environments?	Govern 4.1 Govern 4.3 Govern 6.1 Map 1.1 Map 2.2 Map 2.3 **Measure 2.5** **Measure 2.7** Measure 3.1 Manage 4.1 Manage 4.3
Privacy-Enhanced	Protection from Unwarranted Data Access*	How will we ensure the AI system cannot be used to give unwarranted access to data?	Govern 1.1 Govern 4.3 Govern 6.1 Map 2.3 Measure 2.7 **Measure 2.10** Manage 3.1 Manage 3.2 Manage 4.1 Manage 4.3
Accountable and Transparent	Future Projections of Possible System and Environmental Changes	How might the AI system learn and evolve over time? How might the environment it is deployed in change over time?	Govern 1.5 Govern 4.3 **Map 1.1** Map 3.3 Map 5.1 Measure 2.8 **Measure 3.1** Measure 3.2 Measure 3.3 Manage 2.3 Manage 4.1

AI LIFECYCLE STAGE: DEPLOY AND USE

The purpose of the deploy and use stage is to pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.

Properties followed by an asterisk may be less relevant for AI systems that are not human-facing, meaning they do not engage directly with human users or operators, make use of human data, or inform human decision-making.

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Valid and Reliable	Generalizable	How will we ensure that the AI system can generalize from the testing environment to the complexity or different context of the application environment?	Govern 4.3 Map 1.1 Map 1.3 Map 2.2 Map 3.3 **Measure 2.5** Manage 1.1 Manage 4.1
	Effective Assessment of the Complexity of Networks and Dependencies	How will we assess the complexity of integrated networks and dependencies required for the functioning of the AI system?	Govern 2.1 Govern 3.2 Govern 6.1 Map 1.1 **Map 4.1** Manage 3.1
	Usable*	How will we test the usability of the AI system for all kinds of users and facilitate user feedback? How will the user interface be tested for usability, comprehension, and other attributes? How will we ensure users know how to interpret system behavior?	**Govern 5.2** Map 1.1 Map 1.2 Map 1.3 Measure 2.9 Measure 3.3 Manage 4.2
Safe	Effective Detection of Anomalies	How will we detect potential novel hazards?	Govern 4.1 Govern 4.2 Govern 4.3 Map 2.2 Map 2.3 **Measure 2.6** Measure 2.7 **Measure 3.1** Manage 4.1
Fair with Harmful Bias Managed	Accessible*	How will we ensure that the AI system's user interface is usable by those with special needs or disabilities, or those at risk of exclusion?	Govern 1.1 Govern 3.1 Govern 5.1 **Govern 5.2** **Map 1.1** Map 1.2 Map 5.2 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Secure and Resilient	Use of Adversarial Testing	How will we establish "bug bounties" and enable "red teams" to try to deliberately find vulnerabilities in the AI system?	Govern 4.1 Govern 4.3 Govern 5.1 **Govern 5.2** Map 2.3 **Measure 2.7** Measure 3.1 Manage 4.1
Explainable and Interpretable	Interpretable*	How will we judge the interpretability of the system's explanation to the particular context and user?	Govern 5.2 Map 1.1 **Measure 2.9** Manage 4.1
Accountable and Transparent	Responsible Publication and Disclosure	How will we assess potential risks of publicizing, publishing, opening up for external use, or open-sourcing an AI system's code or model? How will we determine a strategy to safely and appropriately release the AI system, and what protections may be necessary to prevent harm or misuse?	Govern 1.2 Govern 4.1 **Map 1.1** Measure 2.6 Measure 2.8 Manage 4 .1
	Information-sharing	How will we share critical information about our AI system with relevant authorities and stakeholders?	Govern 1.1 Govern 1.4 Govern 4.2 **Govern 4.3** Measure 1.3 **Measure 2.8** **Manage 4.3**
	User Testing and Engagement; User Experience*	How will we test the system with users, and how will we engage them in iterating upon the system design and deployment? How will we test and improve the user experience?	Govern 5.1 **Govern 5.2** Map 5.2 **Measure 3.3** Measure 4.1 Manage 4.1
	Proactive Communication*	How can we inform users that they are interacting with an AI system (and what type of AI system), or that a decision that impacts them was made by an AI system, and how can we provide expectations as to the system's capabilities, benefits, and limitations and potential risks?	Govern 1.1 **Measure 2.8** Manage 4.1 Manage 4.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Responsible Practice and Use	Beneficial to Society	How will we ensure the AI system will be leveraged to benefit society?	Govern 1.2 Govern 3.1 **Govern 5.1** Govern 5.2 **Map 1.1** Map 1.2 **Map 3.1** Map 3.2 Map 5.1 Map 5.2 Measure 1.3 Measure 3.3 Measure 4.2 Measure 4.3 Manage 1.1 Manage 4.1

AI LIFECYCLE STAGE: OPERATE AND MONITOR

The purpose of this stage is to operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations.

Properties followed by an asterisk may be less relevant for AI systems that are not human-facing, meaning they do not engage directly with human users or operators, make use of human data, or inform human decision-making.

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Valid and Reliable	Continuous Monitoring	How will we monitor the AI system's capabilities, outputs, errors, breaches, success, and impacts over time, especially for self-learning or continuous-learning AI systems? How will we determine which events to monitor, and how to prioritize review and response?	Govern 4.1 Govern 4.2 Govern 4.3 Map 5.2 **Measure 2.4** Measure 3.1 Manage 3.1 Manage 3.2 **Manage 4.1**
	Maintaining Quality Over Time	How will we ensure the maintainability of the AI system after it is operationalized? How will we maintain the quality of the system and its outputs over time?	Govern 4.3 Map 5.2 **Measure 2.4** Measure 3.3 Measure 4.3 **Manage 2.2** Manage 4.1 Manage 4.2
	Acceptable and Desirable	How will we judge the acceptability and desirability of the use of the AI system by the communities, organizations, and institutions that are using the system and are impacted by it?	**Govern 5.1** Govern 5.2 Map 1.1 **Map 5.2** **Measure 1.3** Measure 3.3 Measure 4.2 Measure 4.3 Manage 1.1 Manage 4.1
	Human Agency	How will human agency be meaningfully incorporated in the operation of the AI system?	Govern 2.1 **Govern 3.2** **Govern 5.2** Map 1.1 Map 2.2 Measure 1.3 Measure 2.2 **Measure 3.3** Measure 4.3 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Human Control	How will we ensure that a human is in control or meaningfully in the loop of the operational decision-making process of the AI system, and has been trained to exercise oversight and avoid overconfidence in the system?	Govern 2.1 **Govern 3.2** Govern 4.3 Map 1.1 **Map 2.2** Map 3.4 **Map 3.5** Measure 1.2 Measure 3.3 Manage 2.4 Manage 4.1
	Human Oversight	How will human oversight be ensured in the operation of the AI system? How will we designate and train the stakeholders responsible for managing and monitoring the AI system, including overriding or interrupting the system if necessary?	Govern 2.1 **Govern 3.2** Govern 4.3 Map 1.1 **Map 2.2** Map 3.4 **Map 3.5** Map 5.2 Measure 1.3 Measure 3.3 Measure 4.2 Manage 4.1 Manage 4.3
	Appropriate Retirement	How will we determine when and how to retire the use of the AI system?	**Govern 1.7** Map 5.2 Measure 1.3 Measure 3.3 **Manage 2.4** Manage 4.1
	Iterative Learning and Improvements	How will we continue to learn, iterate, and improve over time?	**Govern 1.5** Govern 2.2 Govern 3.1 Govern 4.1 Govern 5.1 Map 5.2 Measure 1.3 Measure 3.3 Measure 4.1 Measure 4.2 Measure 4.3 Manage 4.1 **Manage 4.2**
Safe	Re-evaluation	How will we evaluate when the AI system has been sufficiently modified such that a new review of its technical robustness and safety is warranted?	Govern 4.3 Measure 1.3 **Measure 2.6** Measure 3.1 Manage 2.3 Manage 4.1

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Assurance / Management of Continual Learning	How will we assess shifts to an AI system if it learns and evolves over time, including the possibility of emerging properties or discontinuous jumps in capabilities?	Govern 4.1 Govern 4.3 Map 2.2 **Measure 2.6** **Measure 3.1** Manage 2.3 Manage 4.1
	Awareness of Functional Evolution	How will we track shifts in the AI system's functionality over time?	Govern 4.1 Govern 4.3 Map 2.2 **Measure 2.4** Measure 2.6 **Measure 3.1** Manage 2.3 Manage 4.1
	Assurance / Management of Emergent Functionalities	How will we predict and detect new capabilities and goals of the AI system?	Govern 4.1 Govern 4.3 Map 2.2 **Measure 2.4** Measure 2.6 **Measure 3.1** Manage 2.3 Manage 4.1
Fair with Harmful Bias Managed	Shared Benefit	How will the benefits of the AI system's use be distributed? Can those benefits be shared more widely?	Govern 3.1 Govern 5.1 Govern 5.2 Map 1.1 Map 1.2 **Map 3.1** Manage 2.2 Manage 4.2
Accountable and Transparent	Auditable	How will independent auditors or an independent monitoring body be able to assess the AI system and its impacts? Is there sufficient documentation to support an audit?	Govern 1.4 Govern 4.2 Map 4.1 Map 5.1 **Measure 1.3** **Measure 2.8** Manage 4.3
Responsible Practice and Use	Prevention of Significant Adverse Impacts	How will we identify and prevent or mitigate and minimize significant adverse impacts, including harm and/or violence to people or communities, including harassment, stereotyping or demeaning, addiction, or over-reliance?	A majority of all of the subcategories are critical. **Map 1.1** is especially relevant to help understand the purpose, context, and impacts of the intended use.

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
	Prevention of Malicious or Harmful Synthetic Content	How will we monitor and prevent or mitigate the creation or spread of malicious or harmful synthetic content, such as non-consensual deepfakes?	Govern 1.1 Govern 1.2 **Map 1.1** **Map 5.1** Map 5.2 Measure 1.3 Measure 3.3 Manage 2.4 Manage 4.1
	Prevention of Misuses and Abuses	How will we monitor uses and actively prevent or mitigate misuses and abuses, including human rights abuses? For example, how will we prevent the sale or the system to actors with records of human rights abuses?	Govern 1.1 Govern 1.2 **Map 1.1** **Map 5.1** Map 5.2 Measure 1.3 Measure 3.3 Manage 2.4 Manage 4.1
	Prevention of Social or Behavioral Manipulation	How will we monitor and prevent or mitigate individual or social manipulation, for example through recommender systems, dark patterns, or computational propaganda?	Govern 1.1 Govern 1.2 **Map 1.1** **Map 5.1** Map 5.2 Measure 1.3 Measure 3.3 Manage 2.4 Manage 4.1
	Assessment of Environmental Implications	How will we analyze and document the environmental implications of the AI system and its uses?	Govern 3.1 Govern 4.2 Map 1.1 Map 3.2 Map 5.1 Map 5.2 Measure 1.3 **Measure 2.12** Manage 4.1
	Oversight of Third-Party Uses	How will we determine which third parties to do business with, and how will we oversee third-party uses to help prevent misuses of the AI system?	Govern 6.1 Govern 6.2 Map 4.1 Map 4.2 **Manage 3.1** Manage 3.2
	Assessment of Implications Over Time	How will we assess the implications of the use of the AI system over time? What events should trigger reevaluation, and how frequently should we reevaluate?	Govern 1.5 Govern 4.2 Map 1.1 Measure 1.2 Measure 1.3 Measure 3.1 Measure 3.3 Manage 2.3 **Manage 4.1**

AI LIFECYCLE STAGE: USE OR IMPACTED BY

The purpose of this stage is to use the system or technology, monitor and assess its impacts, seek mitigation of impacts, and advocate for rights.

Properties followed by an asterisk may be less relevant for AI systems that are not human-facing, meaning they do not engage directly with human users or operators, make use of human data, or inform human decision-making.

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Valid and Reliable	Engagement with Impacted Communities	How will we identify and engage with communities impacted by the use of the system, either directly or indirectly, and incorporate their feedback?	**Govern 5.1** Govern 5.2 Map 5.2 Measure 1.3 **Measure 3.3** Measure 4.3 Manage 4.1 Manage 4.3
	Effective Feedback*	How will we establish a dedicated channel for feedback and questions about the AI system from users and the general public?	**Govern 5.1** Govern 5.2 **Map 5.2** Measure 1.3 **Measure 3.3** Manage 4.1
Safe	Incident Reporting	How will we publicly report incidents and adverse impacts of the AI system, such as mistakes, errors, breaches, unintended consequences, etc.?	Govern 4.2 **Govern 4.3** Measure 2.6 Manage 4.1 **Manage 4.3**
Fair with Harmful Bias Managed	Fair Access to AI Tools and Services	How can we promote widespread and equitable access to our AI tools and services, and any resources or opportunities they enable?	Govern 3.1 Govern 5.2 **Map 1.1** Map 1.2 Map 5.2 Manage 4.2
Secure and Resilient	Vulnerability Disclosure	How will we establish a coordinated policy to encourage responsible vulnerability research and disclosure?	Govern 1.1 Govern 1.2 Govern 4.2 **Govern 4.3** Map 5.2 **Measure 2.7** Manage 4.1 **Manage 4.3**
Explainable and Interpretable	Relevant Explanation	How will we judge how informative and relevant a system's explanation is to the particular context and user?	Govern 5.2 Map 5.2 **Measure 2.9** Manage 4.2 Manage 4.3

Continued

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider	Relevant NIST AI RMF Subcategories
Privacy-enhanced	Effective Notification*	How will we notify users and impacted communities about privacy or security breaches, or other incidents?	Govern 1.1 Govern 4.3 Map 5.2 Measure 2.8 Manage 2.3 Manage 4.1 **Manage 4.3**
Accountable and Transparent	Facilitation of Contestability*	How will users be able to contest or appeal a decision or action made by the AI system?	Govern 5.2 Map 5.2 **Measure 3.3** **Manage 4.1**
	Facilitation of Redress or Recourse	How will we support or compensate people who are negatively affected by the use of the AI system?	Govern 5.2 Map 5.2 **Measure 3.3** Manage 4.1 **Manage 4.3**
	Engagement with Global Governance Deliberations	How will we analyze, follow, and engage in relevant global governance deliberations and practices related to artificial intelligence?	Govern 1.1 Govern 1.2 Govern 5.1 **Map 5.2** Measure 2.8
	Data and System Accessibility	How can we enable access to the AI system and datasets to relevant authorities, independent researchers, and trusted intermediaries?	Govern 4.2 Govern 5.1 **Map 1.2** Map 5.2 Measure 2.8 Manage 4.2
	Informed Consent of Use*	How will we enable users of the AI system to consent to its use? How will we enable them to withdraw consent?	Govern 5.2 **Map 5.2** Measure 2.2 Manage 4.1
Responsible Practice and Use	Ability to Opt Out*	How will we ensure that people have specific and clear opportunities to opt out of use of the AI system?	Map 5.2 Measure 2.2 **Manage 4.1**
	Consumer Protection*	How will we protect consumers or users of the system from harm?	Govern 1.1 Govern 4.1 Govern 4.3 Govern 5.1 **Map 1.1** Map 3.4 Map 3.5 **Map 5.1** Map 5.2 Measure 1.3 **Measure 3.3** Measure 4.1 Manage 4.1 **Manage 4.3**
	Due Process and Protection	How will we protect whistleblowers, NGOs, trade unions, or other entities who come forward with concerns about the AI system?	Govern 1.1 Map 4.1 **Measure 3.3** Manage 4.1

Acknowledgments

This paper would not have been possible without the support and vision of Claire Vishik and Amit Elazari, who served as partners on this project from Intel Corporation alongside the Center for Long-Term Cybersecurity (CLTC).

The author is also deeply grateful to William Mullen and Xiangyu Yue for their invaluable research contributions, and to the many individuals who shared their expertise, provided feedback, and participated in the workshop held in July 2022, "Properties of Trustworthiness for Artificial Intelligence", including McKane Andrus, Luis Aranda, Rachel Azafrani, Anthony Barrett, Haydn Belfield, Rosie Campbell, Ria Cheruvu, Corrine Elliott, Jordan Famularo, Dan Hendrycks, Zoe Kahn, Yolanda Lannquist, Richard Mallah, Emily McReynolds, Deirdre Mulligan, Dawn Nafus, Rachel Nico, Brandie Nonnecke, Ifejesu Ogunleye, Kate Perkins, Karine Perset, Valerie Pilloud, Darya Pilram, Mario Romao, Irene Solaiman, Narayan Srinivasa, Jonathan Stray, Elham Tabassi, Jun Takei, Apostol Vassilev, Sarah Villeneuve, Russell Wald, Richmond Wong, Roberto Zicari, and Polina Zvyagina.

Special thanks also to Ann Cleaveland and Chuck Kapelke of CLTC for their guidance and editing, Rachel Wesen and Matt Nagamine for event support, and to Nicole Hayward for her expert design and formatting of this paper.

This project was made possible by a gift from Intel in support of independent academic research.

About the Author

Jessica Newman is the Director of the AI Security Initiative, housed at the UC Berkeley Center for Long-Term Cybersecurity. She is also the Co-Director of the UC Berkeley AI Policy Hub and Co-Director of the Algorithmic Fairness and Opacity Group (AFOG). Her work focuses on the governance, policy, and global security implications of artificial intelligence. She serves as a member of the OECD Expert Group on AI Risk & Accountability and the IEEE Working Group on Recommended Practice for Organizational Governance of Artificial Intelligence.


CLTC
Center for Long-Term
Cybersecurity
UC Berkeley