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

JESSICA NEWMAN

JANUARY 2023

CLTC
Center for Long-Term
Cybersecurity
UC Berkeley

Contents

Introduction

Improving the trustworthiness of artificial intelligence (AI) systems is a shared priority for the private and public sectors, as indicated by prolific research and guidelines in recent years.[1,2,3] However, the meaning of trustworthy AI, and potential approaches to attain it, remain contested.[4,5,6] Stakeholders often lack a shared vocabulary or set of questions to consider, and guidance that speaks both to AI developers and policymakers is rare. Existing frameworks for trustworthy AI have additionally tended to focus on a relatively narrow set of AI models and applications that directly interact with people.[7] Since 2021, the National Institute of Standards and Technology (NIST), in collaboration with diverse stakeholders, has been developing an AI Risk Management Framework (RMF) intended to promote trustworthy AI. In this paper, we analyze the landscape of trustworthy AI and introduce a taxonomy of trustworthiness for artificial intelligence that is intended to complement and support the use of the NIST AI RMF.

The taxonomy introduced in this paper 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. These characteristics of trustworthiness mirror well-established international AI principles, and because the properties included in the taxonomy provide greater nuance about implementing each characteristic, the paper builds upon a body of

[1] "Ethics Guidelines for Trustworthy AI," April 8, 2019, High-Level Expert Group on Artificial Intelligence, European Commission, https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[2] "Advancing Trustworthy AI," National Artificial Intelligence Initiative Office, United States White House, https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/.

[3] Haochen Liu et al., "Trustworthy AI: A Computational Perspective," arXiv, July 12, 2021, https://arxiv.org/abs/2107.06641.

[4] Michele Loi et al., "'Trustworthy AI' is not an appropriate framework," AlgorithmWatch, February 6, 2019, https://algorithmwatch.org/en/trustworthy-ai-is-not-an-appropriate-framework/.

[5] Gernot Rieder, Judith Simon, and Pak-Hang Wong, "Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums," October 23, 2020, In: Marcello Pelillo and Teresa Scantamburlo (Eds.). *Machines We Trust: Perspectives on Dependable AI*, Cambridge, MA: MIT Press, Forthcoming, https://ssrn.com/abstract=3717451.

[6] Matthias Braun, Hannah Bleher, and Patrik Hummel, "A Leap of Faith: Is There a Formula for "Trustworthy" AI?," The Hastings Center Report, Volume 51, Issue 3, Pages 17–22, February 19, 2021, https://onlinelibrary.wiley.com/doi/full/10.1002/hast.1207.

[7] "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," July 17, 2020, High-Level Expert Group on Artificial Intelligence, European Commission, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

literature related to making AI principles more concrete and actionable.[8,9,10,11] We sourced the properties from dozens of papers, policy documents, and extensive interviews and feedback.[12] We also held an expert workshop in July 2022 that further informed the properties and framework.[13]

The taxonomy of trustworthiness is organized by the seven stages of the AI lifecycle depicted in the NIST AI RMF. Our hope is that this organization 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. However, we realize that this approach has limitations. Importantly, trustworthiness is not only tied to a product lifecycle. Indeed, many of the properties of trustworthiness are related to the people and organization associated with an AI technology, rather than just the product. Moreover, each listed property is unlikely to only be important during one particular time, and may need to be revisited at regular intervals throughout the AI lifecycle. Nonetheless, we believe that the properties are likely to have a greater relative importance at particular stages of the AI lifecycle, and that there may be unique windows of opportunity in which to address them.

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

[8] Jessica Morley et al., "What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," Sci Eng Ethics 26, 2141–2168, 2020. https://doi.org/10.1007/s11948-019-00165-5.

[9] Miles Brundage et al. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," arXiv, April 20, 2020, https://arxiv.org/pdf/2004.07213.pdf.

[10] "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems," OECD Digital Economy Papers, No. 312, OECD Publishing, Paris, 2021, https://doi.org/10.1787/008232ec-en.

[11] Ben Shneiderman, "Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems," ACM Transactions on Interactive Intelligent Systems, Volume 10, Issue 4, December 2020, https://dl.acm.org/doi/abs/10.1145/3419764.

[12] For example, see: "Blueprint For an AI Bill of Rights," The White House Office of Science and Technology Policy, whitehouse.gov/ostp/ai-bill-of-rights; "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," July 17, 2020, High-Level Expert Group on Artificial Intelligence, European Commission, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment; "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities," U.S. Government Accountability Office, June 2021, https://www.gao.gov/assets/gao-21-519sp.pdf; "Recommendation on the Ethics of Artificial Intelligence," United Nations Educational, Scientific and Cultural Organization, 2022, https://unesdoc.unesco.org/ark:/48223/pf0000381137; Jessica Fjeld et al. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," January 15, 2020, Berkman Klein Center Research Publication No. 2020-1, https://ssrn.com/abstract=3518482.

[13] The virtual workshop, "Properties of Trustworthiness for Artificial Intelligence," was held in July 2022 and was co-organized by the UC Berkeley Center for Long-Term Cybersecurity and Intel. Participants included more than 40 experts from academia, government, industry, and civil society. The names of some of the workshop participants are included in the acknowledgments.

of trustworthiness that are relevant at that stage of the lifecycle. We also added an eighth characteristic of trustworthiness to the taxonomy, which varies slightly from the NIST AI RMF. This 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. Additionally, each property is accompanied by a list of the most relevant sections of the NIST AI RMF core. This serves to provide guidance to the reader about where to go to find additional information and resources about the property.

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. The paper includes further discussion of the spectrum of human-AI engagement and how this relates to trustworthiness. A subset of the properties of trustworthiness in the taxonomy are likely to only be relevant to AI systems that are human-facing. Human-facing AI systems 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 paper aims to provide a resource that is useful for AI organizations and teams developing AI technologies, systems, and applications. It is designed to specifically assist users of the NIST AI RMF, however 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.

This work highlights a number of key findings:

1. **Many stakeholders have a role to play in developing and ensuring trustworthy AI.** Fully considering the trustworthiness of an AI system requires diverse and multidisciplinary expertise. The process should include a broad range of roles from within an organization as well as outside experts, including members of impacted communities and independent verification and auditing bodies.
2. **The consideration of trustworthiness should not wait until after an AI system has been developed.** Many properties of trustworthiness are most critical in the early design phase.
3. **Many properties of trustworthiness are relevant regardless of whether an AI system is "high risk."** For example, properties related to safety, quality, and sustainability tend to matter regardless of application area. This means that it is critical to consider trustworthiness even for AI applications that do not qualify as "high risk," and that frameworks for trustworthy AI that primarily focus on high-risk applications may not be sufficient.
4. **Some properties of trustworthiness are less relevant for AI applications that are not human-facing.** For example, some properties of trustworthiness relate to interactions with users, but not all AI systems call for interactions with users.
5. **Striving for trustworthy AI is a complex and ongoing process, not an easily achievable outcome.** Organizations should be wary of applying easy-fix solutions to complex technical and social problems. There are numerous properties of trustworthiness, some of which are active areas of research that may not yet have obvious and available solutions. Building trustworthy AI systems should be seen as an ongoing process to earn trust, rather than an easily achievable outcome.

Trustworthy AI

WHAT IS TRUSTWORTHY AI?

The notion of "trustworthy AI" builds upon longer histories of trust in computing, cyberspace, and automation.[14,15,16] The NIST Framework for Cyber-Physical Systems describes trustworthiness in the following way:

> Trustworthiness is the demonstrable likelihood that the system performs according to designed behavior under any set of conditions as evidenced by characteristics including, but not limited to, safety, security, privacy, reliability and resilience. In computer security, a chain of trust is established by validating each component of hardware and software from the bottom up.[17]

The Oxford English Dictionary defines "trustworthy" as "worthy of trust or confidence; reliable, dependable." People typically only place trust in those who have repeatedly and exclusively proven these characteristics. A single failure to live up to these expected characteristics can break trust between two people. Taken as a whole, AI technologies often fail to live up to our expectations. They can be inaccurate, unreliable, and discriminatory.[18,19]

A large body of literature specifically on "trustworthy AI" has emerged, and the phrase is now commonly used in multistakeholder forums that span government, industry, academia, and

[14] National Research Council. *Trust in Cyberspace*. The National Academies Press, 1999; https://doi.org/10.17226/6161.

[15] Bonnie M. Muir. "Trust in Automation: Part 1. Theoretical Issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994, https://www.tandfonline.com/doi/abs/10.1080/00140139408964957.

[16] Kevin Anthony Hoff and Masooda Bashir. "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2006.

[17] "Framework for Cyber-Physical Systems: Volume 2, Working Group Reports," NIST Special Publication 1500-202, National Institute of Standards and Technology, June 2017, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-202.pdf.

[18] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Proceedings of Machine Learning Research 81:1–15, 2018, Conference on Fairness, Accountability, and Transparency, proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[19] Khari Johnson, "How Wrongful Arrests Based on AI Derailed 3 Men's Lives," *Wired*, March 7, 2022, https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/.

civil society.[20,21,22] For example, the OECD has stated, “Trustworthy AI refers to AI systems that embody the OECD AI Principles; that is, AI systems that respect human rights and privacy; are fair, transparent, explainable, robust, secure and safe; and the actors involved in their development and use remain accountable.[23] However, the definition of “trustworthy AI” is contested.[24,25] For example, some argue the term is too vague to be helpful and that it is typically not clear who is being asked to place trust in whom or what.[26] There are also different cultural and political understandings of trustworthiness around the world. The prominence of western democratic nations in high-profile deliberations of trustworthy AI may fail to account for global differences, and could entrench unacknowledged and unquestioned systems of values and power.[27]

We recognize the imperfections of the term “trustworthy AI,” and use it in this paper to refer to an aspiration and an ongoing process, rather than a precise descriptive quality or easily attainable goal. Importantly, the “trust” in question is not only to be placed in a given technical system, but also with the actors and processes that develop, deploy, and monitor that system. Meaningful “trustworthiness” may be unwarranted or unattainable. Self-assessment of trustworthiness is certainly flawed, and independent auditing,[28] technical standards,[29] and federal regulation[30] will all be critical. Nonetheless, countless industries and domains have already adopted AI technologies in their operations, in some cases reaching millions of people

[20] “Ethics Guidelines for Trustworthy AI,” April 8, 2019, High-Level Expert Group on Artificial Intelligence, European Commission, https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[21] “Advancing Trustworthy AI,” National Artificial Intelligence Initiative Office, United States White House, https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/.

[22] Haochen Liu et al., “Trustworthy AI: A Computational Perspective,” arXiv, July 12, 2021, https://arxiv.org/abs/2107.06641.

[23] “Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems,” OECD Digital Economy Papers, No. 312, OECD Publishing, 2021, https://doi.org/10.1787/008232ec-en.

[24] Gernot Rieder, Judith Simon, and Pak-Hang Wong, “Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums,” October 23, 2020, In: Marcello Pelillo and Teresa Scantamburlo (Eds.). *Machines We Trust: Perspectives on Dependable AI*, Cambridge, MA: MIT Press, Forthcoming, https://ssrn.com/abstract=3717451.

[25] Matthias Braun, Hannah Bleher, and Patrik Hummel. “A Leap of Faith: Is There a Formula for ‘Trustworthy’ AI?,” The Hastings Center Report, Volume 51, Issue 3, Pages 17–22, February 19, 2021, https://onlinelibrary.wiley.com/doi/full/10.1002/hast.1207.

[26] Michele Loi et al. “‘Trustworthy AI’ is not an appropriate framework,” AlgorithmWatch, February 6, 2019, https://algorithmwatch.org/en/trustworthy-ai-is-not-an-appropriate-framework/.

[27] Shakir Mohamed, Marie-Therese Png, and William Isaac, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence,” Philosophy and Technology (405), July 2020, https://arxiv.org/pdf/2007.04068.pdf.

[28] Inioluwa Deborah Raji et al., “Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance,” arXiv, June 9, 2022, https://arxiv.org/abs/2206.04737.

[29] Peter Cihon, “Standards for AI Governance: International Standards to Enable Global Coordination in AI Research and Development,” Future of Humanity Institute, University of Oxford, April 2019, https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

[30] Lindsey Barrett, “Ban Facial Recognition Technologies for Children—And for Everyone Else,” *Boston University Journal of Science and Technology Law*, Volume 26.2, July 24, 2020, https://ssrn.com/abstract=3660118.

every day.[31] Having a greater number of actors working toward the development of high-quality and responsible AI systems may help reduce the risks that AI systems pose, which are often disproportionately experienced by marginalized or vulnerable communities.[32]

EXISTING FRAMEWORKS FOR TRUSTWORTHY AI

Many trustworthy AI frameworks and documents have helped inform this work. Some notable examples include ongoing standards efforts, such as those of the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE), including the IEEE 7000™-2021 Standard on Addressing Ethical Concerns During Systems Design;[33] the IEEE CertifAIEd™ Program, a risk-based framework supported by a suite of AI ethical criteria to support a trustworthy experience for users;[34] and the "Overview of trustworthiness in artificial intelligence," published by ISO Technical Committee: ISO/IEC JTC 1/SC 42 on artificial intelligence.

Other works that have helped inform this effort include the OECD AI Recommendation;[35] the UNESCO Recommendation on the Ethics of Artificial Intelligence;[36] the Responsible AI Certification;[37] the "Trustworthy AI" white paper by the China Academy for Information and Communication Technology (CAICT);[38] the Principled Artificial Intelligence project at the Berkman Klein Center;[39] Z-Inspection®;[40] the Indigenous Protocol and Artificial Intelligence

[31] Prabhakar Raghavan, "How AI is powering a more helpful Google," The Keyword, Google, October 15, 2020, https://blog.google/products/search/search-on/.

[32] Meredith Whittaker et al., "Disability, Bias, and AI," AI Now Institute, November 2019, ainowinstitute.org/disabilitybiasai-2019.pdf.

[33] "IEEE 7000™-2021 Standard: Addressing Ethical Concerns During Systems Design," IEEE Standards Association, September 2021, https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html.

[34] "IEEE CertifAIEd™ The Mark of AI Ethics," IEEE Standards Association, https://engagestandards.ieee.org/ieeecertifaied.html.

[35] OECD Recommendation of the Council on Artificial Intelligence, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[36] "Recommendation on the Ethics of Artificial Intelligence," United Nations Educational, Scientific and Cultural Organization, 2022, https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[37] "Responsible AI Certification," The Responsible Artificial Intelligence Institute, https://www.responsible.ai/how-we-help.

[38] Matt Sheehan, "Beijing's Approach to Trustworthy AI Isn't So Dissimilar from the World's," MACRO POLO, August 18, 2021, https://macropolo.org/beijing-approach-trustworthy-ai/.

[39] Jessica Fjeld et al. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," January 15, 2020, Berkman Klein Center Research Publication No. 2020-1, https://ssrn.com/abstract=3518482.

[40] Roberto V. Zicari et al. "Z-Inspection®: A Process to Assess Trustworthy AI," IEEE Transactions on Technology and Society, VoL. 2, No. 2, June 2021, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=andarnumber=9380498. Roberto V. Zicari et al., "How to Assess Trustworthy AI in Practice," arxiv, June 20, 2022, https://arxiv.org/pdf/2206.09887.pdf.

Position Paper;[41] the Predictability, Computability, and Stability (PCS) Framework;[42] the Framework for Supply Chain Trust in Hardware and Software;[43] the Digital Catapult AI Ethics Framework;[44] and the Trustworthy AI process developed by Deloitte.[45] While these works vary in scope and goals, they all articulate visions of what may be required to achieve trustworthy AI.

While all of these efforts are important, below we highlight four high-profile frameworks for trustworthy AI led by or affiliated with government actors. The first was led by the High-Level Expert Group on Artificial Intelligence, set up by the European Commission; the second is the European Union Artificial Intelligence Act (EU AI Act); the third is the White House Blueprint for an AI Bill of Rights; and the fourth is the NIST AI RMF, which is most explicitly linked to the taxonomy introduced in this paper. These four efforts are highly interrelated and influential within the US in particular. The EU AI Act, which was informed by the work of the High-Level Expert Group on Artificial Intelligence, is establishing legal requirements that many American companies will need to meet, and could influence other countries' legal approaches to AI.[46] The AI Bill of Rights and the NIST AI RMF are both voluntary, but are likely to be adopted by organizations and companies across the United States and beyond. We expect this work to be complementary to these four efforts, and have explicitly designed the taxonomy to be usable alongside the NIST AI RMF.

The High-Level Expert Group on Artificial Intelligence Assessment List for Trustworthy AI

In 2019, the High-Level Expert Group on Artificial Intelligence, an independent body set up by the European Commission, published "Ethics Guidelines for Trustworthy AI," a foundational

[41] Jason Edward Lewis et al., "Indigenous Protocol and Artificial Intelligence Position Paper," Indigenous Protocol and Artificial Intelligence Working Group and the Canadian Institute for Advanced Research, 2020, https://www.indigenous-ai.net/position-paper.

[42] Bin Yu and Karl Kumbier, "Veridical data science," PNAS Vol. 117, No. 8, February 13, 2020, https://www.pnas.org/doi/10.1073/pnas.1901326117.

[43] Paul Rosenzweig et al., "Creating a Framework for Supply Chain Trust in Hardware and Software," Lawfare Institute's Trusted Hardware and Software Working Group, May 2022, https://www.documentcloud.org/documents/21831749-creating-a-framework-for-supply-chain-trust-in-hardware-and-software.

[44] "Ethics Framework," Digital Catapult's Ethics Committee, https://migarage.digicatapult.org.uk/ethics/ethics-framework/.

[45] Trustworthy AI™, Deloitte, 2020, https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html.

[46] Charlotte Siegmann and Markus Anderljung, "The Brussels Effect and Artificial Intelligence," Centre for the Governance of AI, August 2022, https://www.governance.ai/research-paper/brussels-effect-ai.

work in articulating the need for trustworthy AI and how stakeholders could try to work toward it. The aim of the guidelines was to promote trustworthy AI. The authors wrote:

> In a context of rapid technological change, we believe it is essential that trust remains the bedrock of societies, communities, economies and sustainable development. We therefore identify Trustworthy AI as our foundational ambition, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place.

The Expert Group defined trustworthy AI as having three components that should be met throughout the system's lifecycle:

1. Trustworthy AI should be **lawful**, complying with all applicable laws and regulations;
2. **ethical**, ensuring adherence to ethical principles and values; and
3. **robust**, both from a technical and social perspective, as even with good intentions, AI systems can cause unintentional harm.[47,48]

The approach is grounded in fundamental rights, including international human rights law and a framework of democracy and the rule of law. The associated framework for achieving trustworthy AI focuses on the second and third components; it identifies and describes the ethical principles required for ethical and robust AI, translates these ethical principles into seven requirements for an AI system to meet throughout the lifecycle, and offers an assessment list to operationalize the requirements. The High-Level Expert Group stressed that the seven requirements outlined in Chapter II of the "Ethics Guidelines for Trustworthy AI" should be continuously evaluated and addressed throughout the AI system's lifecycle.

The third chapter of the framework, the Trustworthy AI Assessment List, is intended to be relevant for AI systems that directly interact with users and is primarily addressed to developers and deployers of AI systems. The original paper included a pilot version of the Assessment List. The final Assessment List for Trustworthy AI (ALTAI) was developed over the next two years, following a piloting phase and with significant stakeholder feedback. It is intended for self-assessment and flexible use, and is designed to be completed by a

[47] "Ethics Guidelines for Trustworthy AI," April 8, 2019, High-Level Expert Group on Artificial Intelligence, European Commission, https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[48] "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," July 17, 2020, High-Level Expert Group on Artificial Intelligence, European Commission, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

multidisciplinary team. It is recommended that a fundamental rights impact assessment is completed prior to completing the Assessment List.

The ALTAI is a rich and useful resource. Each of the seven requirements has between two and four sub-sections, with several questions designed to test whether an AI system might meet that requirement. A first question probes at a particular issue while a second question asks if steps have been taken to manage it. For example:

- Could the AI system generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI system?
    - Are end-users or subjects informed that they are interacting with an AI system?

All questions in the ALTAI are important and continue to be highly relevant for teams developing, procuring, or deploying AI systems. However, as mentioned above, the ALTAI is intended to be primarily relevant for AI systems that directly interact with users, and many AI systems do not directly interact with users. The taxonomy of trustworthiness provided in this paper draws inspiration from the important and foundational work of the ALTAI, but is designed for use with a broader set of AI systems, with varying degrees of human interaction and engagement. The taxonomy in this paper also provides several novel details, including the most relevant stages of the AI lifecycle for each property, and can be used to supplement the ALTAI if desired.

The European Union Artificial Intelligence Act

The European Union Artificial Intelligence Act (EU AI Act) builds upon the notion of trustworthy AI described by the High-Level Expert Group on Artificial Intelligence and establishes a “legal framework for trustworthy AI.” The AI Act is a regulatory proposal from the European Parliament and the European Council to develop harmonized rules on AI, and is one of the world’s first overarching regulations for AI technologies. The primary objective of the EU AI Act is to facilitate the development and use of trustworthy AI in the European Union. The Act’s Explanatory Memorandum explains, “It supports the objective of the Union being a global leader in the development of secure, trustworthy and ethical artificial intelligence as stated by the European Council and ensures the protection of ethical principles as specifically requested by the European Parliament.”

The EU AI Act establishes a set of horizontal mandatory requirements for trustworthy AI, which include a prohibition against a small number of AI uses that create “unacceptable risk,” including AI uses that violate fundamental rights; that have a significant potential to manipulate people through subliminal techniques or exploit vulnerabilities of specific vulnerable groups such as children in a way that is likely to cause psychological or physical harm; that enable social scoring by public authorities; or that enable remote biometric identification for law enforcement in public spaces.

The EU AI Act also contains requirements for “high-risk” AI systems, including data and data governance, documentation and record keeping, transparency and provision of information to users, human oversight, robustness, accuracy, and security. However, the Act allows flexibility in managing these and does not provide precise technical solutions to achieve compliance with the requirements.

The classification of high-risk AI systems is based on the intended purpose of the AI system, and only eight pre-listed areas are considered high-risk. These areas include: biometric identification and categorization of natural persons; management and operation of critical infrastructure; education and vocational training; employment, workers management, and access to self-employment; access to and enjoyment of essential private services and public services and benefits; law enforcement; migration, asylum, and border control management; and administration of justice and democratic processes. High-risk AI systems are required to undergo an *ex ante* conformity assessment.

Certain AI systems that pose limited risks are subject to transparency requirements. For example, AI systems that interact with humans, that are used to detect emotions or determine association with social categories based on biometric data, or that generate or manipulate content are all subject to transparency obligations. For example, people will need to be informed if they are interacting with an AI system.

All other AI systems are not subject to requirements under the AI Act. These AI systems are seen as posing “minimal risk” and are only subject to voluntary codes of conduct. For example, commitments related to environmental sustainability, accessibility for people with disabilities, stakeholders’ participation in the design and development of AI systems, and diversity of development teams are all considered voluntary.

The EU AI Act establishes the world’s first legal framework for trustworthy AI and will surely contribute to greater accountability for high-risk AI technologies. However, the requirements it establishes do not apply to most AI systems, since most will be seen as posing “minimal risk.” The EU AI Act may thus leave a gap for ensuring the trustworthiness of AI applications that do not fall into a relatively small number of predetermined high-risk areas. Notably, almost all of the AI systems that are subject to requirements under the AI Act are likely to be those that have direct or significant interaction with people. The EU AI Act will not necessarily raise the bar of quality and governance for most AI systems unless there is significant work done to develop and facilitate the use of voluntary codes of conduct.

The EU AI Act has a unique approach to trustworthiness because it primarily focuses on the intended purpose of an AI system, and only applies mandatory requirements for pre-defined domains of use. This approach has gaps because AI systems may also be misused, abused, or simply applied in novel areas other than those originally imagined by the developers. Moreover, there are many components of trustworthiness that should apply to AI systems regardless of the riskiness of their use — for example, if an AI system relies upon human data or has an unsustainable environmental footprint.

In this paper, we propose that trustworthiness should not be limited to high-risk applications. Instead, we argue that trustworthiness requirements should be based upon properties of an AI system that include design, development, testing, and impact considerations in addition to the intended application area. The taxonomy provided in this paper can supplement the EU AI Act by providing a voluntary framework that AI stakeholders can use to consider general trustworthiness of AI systems, even if considered to be minimal risk.

The White House Blueprint for an AI Bill of Rights

The United States White House Office of Science and Technology Policy published "The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People" in October 2022, following a year of engagement with policymakers throughout the Federal government and with the American public. The Blueprint is non-binding, but is intended to inform policy decisions. It identifies five principles and associated practices intended to "guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence."[49] The blueprint aims to align AI with democratic values and protect civil rights, civil liberties, and privacy.

[49] "Blueprint For an AI Bill of Rights: Making Automated Systems Work for the American People," The White House Office of Science and Technology Policy, October 2022, whitehouse.gov/ostp/ai-bill-of-rights.

The five principles are the following:

1. **You should be protected from unsafe or ineffective systems.** For example, systems should have ongoing monitoring and evaluation, and be removed from use if necessary.
2. **You should not face discrimination by algorithms and systems should be used and designed in an equitable way.** For example, designers should conduct proactive equity assessments and ensure the use of representative data and protection against proxies for demographic features.
3. **You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used.** For example, designers, developers, and deployers should seek permission about collection and use of your data where possible, and not use defaults that are privacy invasive. Surveillance technologies should be subject to heightened oversight, including restrictions in high-stakes settings.
4. **You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.** For example, you should be given notice that AI systems are in use, the individual or organization responsible for the system, and explanations of outcomes in a clear, timely, and accessible way.
5. **You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.** For example, you should be able to choose a human alternative when possible.

A Technical Companion is provided alongside the blueprint, which provides further information about each principle, including why it is important, expectations associated with the principle, and real-life examples of how different stakeholders are working to implement the principle through policy or practice.

The Blueprint for an AI Bill of Rights speaks primarily to the American public, providing expectations of rights they should expect from AI designers, developers, and deployers. This provides a valuable addition to the AI governance landscape because it centers people rather than AI technologies and their developers. The Blueprint does not ask people to trust AI technologies, but rather establishes a new set of expectations about how AI designers, developers, and deployers must protect the American public in the age of artificial intelligence.

The taxonomy introduced in this paper is complementary with the Blueprint for an AI Bill of Rights and can provide support for the realization of the five principles. For example, the Blueprint's call for safe and effective systems incorporates elements of several characteristics of trustworthiness including valid and reliable, safe, and secure and resilient.

The properties of trustworthiness provided for each of these characteristics can help achieve the principle.

The National Institute of Standards and Technology AI Risk Management Framework

The National Institute of Standards and Technology (NIST) was mandated by the United States Congress to develop an AI Risk Management Framework (AI RMF) to offer guidance for the development and use of trustworthy AI.[50] In July 2021, NIST first requested input from stakeholders on the idea of an AI risk management framework. Since then, NIST has published an AI RMF concept paper, followed by a first and second draft. In January 2023, NIST published the official first version of the NIST AI RMF, which is the primary reference used in this paper.[51] At the same time, NIST also released a companion AI RMF Playbook, which includes suggested actions, references, and documentation guidance.[52]

The AI RMF is intended for voluntary use to address risks in the design, development, use, and evaluation of AI products, services, and systems in support of trustworthy AI. The AI RMF defines trustworthy AI as being "valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed."[53]

The core of the AI RMF is composed of four functions: govern, map, measure, and manage. Each of the functions is then broken down into categories and subcategories. The govern function is intended to help cultivate a culture of risk management; the map function is intended to help recognize context and identify risks; the measure function is intended to help assess, analyze, or track risks; and the manage function is intended to help prioritize and act upon identified risks. The categories and subcategories break these functions down into numerous components, while the playbook provides actions, documentation guidance, and references for each subcategory. The AI RMF is designed to be applied in an iterative manner and used throughout the AI lifecycle.

[50] "AI Risk Management Framework: Initial Draft," National Institute of Standards and Technology, March 17, 2022, https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf.

[51] "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, January 26, 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf."

[52] "NIST AI Risk Management Framework Playbook," National Institute of Standards and Technology, August 18, 2022, https://pages.nist.gov/AIRMF/.

[53] "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, January 26, 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf."

This paper and the taxonomy of trustworthiness it introduces can ideally supplement use of the NIST AI RMF because it provides an approach for connecting what are currently three relatively disparate elements: the core framework, the AI lifecycle, and the characteristics of trustworthiness. This work highlights how these elements connect to each other, and provides examples of how someone might want to reference multiple parts of the AI RMF core at each stage of the lifecycle as they work on particular challenges.

Properties of Trustworthiness

This paper introduces a taxonomy of trustworthiness for artificial intelligence that includes 150 properties. Each property relates to one of seven "characteristics of trustworthiness" as defined in the NIST AI RMF: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. The properties provide greater nuance about each characteristic, and are linked to a particular stage of the AI lifecycle.[54] Each property of trustworthiness offers a distinct lens through which to assess the trustworthiness of an AI system and points to a set of questions and decisions to be made.

The taxonomy also includes an eighth characteristic of trustworthiness, "responsible practice and use." The NIST AI RMF discusses the importance of responsible practice and use of AI systems and suggests that it enhances trustworthiness, but it is not included as one of the characteristics of trustworthiness. We include it as a characteristic of trustworthiness in the taxonomy because we find that human decisions and practices play a critical role in the realization of trustworthy AI and that numerous properties align more closely with this concept than with the other characteristics.

The overall trustworthiness of an AI system is dependent upon the holistic consideration of all properties. For example, an AI system that is reliable and safe, but is not made transparent or explainable to users, is unlikely to be trusted. There are also many interdependencies between the properties. For example, the protection of human dignity may not be possible without also ensuring other properties, such as human control and the prevention of social or behavioral manipulation.

There are also tensions between properties that can arise,[55] for example between explainability and security.[56] In some cases an organization will need to make tradeoffs between some of the properties, either due to resource constraints or conflicts between two or more properties. Which properties to prioritize will depend on the context of the particular organization, AI

[54] The taxonomy is organized into seven stages of the AI lifecycle, as defined 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, and Use or Impacted By.

[55] Jess Whittlestone et al., "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, January 2019, https://dl.acm.org/doi/abs/10.1145/3306618.3314289.

[56] Dang Minh et al. "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, volume 55, 2022, link.springer.com/article/10.1007/s10462-021-10088-y.

system, and use case. Previous work highlights the extent to which prioritization of AI principles already occurs, and the importance of enabling transparency about what is prioritized and why.[57]

The people best suited to address a particular property will vary depending on the organization and the type of system being developed. In general, we expect that multidisciplinary and diverse teams (characteristics that are properties of trustworthiness themselves) will be critical, and that many roles will have relevant input on how to effectively consider the properties. Some of the properties, such as those relating to organizational processes and policies, will require involvement of organizational leadership. Other properties, such as those relating to data privacy and security, will require involvement of legal teams. We expect that in many cases, decisions about a particular property will not be made by a single person or team, but rather there will be chains of interactions among people with different roles who bring unique expertise. Although each property is linked to one of seven stages of the AI lifecycle (to signify when it is likely to be especially important or have a unique window of opportunity), many properties will also require ongoing oversight and management throughout the lifecycle. It will also be critical that the teams exploring the trustworthiness of AI systems have the power to influence or implement necessary changes.

Some of the properties have relevant standards, metrics, or benchmarks that provide guidance about how to address them to a sufficient degree. For example, there are AI wellbeing metrics,[58] AI fairness metrics,[59] and benchmarks for the performance and energy efficiency of AI processors,[60] among many others. To the extent that these exist, they will ideally be included as resources in the NIST AI RMF. Other properties represent emerging areas of inquiry and do not yet have established standards, metrics, or benchmarks. In these cases, quantification and measurement may still help track changes over time, but the methods are more likely to change.

[57] "Bridging AI's trust gaps: Aligning policymakers and companies," EY and The Future Society, July 2020, https://thefuturesociety.org/wp-content/uploads/2020/07/tfs-bridging-ais-trust-gaps-report.pdf.

[58] "IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being," in *IEEE Std 7010-2020* , pp.1–96, 1 May 2020, https://ieeexplore.ieee.org/document/9084219.

[59] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1-4:15, 1 July-Sept. 2019, ieeexplore.ieee.org/abstract/document/8843908.

[60] Y. Wang et al., "Benchmarking the Performance and Energy Efficiency of AI Accelerators for AI Training," 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, 2020, https://ieeexplore.ieee.org/abstract/document/9139681.

We sourced the properties from dozens of documents from AI safety, security, ethics, and governance literature.[61] The properties were also informed by interviews, feedback, and an expert workshop.[62] The list of properties is long, but should not be assumed to be fully comprehensive. Many of the trustworthy AI documents reviewed originated in western democratic nations and may fail to account for global variation of values.[63] The list also does not incorporate regional regulations and should not be interpreted as providing a list of properties that would ensure compliance with any regulation. The fields that inform trustworthy AI are actively growing and evolving and the properties will change with them. We hope others will supplement the taxonomy over time.

DO THE PROPERTIES APPLY TO ALL AI SYSTEMS?

Some frameworks and policies for trustworthy AI encourage users to focus on AI uses that pose the greatest risk. While we agree that the properties of trustworthiness are especially important for high-risk use cases, we find that they remain relevant to AI systems with lower risk. This approach is consistent with the NIST AI RMF, which indicates that trustworthiness is important for all AI systems, though the mechanisms of risk management may vary depending on the severity of risk. For example, protecting human dignity may be especially critical for a system that scales to millions of people, but it still matters at small scales, such as the organizational level. Many of the properties are also critical regardless of the final use case, such as those relating to data practices, the security of a system, or the environmental footprint of a system. Another challenge with focusing on the intended use of an AI system

[61] For example, see: “Blueprint For an AI Bill of Rights,” The White House Office of Science and Technology Policy, whitehouse.gov/ostp/ai-bill-of-rights; “Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment,” July 17, 2020, High-Level Expert Group on Artificial Intelligence, European Commission, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment; “Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities,” U.S. Government Accountability Office, June 2021, https://www.gao.gov/assets/gao-21-519sp.pdf; “Recommendation on the Ethics of Artificial Intelligence,” United Nations Educational, Scientific and Cultural Organization, 2022, https://unesdoc.unesco.org/ark:/48223/pf0000381137; Jessica Fjeld et al. “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI,” January 15, 2020, Berkman Klein Center Research Publication No. 2020-1, https://ssrn.com/abstract=3518482.

[62] The virtual workshop held in July 2022, “Properties of Trustworthiness for Artificial Intelligence,” was co-organized by the UC Berkeley Center for Long-Term Cybersecurity and Intel. More than 40 experts participated, from academia, government, industry, and civil society. The names of some of the workshop participants are included in the acknowledgments.

[63] Shakir Mohamed, Marie-Therese Png, and William Isaac, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence,” Philosophy and Technology (405), July 2020, https://arxiv.org/pdf/2007.04068.pdf.

to determine a risk level is that it may not be its only use. Indeed, some AI systems may have hundreds or thousands of potential uses.[64]

Rather than looking at a small number of final high-risk use cases to determine the relevancy of properties, we opted instead to focus on a set of questions that reveal more about the nature of the interaction between the system and people. This approach offers a way of differentiating the extent of care needed based upon the context of use (e.g., that accessibility matters more for systems that have human users), but does not amount to an assertion that only a small subset of AI uses require trust.

Spectrum of Human-AI Engagement

Human engagement is often assumed for AI technologies because they are designed by people, and we typically imagine them being used by people. However, human engagement with AI systems exists on a spectrum, and it can be useful to consider a set of questions about a particular AI system and its possible uses to assess the degree of human engagement involved. This can help generate a more nuanced picture of what trustworthiness entails in each specific context.

The following questions offer examples that help clarify different elements of potential human-machine engagement. Additional questions may help provide further detail.

Questions to consider:

1. Does the AI system rely upon or generate data about people, including sensitive or personally identifiable information?
2. Does the AI system have human users or operators, or otherwise engage with people?
3. Does the AI system inform human decision-making?

Different answers to these questions would place an AI system at different points along the spectrum of human-AI engagement. For example, if the answer to all of these questions is no, the AI system is probably not human-facing to a meaningful degree, and some properties of trustworthiness may not be relevant. Examples of non-human facing AI systems may include

[64] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arxiv, August 2021, https://arxiv.org/abs/2108.07258.

systems that provide terrain analysis or weather metrics. In these cases, the AI systems may not rely upon data about people and may only produce analytics for other AI systems. If the answers to the questions are mixed, the AI system may be partially human-facing, and a different set of properties of trustworthiness may be relevant.

An AI system may change its degree of human engagement at some point during its lifecycle, and so any assessment of human engagement should be revisited at regular intervals. It is not the point of this paper to define discrete points along the spectrum of human-AI engagement, but rather to highlight that understanding the degree of engagement an AI system has with people plays a role in what is likely to matter for people to trust that system.[65]

Importantly, consideration of where an AI system is situated on the spectrum of human-AI engagement is distinct from consideration of its impact on people. AI systems may be considered non-human facing and still be highly impactful for people, society, or the environment. Conducting a risk and impact assessment is still a critical process for any AI system, and should be done in addition to considering the degree to which an AI system engages with people.

To help demarcate the properties of trustworthiness that are especially impacted by the degree of human engagement an AI system has, a subset of the properties in the taxonomy have an asterisk after their name. This signifies that the property is likely to be less relevant for AI systems that are not human-facing. We hope that future work will build upon these ideas and further explore the relationships between risk, trust, and human engagement for AI systems.

TAXONOMY OF TRUSTWORTHINESS FOR ARTIFICIAL INTELLIGENCE

The taxonomy of trustworthiness for artificial intelligence introduced below provides a novel tool for AI developers, policymakers, and others to consider how the complex notion of trustworthiness may be integrated into a risk management process carried out throughout the AI lifecycle.

[65] The importance of whether an AI system is human-facing or not human-facing, or somewhere along the continuum, is explored further in an August 2022 Response to the Request for Comments on the NIST AI RMF Playbook provided by Intel Corporation. The response includes helpful use cases for illustration.

There are many different depictions of the AI lifecycle, which have been defined by different stakeholders.[66] There is some variation between the number and names of different stages, but also significant similarity among them. The taxonomy below uses the depiction of the AI lifecycle given in the NIST AI RMF. Keep in mind that not every characteristic of trustworthiness has properties that map to it for every lifecycle stage.

The Seven AI System Lifecycle Stages, as defined 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

Each property included in the taxonomy of trustworthiness is 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 in the NIST AI RMF Playbook 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 the AI system development and use.

[66] There are other notable depictions of the AI lifecycle, including from the OECD AI Recommendation, the ISO/IEC/IEEE 12207:2017 standard, and, by extension, the older ISO/IEC TR 24748-1:2010 standard, which the OECD partially draws upon. Many AI lifecycles have substantial overlap, as noted in a paper from Oxford University, "CapAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act," which includes a comparison of six prominent models of the AI lifecycle.

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

Implications and Further Research

The Taxonomy of Trustworthiness includes 150 properties of trustworthiness, each of which has a set of questions to prompt consideration at a particular point in the AI lifecycle. This provides a thorough starting point for a team to begin to map their expectations and responsibilities to support trustworthy development and use of an AI system. The taxonomy is designed to be compatible with the NIST AI RMF, as it makes use of the same AI lifecycle and core characteristics of trustworthiness. The taxonomy also links each property of trustworthiness to particular sections ("subcategories") of the NIST AI RMF and associated playbook, where a reader can find additional resources and guidance about how to govern, map, measure, and manage the particular property.

The taxonomy may serve as a resource and tool for organizations developing AI, as well as for standards-setting bodies, policymakers, independent auditors, and civil society organizations working to evaluate and promote trustworthy AI. We hope it serves as a useful resource alongside the NIST AI RMF, helping users to further explore the relationships between NIST's core framework, characteristics of trustworthiness, and depiction of the AI lifecycle.

Through the development of the taxonomy, we found that most of the properties of trustworthiness typically remain relevant for all AI systems, but that some properties are less relevant, or not at all relevant, to AI systems that are less visible to human users. This concept is explored further in the section above, "Spectrum of Human-AI Engagement," in which a set of questions are provided to prompt consideration of the degree of human-AI engagement a particular system might have. Any categorization of a system at a particular degree of human engagement should be regularly revised to account for shifts in use or context.

One of the interesting discoveries of this research is that there are properties of trustworthiness that are unlikely to be relevant to AI systems that have minimal human engagement. For example, properties relating to privacy and mitigating computational bias are less relevant to AI systems that do not rely upon or generate data about people, including sensitive information or personally identifiable information. Similarly, properties relating to user testing and engagement, usability, communication, and accessibility are less relevant to AI systems that do not have human users or operators.

However, it is notable that the majority of the properties remain important across the spectrum of human engagement. This is largely because effectively all AI systems are designed and developed by people, and can impact our shared environment or high-stakes settings, even if they are not ultimately used or operated by people. For example, it is still important to conduct risk and impact assessments and ensure continuous monitoring for non-human-facing systems to ensure ongoing effectiveness, safety, and security. It is also still important to consider properties such as justice and the protection of human rights, which include aspects of design and impact beyond immediate engagement with or use of the system.

It is interesting that all of the properties related to safety and security are likely to be relevant irrespective of the degree of human engagement. We believe this to be the case because people expect technological tools to be safe and secure regardless of use. Nonetheless, the impact of safety or security failures will vary depending on how the AI system is used, and higher standards are typical for more high-stakes settings.

The segmentation of the properties of trustworthiness across parts of the AI lifecycle, including connecting them to available tools and resources for implementation as found in the NIST AI RMF, is intended to provide further nuance and practicality. Further research would be useful to pilot the use of this framework with organizational teams and to develop case studies for using the taxonomy across different domains.

Conclusion

This paper recognizes the growing interest in developing and using trustworthy AI, but notes that significant challenges remain in achieving the aspiration of trust in AI. The recent development of the NIST AI Risk Management Framework (RMF) and its articulation of trustworthy characteristics provides an important mechanism for organizations hoping to reduce risk and improve trust in their AI technologies. This paper introduced a taxonomy of trustworthiness for artificial intelligence that can be used as a resource alongside the NIST AI RMF. The taxonomy connects properties of trustworthiness to particular stages of the AI lifecycle, and guides the reader to relevant sections of the core NIST AI RMF, where they can find additional resources about implementing the property.

Finding that many previous frameworks for trustworthy AI primarily focus on human-facing applications of AI, but that many aspects of trustworthiness remain critical for all AI systems, this paper also takes an approach that is intended to apply to a broader spectrum of AI types and applications. The paper includes a discussion of the spectrum of human-AI engagement and encourages consideration of a set of questions that can help determine whether some properties of trustworthiness may be more or less relevant.

The organization and classification of the properties, while imperfect, offers a starting point to teams attempting to build, use, or procure AI systems for a wide variety of purposes. Taken as a whole, the taxonomy builds on efforts to improve the practicality of ethical and responsible AI guidance in a way that is fully flexible to context without being so vague as to invite misuse.[67,68] The goal of the taxonomy is to help support efforts toward developing trustworthy AI and using it in a responsible way, today and into the future.

Key findings of this paper include the following:

1. **Many stakeholders have a role to play in developing and ensuring trustworthy AI.** Fully considering the trustworthiness of an AI system requires diverse and multidisciplinary expertise. The process should include a broad range of roles from within an organization

[67] Jessica Morley et al., “Operationalising AI ethics: barriers, enablers and next steps,” AI and SOCIETY, 2021, https://link.springer.com/article/10.1007/s00146-021-01308-8.

[68] Jessica Morley et al., “Ethics as a service: a pragmatic operationalisation of AI Ethics,” arxiv, February 2021, https://arxiv.org/pdf/2102.09364.pdf.

as well as outside experts, including members of impacted communities and independent verification and auditing bodies.

2. **The consideration of trustworthiness should not wait until after an AI system has been developed.** Many properties of trustworthiness are most critical in the early design phase.
3. **Many properties of trustworthiness are relevant regardless of whether an AI system is “high risk.”** For example, properties related to safety, quality, and sustainability tend to matter regardless of application area. This means that it is critical to consider trustworthiness even for AI applications that do not qualify as “high risk,” and that frameworks for trustworthy AI that primarily focus on high-risk applications may not be sufficient.
4. **Some properties of trustworthiness are less relevant for AI applications that are not human-facing.** For example, some properties of trustworthiness relate to interactions with users, but not all AI systems call for interactions with users.
5. **Striving for trustworthy AI is a complex and ongoing process, not an easily achievable outcome.** Organizations should be wary of applying easy-fix solutions to complex technical and social problems. There are numerous properties of trustworthiness, some of which are active areas of research that may not yet have obvious and available solutions. Building trustworthy AI systems should be seen as an ongoing process to earn trust, rather than an easily achievable outcome.

Appendix 1

CONNECTING THE TAXONOMY OF TRUSTWORTHINESS TO INTERNATIONAL AI STANDARDS

The table below includes select international AI standards related to each of the AI trustworthy characteristics. These standards can provide guidance about achieving the characteristics of trustworthiness and their associated properties, which may help supplement guidance in the relevant NIST AI RMF sections, which are documented in the primary taxonomy.

Only a select number of high-profile standards-setting and governmental institutions are included: the Organisation for Economic Co-operation and Development (OECD), European Commission, National Institute of Standards and Technology (NIST), International Organization for Standardization (ISO), Institute of Electrical and Electronics Engineers (IEEE), ITU Telecommunication Standardization Sector (ITU-T), and European Telecommunications Standards Institute (ETSI). The mapping of relevant standards to the characteristics of trustworthiness is inspired and informed by an earlier effort to map AI standards to the developing regulatory framework for AI in the EU.[69]

AI Trustworthy Characteristics	Relevant International AI Standards
Valid and Reliable	"Ethics guidelines for trustworthy AI," High-Level Expert Group on AI, European Commission "Proposal for a Regulation laying down harmonised rules on artificial intelligence," European Commission "AI Risk Management Framework," National Institute of Standards and Technology "Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems," ISO/IEC 25059 (in development) "Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality," ISO/IEC 25024:2015 "Artificial intelligence — Overview of trustworthiness in artificial intelligence," ISO/IEC 24028:2020 (in development) "Artificial Intelligence — Assessment of machine learning classification performance," ISO/IEC 4213.2 (in development) "Securing Artificial Intelligence (SAI); Data Supply Chain Security," ETSI GR SAI 002 "Requirements for machine learning-based quality of service assurance for the IMT-2020 network," ITU-T Y.3170 "Cloud computing - Functional requirements for machine learning as a service," ITU-T Y.3531

[69] "AI Watch: AI Standardisation Landscape state of play and link to the EC proposal for an AI regulatory framework," European Commission Joint Research Centre, 2021, https://www.standict.eu/sites/default/files/2021-07/jrc125952_ai_watch_task_9_standardization_activity_mapping_v5.1%281%29.pdf.

Continued

Safe	"Recommendation of the Council on Artificial Intelligence," OECD "Ethics guidelines for trustworthy AI," High-Level Expert Group on AI, European Commission "Proposal for a Regulation laying down harmonised rules on artificial intelligence," European Commission "AI Risk Management Framework," National Institute of Standards and Technology "Artificial intelligence — Functional safety and AI systems," ISO/IEC 5469 (in development) "Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview," ISO/IEC 24029-1:2021 (in development) "Artificial intelligence — Overview of trustworthiness in artificial intelligence," ISO/IEC 24028:2020 (in development) "Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems," IEEE P7009
Fair with Harmful Bias Managed	"Recommendation of the Council on Artificial Intelligence," OECD "Ethics guidelines for trustworthy AI," High-Level Expert Group on AI, European Commission "Proposal for a Regulation laying down harmonised rules on artificial intelligence," European Commission "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," National Institute of Standards and Technology "AI Risk Management Framework," National Institute of Standards and Technology "Artificial intelligence (AI) — Bias in AI systems and AI aided decision making," ISO/IEC 24027:2021 (in development) "Artificial intelligence — Overview of trustworthiness in artificial intelligence," ISO/IEC 24028:2020 (in development) "Artificial intelligence — Overview of ethical and societal concerns," ISO/IEC 24368 (in development) "The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)," IEEE SA "Model Process for Addressing Ethical Concerns During System Design," IEEE 7000-2021 "Algorithmic Bias Considerations," IEEE P7003 "IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being," IEEE Std 7010-2020
Secure and Resilient	"Recommendation of the Council on Artificial Intelligence," OECD "Ethics guidelines for trustworthy AI," High-Level Expert Group on AI, European Commission "Proposal for a Regulation laying down harmonised rules on artificial intelligence," European Commission "AI Risk Management Framework," National Institute of Standards and Technology "Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview," ISO/IEC 24029-1:2021 (in development) "Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods," ISO/IEC 24029-2 (in development) "Artificial intelligence — Overview of trustworthiness in artificial intelligence," ISO/IEC 24028:2020 (in development) "Recommended Practices for Virtual Classroom Security, Privacy and Data Governance," IEEE P7004.1 "Securing Artificial Intelligence (SAI); AI Threat Ontology," ETSI GR SAI 001 "Securing Artificial Intelligence (SAI); Data Supply Chain Security," ETSI GR SAI 002 "Securing Artificial Intelligence (SAI); Security Testing of AI," ETSI DGR/SAI-003 "Securing Artificial Intelligence (SAI); Problem Statement," ETSI GR SAI 004 "Securing Artificial Intelligence (SAI); Mitigation Strategy Report," ETSI GR SAI 005 "Securing Artificial Intelligence (SAI); The role of hardware in security of AI," ETSI GR SAI 006 "Securing Artificial Intelligence (SAI); Explicability and transparency of AI processing," ETSI GR SAI 007

Continued

Explainable and Interpretable	“Recommendation of the Council on Artificial Intelligence,” OECD “Ethics guidelines for trustworthy AI,” High-Level Expert Group on AI, European Commission “Proposal for a Regulation laying down harmonised rules on artificial intelligence,” European Commission “Four Principles of Explainable Artificial Intelligence,” National Institute of Standards and Technology “AI Risk Management Framework,” National Institute of Standards and Technology “Artificial intelligence — Overview of trustworthiness in artificial intelligence,” ISO/IEC 24028:2020 (in development) “Securing Artificial Intelligence (SAI); Problem Statement,” ETSI GR SAI 004 “Securing Artificial Intelligence (SAI); Explicability and transparency of AI processing,” ETSI GR SAI 007
Privacy-Enhanced	“Recommendation of the Council on Artificial Intelligence,” OECD “Ethics guidelines for trustworthy AI,” High-Level Expert Group on AI, European Commission “Proposal for a Regulation laying down harmonised rules on artificial intelligence,” European Commission “AI Risk Management Framework,” National Institute of Standards and Technology “Artificial intelligence — Overview of trustworthiness in artificial intelligence,” ISO/IEC 24028:2020 (in development) “Data Privacy Process,” IEEE 7002-2022 “Standard on Child and Student Data Governance,” IEEE P7004 “Recommended Practices for Virtual Classroom Security, Privacy and Data Governance,” IEEE P7004.1 “Standard on Employer Data Governance,” IEEE 7005-2021 “IEEE Guide for Architectural Framework and Application of Federated Machine Learning,” IEEE 3652.1-2020 “Cloud computing - Functional requirements for machine learning as a service,” ITU-T Y.3531
Accountable and Transparent	“Recommendation of the Council on Artificial Intelligence,” OECD “Ethics guidelines for trustworthy AI,” High-Level Expert Group on AI, European Commission “Proposal for a Regulation laying down harmonised rules on artificial intelligence,” European Commission “AI Risk Management Framework,” National Institute of Standards and Technology “Four Principles of Explainable Artificial Intelligence,” National Institute of Standards and Technology “Artificial intelligence — Overview of trustworthiness in artificial intelligence,” ISO/IEC 24028:2020 (in development) “Artificial intelligence — Process management framework for big data analytics,” ISO/IEC 24668 (in development) “Artificial intelligence — Guidance on risk management,” ISO/IEC 23894 (in development) “Governance of IT — Governance implications of the use of artificial intelligence by organizations,” ISO/IEC 38507:2022 “Artificial intelligence — Management system,” ISO/IEC 42001.2 (in development) “Model Process for Addressing Ethical Concerns During System Design,” IEEE 7000-2021 “Standard for Data and Artificial Intelligence (AI) Literacy, Skills, and Readiness,” IEEE P7015 “Recommended Practice for Organizational Governance of Artificial Intelligence,” IEEE P2863 “The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS),” IEEE SA “Transparency of Autonomous Systems,” IEEE 7001-2021 “Recommended Practice for Environmental Social Governance (ESG) and Social Development Goal (SDG) Action Implementation and Advancing Corporate Social Responsibility,” IEEE P7010.1 “Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems,” IEEE P7008 “Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems,” IEEE P7014 “Securing Artificial Intelligence (SAI); Explicability and transparency of AI processing,” ETSI GR SAI 007

Appendix II

THE PROPERTIES OF TRUSTWORTHINESS WITHOUT SEGMENTATION BY LIFECYCLE STAGE

NIST Characteristics of Trustworthiness	Properties of Trustworthiness	Question(s) to Consider
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?
	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?
	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?
	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)?
	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?
	Trusted Dependencies on External Parties	How will we identify, assess, and monitor our dependencies on external parties?
	Foresight and Scenario Planning	How will we assess and navigate possible futures and the evolving risk landscape?
	Data Completeness	How will we assess and improve the completeness, quantity, suitability, and representativeness of the data?
	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?
	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?
	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?
	Reproducible	How will we test whether desirable outputs of the AI system can be reproduced in different circumstances?
	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?
	Verifiable	How will we verify that the system is behaving as expected?

Continued

	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?
	Replayable	How can we replay the behavior of the system to see if the same input generates the same output?
	Effective	How will we judge sufficient effectiveness of the AI system, in the lab and in the real world?
	Valid	How will we validate the outputs of the AI system, including through external validation?
	Appropriate Capabilities for the Tasks	How will we review whether the capabilities of the AI system are appropriate for a particular use and context?
	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?
	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?
	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?
	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?
	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?
	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?
	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?
	Human Agency	How will human agency be meaningfully incorporated in the operation of the AI system?
	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?
	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?
	Appropriate Retirement	How will we determine when and how to retire the use of the AI system?
	Iterative Learning and Improvements	How will we continue to learn, iterate, and improve over time?
	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?

Continued

	Effective Feedback*	How will we establish a dedicated channel for feedback and questions about the AI system from users and the general public?
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?
	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?
	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?
	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?
	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?
	Data Stability	How will we analyze and monitor for data drift over time?
	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?
	Loyal	To whom or what will the AI system be “loyal,” and will that be optimal and made transparent?
	Power-averse	How will we incentivize models to avoid power or avoid gaining more power than is necessary?
	Containment	How can we contain the AI system to prevent safety and security breaches?
	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?
	Review	How will we review any errors or inconsistencies with the AI system that emerge?
	Effective Detection of Anomalies	How will we detect potential novel hazards?
	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?
	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?
	Awareness of Functional Evolution	How will we track shifts in the AI system’s functionality over time?
	Assurance / Management of Emergent Functionalities	How will we predict and detect new capabilities and goals of the AI system?
	Incident Reporting	How will we publicly report incidents and adverse impacts of the AI system, such as mistakes, errors, breaches, unintended consequences, etc.?

Continued

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?
	Inclusive	How will we ensure inclusivity of all relevant experts and communities in the design and development of the AI system?
	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?)
	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?)
	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?
	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?
	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?
	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?
	Non-Discrimination*	How will we ensure the AI system is not discriminatory across gender, racial, ability, age, political beliefs, religion, or other dimensions?
	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?
	Shared Benefit	How will the benefits of the AI system's use be distributed? Can those benefits be shared more widely?
	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?
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?
	Availability	How will we ensure that information for and about the AI system is available to authorized personnel when it is needed?
	Confidentiality	How will we ensure that information is not made available or disclosed to unauthorized individuals, entities, or processes?
	Integrity	How will we maintain and ensure the accuracy, completeness, and appropriateness of data, models, and procedures informing the AI system?
	Data Security	How will the security of data that is used for training or created be ensured?
	Protection Against Trojans	How will we detect if there is hidden functionality embedded in our models?

Continued

	Built-in Defenses	How will the AI system respond to attacks as they occur?
	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?
	Resilient	How will we assess the AI system's ability to handle uncertainty and unknown environments?
	Use of Adversarial Testing	How will we establish "bug bounties" and enable "red teams" to try to deliberately find vulnerabilities in the AI system?
	Vulnerability Disclosure	How will we establish a coordinated policy to encourage responsible vulnerability research and disclosure?
Explainable and Interpretable	Intelligible*	How will we assess the system for intelligible explanations and select a model to support this?
	Positive Human-Machine Interaction*	How will we enable positive human-machine interactions throughout the AI system's operation?
	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?
	Interpretable*	How will we judge the interpretability of the system's explanation to the particular context and user?
	Relevant Explanation	How will we judge how informative and relevant a system's explanation is to the particular context and user?
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?
	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?
	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?
	Data Processing Oversight*	How will we establish data oversight mechanisms, such as limiting and logging data access?
	Consent to Use of Data*	How will we enable people to consent to the uses of their data?
	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?
	Model Protection*	How will we protect model access that could reveal sensitive information?
	Protection from Unwarranted Data Access*	How will we ensure the AI system cannot be used to give unwarranted access to data?
	Effective Notification*	How will we notify users and impacted communities about privacy or security breaches, or other incidents?
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?
	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?

Continued

	Coordination (Public-Private; International)	How will we identify and coordinate with relevant institutions, nationally and internationally?
	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?
	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?
	Open	How can we promote openness and transparency about our development and governance of AI technologies, internally and externally?
	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.?
	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?
	Internal Reviews	How will internal reviews be conducted to assess trustworthy AI practices?
	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?
	Traceable	How will we document the provenance of data, processes, and artifacts involved in the production of the AI system?
	System Honesty	How will we ensure the AI system only presents outputs that are accurate and not intentionally deceptive?
	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?
	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?
	Information-sharing	How will we share critical information about our AI system with relevant authorities and stakeholders?
	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?
	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?
	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?
	Facilitation of Contestability*	How will users be able to contest or appeal a decision or action made by the AI system?

Continued

	Facilitation of Redress or Recourse	How will we support or compensate people who are negatively affected by the use of the AI system?
	Engagement with Global Governance Deliberations	How will we analyze, follow, and engage in relevant global governance deliberations and practices related to artificial intelligence?
	Data and System Accessibility	How can we enable access to the AI system and datasets to relevant authorities, independent researchers, and trusted intermediaries?
	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?
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?
	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?
	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?
	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?
	Verified Supply Chain	How will we assess and verify the relevant components of the supply chain?
	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?
	Effective Capabilities	How will we obtain the necessary resources and knowledge to achieve our trustworthy AI objectives?
	Collaboration	How will we enable multi-stakeholder collaboration?
	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?
	Effective Hiring and Training	How will we support the hiring and training of individuals who can carry out trustworthy AI objectives?
	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?
	Leadership Commitment	How will we ensure long-term commitment to trustworthy AI from organizational leadership?
	Supportive Organizational Culture	How will our organizational culture support our trustworthy AI objectives? Are changes needed?
	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?

Continued

	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?
	Alignment with Organizational Vision, Mission, and Values	How will we ensure the AI system is true to our vision, mission, and values?
	Socially Responsible	How will our AI system and its use align with our social responsibility efforts?
	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?
	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)?
	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?
	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?
	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?
	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?
	Supportive of Wellbeing	How will we ensure the AI system supports individual, community, and societal wellbeing, including mental or emotional wellbeing?
	Reduction of Carbon Emissions	How can we reduce the carbon emissions from the design and use of AI systems in general?
	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?
	Efficient Data Centers	How can we make our use of data centers more energy-efficient?
	Reduction of Computational Requirements	How can we reduce the computational requirements of the AI system?
	Beneficial to Society	How will we ensure the AI system will be leveraged to benefit society?
	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?

Continued

	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?
	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?
	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?
	Assessment of Environmental Implications	How will we analyze and document the environmental implications of the AI system and its uses?
	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?
	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?
	Ability to Opt Out*	How will we ensure that people have specific and clear opportunities to opt out of use of the AI system?
	Consumer Protection*	How will we protect consumers or users of the system from harm?
	Due Process and Protection	How will we protect whistleblowers, NGOs, trade unions, or other entities who come forward with concerns about the AI system?

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 (AISI), housed at the UC Berkeley Center for Long-Term Cybersecurity. She is also the Co-Director of the UC Berkeley AI Policy Hub. Her work focuses on the governance, policy, and global security implications of artificial intelligence. She was previously an AI Policy Specialist with The Future of Life Institute and has held research positions with Harvard's Program on Science, Technology & Society, The Institute for the Future, The Center for Genetics and Society, and The Belfer Center for Science and International Affairs. Jessica received her master's degree in public policy from the Harvard Kennedy School and her bachelor's in anthropology from the University of California, Berkeley with highest distinction honors. Jessica serves as a member of the OECD Network of Experts on AI (ONE AI) and the IEEE Working Group on Recommended Practice for Organizational Governance of Artificial Intelligence. She is also a Research Advisor with The Future Society, a Senior Researcher with the Algorithmic Fairness and Opacity Working Group (AFOG), and an Affiliated Scholar with the CITRIS Policy Lab. She previously served as a Co-Chair for the University of California Presidential Working Group on Artificial Intelligence, the CNAS Task Force on Artificial Intelligence and National Security, and the Partnership on AI Expert Group on Fair, Transparent, and Accountable AI.

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

Center for Long-Term Cybersecurity | cltc.berkeley.edu | @CLTCBerkeley