MITRE

The AI Assurance Landscape (v1.0)

Toward a Standardized Framework of Unique and Differentiated AIA Concepts

Paul Ward, Jeff Stanley, Ron Ferguson, Joanna Korman

September 2024

McLean, VA

Contents

1 Introduction

An AI-enabled system is "*…a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*" (OECD, 2024, p.4). A wide range of AI methods and technologies—that transcend disciplinary boundaries—can be used either individually or as an ensemble to generate those inferences, make predictions, decisions and recommendations, or take actions. Hence, AI is not a single or simple system. Instead, it is multifaceted and typically complex and, invariably, part of a larger sociotechnical ecosystem (Robbins, Eris, Kapusta, Booker & Ward, 2024).

Recent investment has focused on the technological aspects of developing *effective* and *reliable* AI solutions. However, research and development are also needed to assure that AI-enabled systems minimize the risks to the sociotechnical system in which they are embedded and, ultimately, provide a benefit to society (Robbins, et al., 2024; Shneiderman, 2022). To achieve these disparate goals—which can trade-off with one another—a comprehensive view of AI assurance is required.

MITRE defines AI assurance[1] as *"…a process for discovering, assessing, and managing risk throughout the life cycle of an AI-enabled system so that it operates effectively to the benefit of its stakeholders"* (Robbins, et al., 2024, p.2). This definition is consistent with generally accepted goals of AI assurance: To ensure that deployed AI systems are (i) free of vulnerabilities across the entire lifecycle, (ii) function as intended, (iii) produce valid, verifiable, and robust outcomes, and (iv) act in a principled and ethical manner (e.g., Baterseh et al., 2021; Fjeld et al., 2020; Freeman et al., 2021; Robbins et al., 2024; Shneiderman, 2022, Tabassi, 2023). Below, we describe MITRE's AI Assurance (AIA) Landscape, its purpose, the process used to create the landscape and accompanying glossary, and the rationale for its development.

2 What is the AI Assurance Landscape?

The AIA Landscape is a visual representation of a synthesis of existing AIA frameworks and reports.[2] It captures a wide range of unique AIA needs and requirements documented in other AIA reports and frameworks developed by government departments/agencies, non-government entities, and industry worldwide. When not adequately addressed, AIA needs can result in potential risks that may have consequential technological, mission-related, and/or societal impacts.

The AIA Landscape is presented in Figure 1. It takes the form of a simple concept map containing a superordinate node (Trustworthy AI), 11 primary *categories* of AIA needs, and 66 sub-ordinate concepts referred to as specific assurance needs. Definitions of each term within the AIA Landscape are included in the accompanying glossary (see Appendix D).

[1] The term AI assurance and the associated assurance goals mentioned here are adapted from the fields of software assurance and quality assurance (as discussed in Freeman et al., 2021).

[2] A complete list of the frameworks reviewed in this synthesis can be found in The AIA Landscape Resource List below. In creating the landscape, MITRE paid an especial focus to some of the more comprehensive assurance frameworks, including NIST's AI Risk Management Framework (see Section 4: *Development of the AIA Landscape* for more details).

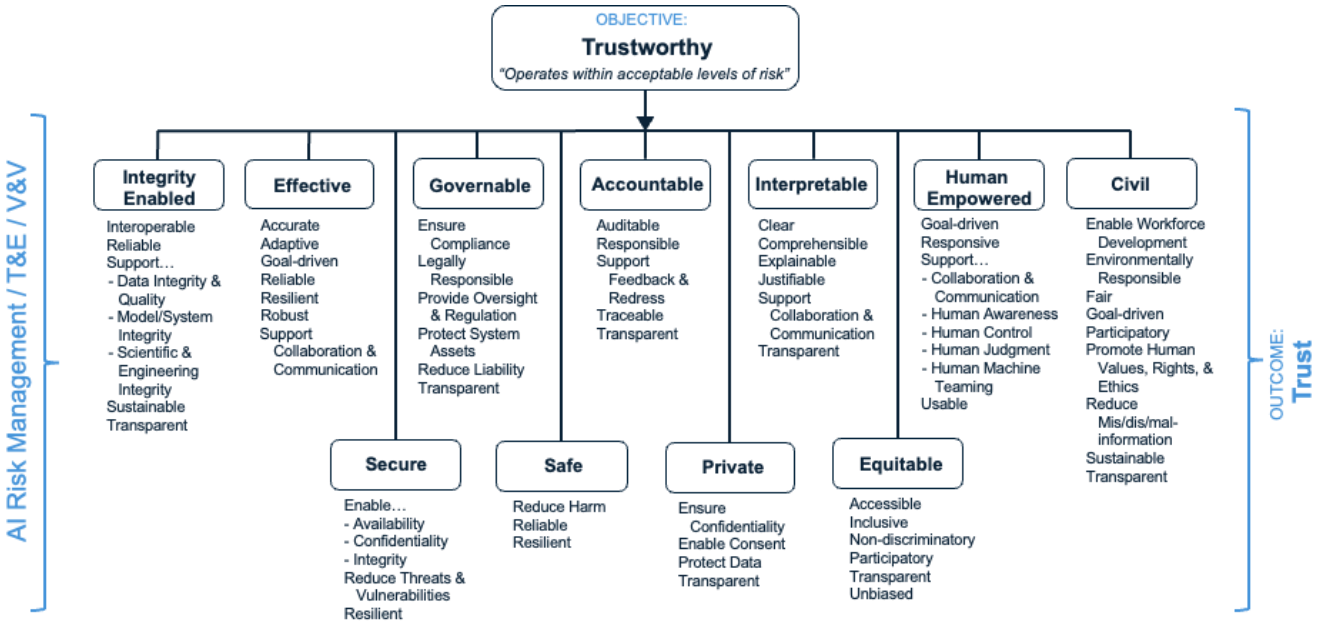

Figure 2. *The AIA Landscape*

3 Purpose of the AIA Landscape

The AIA landscape was developed as a first step toward creating a standard set of AIA concepts. The primary intent was to capture a comprehensive set of unique and specific needs that are documented across the collection of existing AIA-related frameworks and reports, and to organize these needs around a set of assurance categories in a single, domain-agnostic representation.

A second, and arguably more important, goal was to define, disambiguate, and disentangle related (and, elsewhere, grouped or synonymous) assurance concepts to permit differentiation between unique needs. Our expectation is that greater clarity and a better understanding of AIA needs and requirements will permit associated risks to be addressed more effectively.

The AIA Landscape and the associated glossary are key tools used in the *Risk Discovery Protocol for AI Assurance* (RDP-AIA) described in Version 1 of the protocol (see Ward, Stanley, Ferguson, Gladding, & Burns, 2023). In conjunction with the RDP-AIA, the landscape is intended to enable Stakeholder exploration of a more complete range of AIA needs and facilitate discovery of potential, emergent, or new risks. The AIA Landscape is not meant to be a replacement for an organization's own AIA framework. Rather, it is intended to act as a supplement to their existing efforts, enable Stakeholders to clarify their specific needs, and broaden their aperture on assurance-related risk assessment.

4 Development of the AIA Landscape

MITRE's RDP-AIA team engaged in an iterative process of analysis and synthesis of existing AIA resources that were available at the time of developing the AIA landscape (i.e., early-mid 2023). These resources included publicly available AIA frameworks and reports, associated AIA literature, input from Stakeholders and their AIA needs, subject-matter expert (SME) knowledge on AI, AIA, AI harms, AI risk management, and human-centered AI, and feedback from pilot testing with Stakeholders of the initial landscape. The complete list of frameworks and reports used to develop the AIA Landscape is available in the AIA Landscape Resources section below. The AIA Landscape development process included the following activities:

1) Environmental scan and review of existing AIA frameworks:

a. Approximately 50 AIA documents (incl. frameworks and reports) were identified that contained an AIA scheme.

b. Each of these frameworks was reviewed to:

   i. Determine the core assurance categories they contained.

   ii. Identify the most comprehensive frameworks amongst them (*n* = 10) (i.e., those that captured broad sets of assurance needs and requirements found across all available frameworks).[3]

2) Generation of a list of unique assurance concepts:

a. As a starting point, an initial list of assurance concepts was extracted from two of the most comprehensive frameworks reviewed (Fjeld et al., 2020; Tabassi, 2023).

   i. Concepts that had been treated as the same or similar (within or across frameworks) or had been grouped into a single category or need (e.g., security & resilience) were disambiguated or separated and given a working definition.

   ii. The same concept(s) referred to using different assurance terms (within or across frameworks) were combined and given a common reference and working definition (See Appendix C and Appendix D).

   iii. The list of unique assurance concepts generated was cross referenced to their source (see Table 1).

b. The remaining *comprehensive* frameworks and reports (*n* = 8) were reviewed, their assurance concepts added to the list of unique concepts (using a similar procedure to that used for the initial comprehensive frameworks) and cross-referenced to their source (see Table 2).

3) Categorization of assurance concepts:

[3] Framework schemes identified as most comprehensive and used as the initial and primary references to identify unique assurance concepts included: Department of Defense (DIB, 2019; DSOD, 2021; DOD, 2022), Dorton & Stanley (2023), Fjeld et al. (2020), GAO (2021), Leslie (2019), ODNI (2020a, 2020b), Shneiderman (2020, 2021, 2022/22), and Tabassi (2023). These were also cross-referenced with the CDAO (2023) Responsible AI Toolkit and White House (2020, 2023) Executive Orders 13960 and 14110—See Table 2.

a. An informal definition was created for each concept based on the source documents(s) (e.g., the original framework citing the concept in question).

b. Initial definitions were cross-validated and/or revised using established standards, academic literature, and/or common parlance as appropriate.[4] These definitions formed the basis for the initial AIA Landscape Glossary (see Appendix C and Appendix D).

c. A thematic analysis was performed of the list of unique assurance concepts to generate an initial set of AIA categories, including the specific AIA needs that belonged to each category[5]. This categorization scheme formed the basis for the initial AIA Landscape (see Figure 1).

4) Iterative refinement of the AIA Landscape:

a. The initial AIA Landscape (see Figure 1) and Glossary (see Appendix D) were peer-reviewed by both the production team and AI SMEs, and their feedback incorporated.

b. These materials were pilot tested with Stakeholders to ensure they captured Stakeholder needs, and their feedback incorporated.

c. Other frameworks (i.e., those not identified as most comprehensive) were randomly sampled and used to cross-validate the categories, concepts, and definitions in the landscape. This continued until no further revisions to categories, concepts, definitions, or their organization were necessary.[6]

Table 1 provides the initial mapping between the assurance categories and specific AIA needs contained in the AIA Landscape assurance and the corresponding principles, characteristics, or functions in Fjeld et al. (2020) and the NIST AI RMF (Tabassi, 2023).[7] Table 2 provides a more cursory mapping to the remaining comprehensive frameworks.

Table 1. *Mapping of assurance categories and concepts from the AIA Landscape to Fjeld et al. (2020) and the NIST AI RMF (Tabassi, 2023).*

[4] In numerous instances, framework definitions were inconsistent with each other or with established definitions, or a definition was not provided in the original framework(s). In each instance, established sources were used to identify or generate an appropriate definition (see Appendix C and Appendix D).

[5] In a few instances, a specific AIA need (e.g., Transparent) appears in multiple AIA categories of the AIA Landscape (e.g., Interpretable, Private). Where this is the case, a general definition is provided in the Glossary for the specific need—which is applicable to all uses of this need, irrespective of category—along with a unique qualifier that provides the necessary context to understand its relation to the category in which it appears.

[6] Future versions of the AIA Landscape will include additional frameworks that were not available at the time of publication and will seek additional feedback from third parties that have developed their own assurance framework.

[7] In numerous instances, some assurance terms are mentioned in multiple themes or characteristics and discussed in different ways. Table 1 captures the main sections of the respective frameworks in which an AIA Landscape concept is discussed.

The AIA Landscape (MITRE)		AI Principles (Fjeld et al., 2020)	NIST AI RMF (Tabassi, 2023)
AIA Category	*Specific AIA Need*	*Theme*	*TAI Characteristic or Framework Core*[8]
Integrity Enabled		**Professional Responsibility; Security & Safety**	**Secure & Resilient; *Govern; Map***
	Interoperable		Valid & Reliable
	Reliable	Professional Responsibility	Valid & Reliable; *Map; Measure*
	Support Data Integrity & Quality	Fairness & Non-Discrimination; Privacy	Fair-Managed Bias; Secure & Resilient; Valid & Reliable; *Govern; Map*
	Support Model Integrity & Quality	Fairness & Non-Discrimination	Secure & Resilient; Valid & Reliable; *Govern; Map*
	Support Scientific & Engineering Integrity	Professional Responsibility	*Govern; Map*
	Sustainable	Accountability	*Measure; Manage*
	Transparent	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Privacy; Transparency & Explainability	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-Enhanced; *Govern; Measure; Manage*
Effective		**Accountability**	**Valid & Reliable; *Govern; Measure; Manage***
	Accurate	Professional Responsibility	Valid & Reliable
	Adaptive	Safety & Security	*Govern*
	Goal-driven	Fairness and Non-Discrimination; Professional Responsibility; Promotion of Human Values; Transparency & Explainability	Valid & Reliable; *Map*
	Reliable	Accountability; Safety & Security	Valid & Reliable; *Map; Measure*
	Resilient	Safety & Security	Secure & Resilient; *Measure*
	Robust	Privacy; Professional Responsibility	Valid & Reliable; Secure & Resilient; *Govern; Measure*
	Support Collaboration & Communication	Transparency & Explainability; Fairness & Non-Discrimination; Professional Responsibility	Explainable & Interpretable; *Govern; Map; Manage*

[8] Trustworthy AI characteristics are provided in normal font. Framework core categories are provided in italics.

The AIA Landscape (MITRE)		AI Principles (Fjeld et al., 2020)	NIST AI RMF (Tabassi, 2023)
AIA Category	*Specific AIA Need*	*Theme*	*TAI Characteristic or Framework Core*[8]
Secure		**Safety & Security; Privacy; Transparency & Explainability**	**Secure & Resilient; Privacy-Enhanced; *Map; Measure***
	Enable Availability	Accountability; Professional Responsibility Promotion of Human Values	Accountable & Transparent; Secure & Resilient; *Map; Measure*
	Enable Confidentiality	Safety & Security	Secure & Resilient; Privacy-Enhanced
	Enable Integrity	Professional Responsibility; Security & Safety	Secure & Resilient; *Govern; Map*
	Reduce Threats & Vulnerabilities	Accountability; Security & Safety	Secure & Resilient
	Resilient	Security & Safety	Secure & Resilient; *Map; Measure*
Governable		**International Human Rights[9]; Privacy; Professional Responsibility; Promotion of Human Values; Transparency & Explainability**	**Accountable & Transparent; *Govern; Map; Measure; Manage***
	Ensure Compliance	Privacy; Transparency & Explainability	*Govern*
	Legally Responsible	Accountability; Human Control of Technology; Privacy; Promotion of Human Values	Accountable & Transparent; *Govern; Map, Measure*
	Provide Oversight & Regulation	Accountability; Privacy; Safety & Security; Transparency & Explainability	*Govern; Map*
	Protect System Assets	Accountability; Privacy; Safety & Security; Fairness & Non-Discrimination; Promotion of Human Values	Secure & Resilient; *Govern; Map; Measure; Manage*
	Reduce Liability	Accountability; Safety & Security;	Accountable & Transparent
	Transparent	Accountability; Fairness & Non-Discrimination; Human Control of Technology;	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-

[9] Fjeld et al.'s International Human Rights theme did not emerge from their analysis of existing frameworks. It was included based on their preference only.

The AIA Landscape (MITRE)		AI Principles (Fjeld et al., 2020)	NIST AI RMF (Tabassi, 2023)
AIA Category	*Specific AIA Need*	*Theme*	*TAI Characteristic or Framework Core*[8]
		Privacy; Transparency & Explainability	Enhanced; *Govern; Measure; Manage*
Safe		**Accountability; Privacy; Safety & Security; Promotion of Human Values**	**Accountable & Transparent; Privacy-Enhanced; Safe; Secure & Resilient; *Govern; Measure***
	Reduce Harm	Accountability; Explainability & Professional Responsibility; Transparency; Privacy; Safety & Security	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-Enhanced; Safe; Valid & Reliable; *Map*
	Reliable	Accountability; Safety & Security	Valid & Reliable; *Measure*
	Resilient	Security & Safety	Secure & Resilient; *Map; Measure*
Accountable		**Accountability**	**Accountable & Transparent; *Govern***
	Auditable	Accountability	Valid & Reliable; *Govern, Map, Measure*
	Responsible	Accountability; Professional Responsibility	Safe; *Govern; Measure*
	Supports Redress & Feedback	Accountability; Human Control of Technology; Privacy	Accountable & Transparent; Privacy-Enhanced; *Measure; Map; Manage*
	Traceable	Accountability	Measure; Manage
	Transparent	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Privacy; Transparency & Explainability	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-Enhanced; *Govern; Measure; Manage*
Private		**Privacy**	**Privacy-Enhanced; *Map; Measure***
	Ensure Confidentiality	Privacy; Safety & Security	Privacy-Enhanced; Secure & Resilient
	Enable Consent	Privacy; Transparency & Explainability	Privacy
	Protect Data	Privacy; Safety & Security; Transparency & Explainability; Fairness &	Accountable & Transparent; Fair-Managed Bias; Secure & Resilient; *Govern*

The AIA Landscape (MITRE)		AI Principles (Fjeld et al., 2020)	NIST AI RMF (Tabassi, 2023)
AIA Category	*Specific AIA Need*	*Theme*	*TAI Characteristic or Framework Core*[8]
		Non-discrimination; Protection of Human Values	
	Transparent	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Privacy; Transparency & Explainability	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-Enhanced; *Govern; Measure; Manage*
Interpretable		**Transparency & Explainability**	**Explainable & Interpretable; *Measure***
	Clear	Accountability; Human Control of Technology; Security & Safety; Transparency & Explainability	Valid & Reliable; Safe; *Govern; Map*
	Comprehensible	Human Control of Technology; Promotion of Human Values; Transparency & Explainability	Accountable & Transparent; Explainable & Interpretable; *Govern; Map; Measure*
	Explainable	Transparency & Explainability	Explainable & Interpretable; Safe; *Measure*
	Justifiable	Privacy; Transparency & Explainability; Fairness & Non-discrimination	Accountable & Transparent; Explainable & Interpretable;
	Support Collaboration & Communication	Transparency & Explainability; Fairness & Non-Discrimination; Professional Responsibility	Explainable & Interpretable; Govern; *Map; Manage*
	Transparent	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Privacy; Transparency & Explainability	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-Enhanced; *Govern; Measure; Manage*
Equitable		**Fairness & Non-Discrimination**	**Fair-Managed Bias; *Govern***
	Accessible	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Privacy; Professional Responsibility; Safety & Security; Transparency & Explainability	Accountable & Transparent; Fair-Managed Bias; Secure & Resilient; *Govern*

The AIA Landscape (MITRE)		AI Principles (Fjeld et al., 2020)	NIST AI RMF (Tabassi, 2023)
AIA Category	*Specific AIA Need*	*Theme*	*TAI Characteristic or Framework Core*[8]
	Inclusive	Fairness & Non-Discrimination	Fair-Managed Bias; *Govern*
	Non-discriminatory	Accountability; Fairness & Non-Discrimination; Privacy; Transparency & Explainability; Promotion of Human Values	Fair-Managed Bias; *Govern*
	Participatory	Accountability; Fairness & Non-Discrimination; Professional Responsibility	Fair-Managed Bias; *Govern; Measure; Manage*
	Transparent	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Privacy; Transparency & Explainability	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-Enhanced; *Govern; Measure; Manage*
	Unbiased	Fairness & Non-Discrimination; Professional Responsibility; Transparency & Explainability;	Fair-Managed Bias; Privacy-Enhanced; *Measure*
Human Empowered		**Human Control of Technology; Promotion of Human Values**	**Valid & Reliable; *Govern; Map***
	Goal-driven	Fairness and Non-Discrimination; Professional Responsibility; Promotion of Human Values; Transparency & Explainability	Valid & Reliable; *Map; Measure*
	Responsive		
	Support Collaboration & Communication	Fairness & Non-Discrimination; Professional Responsibility; Transparency & Explainability	Explainable & Interpretable; *Govern; Map; Manage*
	Support Human Awareness	Accountability; Human Control of Technology; Privacy; Safety & Security; Transparency & Explainability	Accountable & Transparent; Fair-Managed Bias; Privacy-Enhanced; *Govern; Map*

The AIA Landscape (MITRE)		AI Principles (Fjeld et al., 2020)	NIST AI RMF (Tabassi, 2023)
AIA Category	*Specific AIA Need*	*Theme*	*TAI Characteristic or Framework Core*[8]
	Support Human Control	Human Control of Technology; Safety & Security	Safe; Valid & Reliable
	Support Human Judgment	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Promotion of Human Values; Privacy	Fair-Managed Bias; Privacy Enhanced; *Govern*
	Support Human Machine Teaming	Human Control of Technology; Transparency & Explainability	Valid & Reliable; *Govern; Map; Measure*
	Usable	Human Control of Technology; Professional Responsibility	Accountable & Transparent; *Govern; Map; Measure*
Civil		**Accountability; Fairness & Non-Discrimination; *International Human Rights;* Professional Responsibility; Promotion of Human Values**	**Accountable & Transparent; Explainable & Interpretable; Valid & Reliable; *Govern; Map; Measure; Manage***
	Enable Workforce Development	Privacy; Promotion of Human Values; Fairness & Non-discrimination; Professional Responsibility; Promotion of Human Values	*Govern*
	Environmental Responsibility	Accountability; Safety & Security; Promotion of Human Values	Safe; *Measure*
	Fair	Fairness & Non-Discrimination; Promotion of Human Values	Accountable & Transparent; Fair-Managed Bias; Privacy-Enhanced; *Measure*
	Goal-driven	Fairness & Non-Discrimination; Professional Responsibility; Promotion of Human Values; Transparency & Explainability	Valid & Reliable; *Map*
	Participatory	Accountability; Fairness & Non-Discrimination; Professional Responsibility	Fair-Managed Bias; *Govern; Measure; Manage*
	Promote Human Values, Rights, & Ethics	Accountability; International Human Rights;	Privacy-Enhanced; *Govern; Map; Measure; Manage*

The AIA Landscape (MITRE)		AI Principles (Fjeld et al., 2020)	NIST AI RMF (Tabassi, 2023)
AIA Category	*Specific AIA Need*	*Theme*	*TAI Characteristic or Framework Core*[8]
		Promotion of Human Values; Privacy	
	Reduce Mis/dis/mal-information	Human Control of Technology; Professional Responsibility; Transparency & Explainability	Accountable & Transparent
	Sustainable	Accountability; Promotion of Human Values	Measure
	Transparent	Accountability; Fairness & Non-Discrimination; Human Control of Technology; Privacy; Transparency & Explainability	Accountable & Transparent; Explainable & Interpretable; Fair-Managed Bias; Privacy-Enhanced; *Govern; Measure; Manage*

5 Rationale for the development of the AIA Landscape

More than 50 frameworks and over 500 reports on AI assurance have been published in recent years by various US and international government departments or agencies as well as by industry and not-for-profit and non-government organizations (Fjeld et al., 2022; Shneiderman, 2022). This is a substantive testament to Stakeholders' growing aspirations to develop trustworthy AI and a necessary step toward assuring AI-enabled systems within their organizations. Despite the continued production line of AIA frameworks there is a conspicuous absence of a standardized approach to AI assurance (cf. Robbins et al., 2024). In particular, there is a lack of a common AI Assurance scheme that would facilitate subsequent generalization and collaboration across technologies and Stakeholders, respectively.

The absence of standardization is evident in the perspective taken by each of the current AIA frameworks and reports, which is reflected in their specific purpose or application and the varied naming conventions of the assurance scheme they adopt.[10] While each framework scheme is similar and comprises a set of AIA categories (e.g., safety, security, governance), they do not converge on a common or standard assurance scheme (i.e., a common set of core assurance categories or common sub-sets of more specific assurance needs).[11] Instead, each scheme varies in the extent to which it captures the full spectrum of AIA needs that are evident across all frameworks. Example mapping between the AIA Landscape's assurance categories

[10] Assurance scheme naming conventions include but are not limited to: *Accountable AI, AI Assurance, AI Principles, AI Risk Management, AI Safety, Auditable AI, Ethical AI, Human-Centered AI, Responsible AI, Social AI, Transparent AI, Trusted AI, Trustworthy AI, and Universal Principles*.

[11] Fjeld et al. (2020) provided some evidence that frameworks were beginning to converge by showing that more recent frameworks covered all eight themes, whereas those produced earlier did not. However, in our review, we did not find consistent evidence for complete convergence.

and the schemes used in some of the more comprehensive AIA frameworks and reports is captured in Table 2.[12]

[12] In situations where a category is implied but not explicitly stated, the corresponding category is added in parentheses and gray text.

Table 2. *Approximate mapping between the AIA Landscape assurance categories and other comprehensive frameworks and reports.*

The AIA Landscape	AI Principles (Fjeld et al., 2020)[13]	NIST AI RMF (Tabassi, 2023)[14]	Accountability Framework (GAO, 2021)[15]	Ethical Principles/RAI Strategy (DIB, 2019; DOD, 2022; DSOD, 2021)[16]	AI Ethics & Safety Guide (Leslie, 2019)[17]	Ethics Principles/Framework ODNI (2020a, 2020b)[18]	Human-Centered AI (Shneiderman (2020, 2021, 2022)[19]	ELATE (Dorton & Stanley, 2023)[20]	RAI Toolkit: DAGR (CDAO, 2023)[21]	Executive Order 14110 White House (2023)	Executive Order 13960 White House (2020)
Integrity Enabled	*(Professional Responsibility)*	*(Valid & Reliable; Govern)*	Governance; Data	*(Responsible; Traceable)*	Sustainability; *(Fairness)*	Respect the Law & Act with Integrity; Informed by Science & Technology	General Virtues of the System; *(Reliable Systems/ Technical Practice)*	*(Continuous Improvement, Feedback Loops, Forensic Support)*			(Responsible & Traceable)

[13] Fjeld et al. (2020) included 8 AI Principles in their synthesis of popular AI assurance frameworks: Privacy, Accountability, Safety & Security, Transparency & Explainability, Fairness & Non-Discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values.
[14] The NIST AI RMF (Tabassi, 2023) Includes 7 characteristics of trustworthy AI: Valid & Reliable, Safe, Secure & Resilient, Accountable & Transparent, Explainable & Interpretable, Privacy-Enhanced, Fair—with Harmful Bias Managed, and 4 core functions: Govern, Map, Measure, and Manage AI.
[15] GAO (2021) includes 4 framework principles: Governance, Data, Performance, and Monitoring.
[16] The DoD included 5 ethical principles: Responsible, Equitable, Traceable, Reliable, and Governable (DIB, 2019; DOD, 2022; DSOD, 2021).
[17] The Turing Report (Leslie, 2019) includes 4 SUM values that underpin Responsible delivery of ethical AI: Respect, Connect, Care, Protect, 4 FAST principles that facilitate ethical design and use: Fairness, Accountability, Sustainability, Transparency, and a Governance framework (Leslie, 2019).
[18] ODNI (2020a,b) offers 6 principles of AI ethics: Respect the law and act with integrity, Transparent and accountable, objective and equitable, human-centered development and use, secure and resilient, informed by science and technology.
[19] Shneiderman (2020, 2021, 2022) outlines four sets of Human-Centered AI attributes: (1) General Virtues of the System (e.g., ethical, humane, benevolent, secure, private), (2) Performs Well in Practice (e.g., robust, reliable, available, adaptive, resilient, testable), (3) Provides Clarity to the Stakeholder (e.g., fair, transparent, interpretable, usable), (4) Enables Independent oversight (e.g., auditable, trackable, certifiable, compliant, redressable), and 4 levels of governance structures: (A) Reliable systems engineering/technical practices, (B) safety culture, organizational design and management strategies, (C) independent oversight and certification via external review, and (D) Government Regulation.
[20] The Evidence-based List of Exploratory Questions for AI Trust Engineering (ELATE) is a lightweight participatory toolkit to help AI development teams proactively identify how AI can go wrong "in the wild," and then mitigate such issues (Dorton & Stanley, 2023).
[21] DAGR utilizes both the NIST AI RMF 7 Trustworthy AI characteristics and the 5 DOD Ethical Principles (CDAO, 2023).

The AIA Landscape	AI Principles (Fjeld et al., 2020)[13]	NIST AI RMF (Tabassi, 2023)[14]	Accountability Framework (GAO, 2021)[15]	Ethical Principles/RAI Strategy (DIB, 2019; DOD, 2022; DSOD, 2021)[16]	AI Ethics & Safety Guide (Leslie, 2019)[17]	Ethics Principles/Framework ODNI (2020a, 2020b)[18]	Human-Centered AI (Shneiderman (2020, 2021, 2022)[19]	ELATE (Dorton & Stanley, 2023)[20]	RAI Toolkit: DAGR (CDAO, 2023)[21]	Executive Order 14110 White House (2023)	Executive Order 13960 White House (2020)
Effective	*(Professional Responsibility)*	*(Valid & Reliable; Manage)*	Governance; Data; Performance	Reliable	Sustainability; Transparency	Secure & Resilient	Performs Well in Practice *(Reliable Systems/ Technical Practice)*	Adaptability; Evolution; Reliability; Operations; Graceful Degradation	***DAGR** is a combination of:* *(i) NIST AI RMF's Trustworthy AI Characteristics* **+** *(ii) DOD's Ethical Principles)*		Purposeful & Performance-driven; Accurate, Reliable, & Effective
Secure	Security & Safety	Secure & Resilient	Data	Reliable	Sustainability; *(Care)*	Secure & Resilient	General Virtues of the System			Safety & Security	Safe, Secure, & Resilient
Governable	*(Accountable; Professional Responsibility)*	Govern	Governance	Governable	*(Accountability ; Respect, Care, Connect, Protect)*	*(Transparent & Accountable)*	Enables Independent Oversight; Government Regulation	Continuous Improvement, Feedback Loops,		*(Government & Industry Consensus Standards)*	Oversight by Responsible Agency & Officials
Safe	Security & Safety	Safe	Data	Reliable	Sustainability; Transparency; *(Care)*	*(Objective & Equitable)*	Safety Culture/Organizational Design *(Performs well in practice)*			Safety & Security	Safe, Secure, & Resilient
Accountable	Accountability	Accountable & Transparent	Governance; Monitoring; Performance	Traceable	Accountability; Transparency	Transparent & Accountable	Independent oversight; External Review/Certification	Continuous Improvement, Feedback Loops, Forensic Support		*(Government & Industry Consensus Standards)*	Accountable; Responsible & Traceable; Regularly Monitored

The AIA Landscape	AI Principles (Fjeld et al., 2020)[13]	NIST AI RMF (Tabassi, 2023)[14]	Accountability Framework (GAO, 2021)[15]	Ethical Principles/RAI Strategy (DIB, 2019; DOD, 2022; DSOD, 2021)[16]	AI Ethics & Safety Guide (Leslie, 2019)[17]	Ethics Principles/Framework ODNI (2020a, 2020b)[18]	Human-Centered AI (Shneiderman (2020, 2021, 2022)[19]	ELATE (Dorton & Stanley, 2023)[20]	RAI Toolkit: DAGR (CDAO, 2023)[21]	Executive Order 14110 White House (2023)	Executive Order 13960 White House (2020)
Private	Privacy	Privacy-Enhanced	Data	*(Equitable)*	Sustainability *(Fairness; Care)*	Respect the Law & Act with Integrity;	General Virtues of the System			Privacy	Transparent
Interpretable	Transparency & Explainability	Explainable & Interpretable	Performance	*(Reliable)*	Transparency	Transparent & Accountable	Provides Clarity to Stakeholders	Explainability			Understandable
Equitable	Fairness & Non-Discrimination	Fair-with Harmful Bias Managed	Governance; Data; Performance	Equitable	Fairness; Sustainability; Transparency; *(Care; Protect)*	Objective & Equitable	General Virtues of the System	Equity; Bias Awareness		Equity & Civil Rights	*(Lawful & Respectful of Our Nation's Values)*
Human Empowered	*(Human Control of Technology)*	*(Accountable & Transparent; Explainable & Interpretable)*	Governance; Performance	*(Responsible)*	*(Respect; Connect; Protect)*	*(Human-Centered Development & Use)*	Provides Clarity to Stakeholders; Independent Oversight	Third-party Acceptance, Human Approval; Stakeholder Participation during Development; Usability; Human Control; Risk		Consumer, Patient, & Student Advocacy; Worker Support; *(Innovation & Competition)*	

The AIA Landscape	AI Principles (Fjeld et al., 2020)[13]	NIST AI RMF (Tabassi, 2023)[14]	Accountability Framework (GAO, 2021)[15]	Ethical Principles/RAI Strategy (DIB, 2019; DOD, 2022; DSOD, 2021)[16]	AI Ethics & Safety Guide (Leslie, 2019)[17]	Ethics Principles/Framework ODNI (2020a, 2020b)[18]	Human-Centered AI (Shneiderman (2020, 2021, 2022)[19]	ELATE (Dorton & Stanley, 2023)[20]	RAI Toolkit: DAGR (CDAO, 2023)[21]	Executive Order 14110 White House (2023)	Executive Order 13960 White House (2020)
Civil	*(Promotion of Human Values)*	*(Fair-with Harmful Bias Managed)*	Governance; Performance	*(Equitable)*	Fairness; Sustainability; Transparency; *(Respect; Care; Connect; Protect; Sustainability)*	Human-Centered Development & Use	General Virtues of the System	Stakeholder Participation during Development		Equity & Civil Rights	Lawful & Respectful of Our Nation's Values

5.1 *Consequences of the lack of standardization*

While there are obvious reasons why assurance frameworks may lack standardization or commonality[22], the lack thereof can have consequential technological, mission-related, and/or societal impacts. For instance, non-standardized use of assurance needs or their definitions can result in different assurance requirements across frameworks and idiosyncratic recommendations about how best to assure against AI risks in any given use case. Similar terminological limitations can contribute to Stakeholders (e.g., technologists, domain specialists, organizations, and impacted communities) talking or working at cross-purposes[23], which can inhibit the ability of individuals, communities, and organizations to maintain resilience and, in turn, can increase the likelihood of complex system failure (Bradshaw, Hoffman, Johnson, & Woods, 2013; Johnson & Vera, 2019; Woods & Branlat, 2011). When the absence of a common lexicon is coupled with terminological inconsistency these effects can be compounded. Beyond well-established terms likely safety, security, and risk, we observed inconsistencies within and across existing AIA frameworks in the definitions and usage of assurance terms, with some terms being defined tautologically (i.e., with reference to other categories that, in turn, reference the original category). The need for a common vocabulary and consistency of use was recently highlighted as a key action limiting effective and sustained US-UK collaboration on trustworthy AI tools (Gunashekar, et al., 2024).

5.2 *Past efforts to create a comprehensive assurance scheme*

A handful of organizations have attempted to synthesize aspects of existing assurance frameworks to create a common and domain-agnostic view of assurance. For instance, in a review of 36 of the most popular AIA frameworks and reports published between 2016 and 2019, Fjeld et al. (2020) identified 47 different assurance 'principles.' These principles were categorized into eight assurance themes, which included *Privacy, Accountability, Safety & Security, Transparency & Explainability, Fairness & Non-Discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values*[24]*.*

Fjeld et al.'s synthesis is considered one of the most comprehensive assurance schemes available (Shneiderman, 2022). However, more technological elements of AIA did not emerge as key assurance themes in their analysis of existing frameworks. Instead, many technical assurance needs—ones that might be assumed to be covered by established processes, such as risk management, test & evaluation, and validation & verification—were subsumed under one or more of the ethically oriented themes (e.g., Accuracy and Scientific Integrity were included under Professional Responsibility; Verifiability and Replicability, and other elements of Governance, were included under Accountability).

[22] Examples of potential reasons for the lack of commonality or standardization across frameworks might include, for instance, internal requirements for consistency with an organization's established lexicon, or unique operational demands within a particular domain.

[23] Talking or working at cross purposes has been shown to be 'locally adaptive but globally maladaptive' (Woods & Branlat, 2011).

[24] Fjeld et al (2020) also assessed the extent to which the frameworks reviewed mentioned international human rights.

NIST's AI Risk Management Framework (AI RMF) (Tabassi, 2023) incorporates, at a high level, some of the more technical assurance functions omitted by Fjeld et al. (2020). These functions are included both as part of the AI RMF *core* (i.e., *Govern, Map, Measure*, and *Manage* AI risks) and as part of its *Trustworthy AI* characteristics (i.e., *Valid* & *Reliable*). Their Trustworthy characteristics include an additional six AI principles (i.e., *Safe, Secure & Resilient, Accountable & Transparent, Explainable & Interpretable, Privacy-Enhanced,* and *Fair—with Harmful Bias Managed*).

5.3 *Translation between existing frameworks*

There are several points of convergence across Fjeld et al. (2020) and Tabassi (2023). Each one includes, as a high-level assurance theme or characteristic, a reference to safety, security, privacy, accountability, transparency, explainability, and fairness. However, there are also substantial differences, making it difficult to align frameworks. For instance, Fjeld et al. (2020) combine *Safety & Security* into a single theme, whereas the Tabassi (2023) combines *Security & Resilience* and treats *Safety* separately. Similarly, whereas Fjeld et al. (2020) presents *Transparency & Explainability* as a combined theme and treats *Accountability separately*, Tabassi (2023) combines *Accountable & Transparent* into one category, and *Explainable & Interpretable* into another. In addition, three of Fjeld et al.'s (2020) primary themes (*Human Control of Technology, Professional Responsibility, and Promotion of Human Values*) are not included as a core element of, or as a *Trustworthy AI Characteristic* in NIST's AI RMF. Instead, these principles are captured as sub-needs in NIST AI RMF, appearing under multiple *core* and *Trustworthy AI Characteristic* categories, albeit in various guises. Add to this the absence of technical assurance needs represented in Fjeld et al.'s themes and it is not immediately obvious how to translate one set of assurance needs into the other.

While mapping is far from straight forward across frameworks, Table 2 presents an initial mapping between these and other popular AIA frameworks, created as part of our efforts to build the AIA Landscape

6 Summary

Despite a growing body of idiosyncratic AIA frameworks there is a notable absence of a common vocabulary and inconsistency of definitions and terminology use across frameworks. A standard assurance scheme does not yet exist—that captures the entire range of AIA needs evident in the collective set of published frameworks and reports—against which all AI-enabled systems can be benchmarked and the associated risks documented. Efforts to identify a common set of principles or create a common and domain-agnostic view of assurance have made positive strides in this regard, yet the commonality, consistency, and standardization issues—both within and across frameworks—remain. The absence of a common vocabulary is likely to impede framework adoption, alignment, and generalizability. It is also likely to inhibit individual and organizational collaboration and resilience in the face of AI risks. AI risks are known to *"emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed"* (Tabassi, 2023, p.1). The AIA Landscape is a first step

towards integrating frameworks to identify (a) a standard set of assurance needs and (b) a common set of definitions that can contribute to assuring that AI-enabled systems minimize the risks to the sociotechnical systems in which they are deployed.

Appendix A: References

[1] Akula, R. & Garibay, I. (2021). Audit and assurance of AI algorithms: A framework to ensure ethical algorithmic practices in artificial intelligence. *arXiv:2107.14046v1 [cs.CY] 14 Jul 2021.*

[2] Batarseh, F. A., Freeman, L., and Huang, C. H. (2021). A survey on artificial intelligence assurance. *Journal of Big Data, 8*(1), 1- 30.

[3] Bradshaw, J. M., Hoffman, R. R., Johnson, M. & Woods, D. D. (2013). The seven deadly myths of "autonomous systems." *IEEE Intelligent Systems, 28*(3), 54-61.

[4] Defence Science and Technology Laboratory (DSTL, 2021). *Assurance of artificial intelligence and autonomous systems: A Dstl biscuit book.* London, UK: Dstl. https://assets.publishing.service.gov.uk/media/61a765be8fa8f503764ed497/Assurance_of_AI_and_Autonomous_Systems_ONLINE_RGB_VERSION.pdf

[5] Freeman, L., Rahman, A., and Batarseh, F. A. (2021). Enabling artificial intelligence adoption through assurance. *Social Sciences, 10*(9), 322.

[6] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication,* 2020-1, http://dx.doi.org/10.2139/ssrn.3518482

[7] Gunashekar, S., van Soest, H., Qu, M., Politi, C., Chiara Aquilino, M., & Smith, G. (2024). *Examining the landscape of tools for trustworthy AI in the UK and the US: Current trends, future possibilities, and potential avenues for collaboration* (Rand Report # RR-A3194-1). Rand Corporation, https://doi.org/10.7249/RRA3194-1

[8] OECD (2024). Explanatory memorandum on the updated OECD definition of an AI system, *OECD Artificial Intelligence Papers*, No. 8, OECD Publishing, Paris, https://doi.org/10.1787/623da898-en.

[9] Robbins, D., Eris, O., Kapusta, A., Booker, L., & Ward, P. (2024). *AI assurance: A repeatable process for assuring AI-enabled systems.* MITRE White Paper.

[10] Shneiderman, B. (2022). *Human-centered AI.* Oxford, UK: Oxford University Press.

[11] Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST Trustworthy and Responsible AI. Gaithersburg, MD: National Institute of Standards and Technology.

[12] Ward, P., Stanley, J., Ferguson, R., Gladding, C. M. & Burns, K. J. (2023). *Risk Discovery Protocol for AI Assurance: Guidance for Administering the Protoco*l. MITRE Technical Report.

[13] Woods D.D. & Branlat, M. (2011). Basic patterns in how adaptive systems fail. In Erik Hollnagel, E., Pariès, J., & Wreathall, J. (Eds.), *Reslience Engineering in Practice: A Guidebook* (pp. 127-143). Aldershot, UK: Ashgate.

Appendix B: AIA Landscape Resources

The following list of resources was used to develop the AIA Landscape (for more detail, see Section 4: Development of the AIA Landscape). References in bold were considered amongst the most comprehensive frameworks and used as primary resources. The remaining references were randomly sampled until no further revisions to AIA Landscape categories, concepts, definitions, or their organization were necessary.

[14] Accenture (2020). *Responsible AI: A framework for building trust in your AI solutions.* https://www.accenture.com/us-en/insights/artificial-intelligence/responsible-ai-principles-practice

[15] Central Digital and Data Office (2020). *Data ethics framework*. London, UK: UK Cabinet Office. https://www.gov.uk/government/publications/data-ethics-framework

[16] Centre for Data Ethics and Innovation (CDEI, 2021). *The roadmap to an effective AI assurance ecosystem.* London, UK: CDEI. https://assets.publishing.service.gov.uk/media/61b0746b8fa8f50379269eb3/The_roadmap_to_an_effective_AI_assurance_ecosystem.pdf

[17] Chief Digital & Artificial Intelligence Office (2023). *Responsible AI Toolkit.* https://rai.tradewindai.com

[18] CISCO (2022). *The CISCO responsible AI framework: Security by design / human rights by design / privacy by design for personal data and consequential decisions.* https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-responsible-artificial-intelligence-framework.pdf

[19] Coalition for Health AI (2023). *Blueprint for trustworthy AI: Implementation guidance and assurance for healthcare* (v 1.0, April, 2023). The MITRE Corporation. https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

[20] **Defense Innovation Board (DIB, 2019). *AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense (Supporting Document).* Washington, DC: US Department of Defense.** https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

[21] **Department of Defense (DOD, 2022). *Responsible artificial intelligence strategy and implementation pathway.* Washington, DC: US DOD Responsible AI Working Council.** https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF

[22] **Deputy Secretary of Defense (2021). *Implementing Responsible Artificial Intelligence in the Department of Defense.* Memorandum for senior Pentagon leadership, Commanders of the Combatant Commands Defense Agency and DOD Field Activity Directors.** https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF

[23] **Dorton, S. L. & Stanley, J. C (2023). *Evidence-based list of exploratory questions for artificial intelligence trust engineering.* MITRE Technical Report (MTR230084).** https://www.mitre.org/sites/default/files/2024-04/PR-23-00018-2-evidence-based-exploratory-list-ai-trust-engineering-4-24.pdf

[24] Dunnmon, J., Goodman, B., Kirechu, P., Smith, C. & van Deusen, A. (2021). *Responsible AI guidelines in practice: Lessons learned from the DIU portfolio.* Defense Innovation Unit. https://assets.ctfassets.net/3nanhbfkr0pc/acoo1Fj5uungnGNPJ3QWy/6ec382b3b5a20ec7de6defdb33b04dcd/2021_RAI_Report.pdf

[25] Elam, M & Reich, R. (2022). *Stanford HAI Artificial Intelligence Bill of Rights: A white paper for Stanford's institute for human-centered artificial intelligence*, January, 2022. https://hai.stanford.edu/white-paper-stanford-hai-artificial-intelligence-bill-rights

[26] European Commission (EU, 2019). *Ethics guidelines for trustworthy AI.* Directorate-General for Communications Networks, Content and Technology, Publications Office, https://data.europa.eu/doi/10.2759/346720

[27] European Commission (EU, 2021). *Proposal for a regulation of the European parliament and of the council: Laying down harmonized rules on artificial intelligence* (Artificial Intelligence Act) and amending certain Union legislative acts. *COM(2021) 206 final; 2021/0106 (COD)*. Brussels, 21.04.2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[28] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication,* 2020-1, http://dx.doi.org/10.2139/ssrn.3518482

[29] G20 (2019). *G20 ministerial statement on trade and digital economy.* https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf

[30] General Accountability Office (GAO) (2021). *Artificial intelligence: An accountability framework for federal agencies and other entities.* GAO-21-519SP, June 2021. https://www.gao.gov/assets/gao-21-519sp.pdf

[31] Global Partnership on Artificial Intelligence (GPAI, 2021). *Responsible AI Working Group Report, GPAI Paris Summit* (November 2021). https://gpai.ai/projects/responsible-ai/

[32] Google (2020). AI principles: Objectives for building beneficial AI. https://ai.google/responsibility/principles/

[33] Hashimi, A. (2019). *AI ethics: The next big thing in government. Anticipating the impact of AI ethics within the public sector.* World Government Summit 2019. https://www.worldgovernmentsummit.org/docs/default-source/default-document-library/deloitte-wgs-report-en-lq.pdf?sfvrsn=1acfc90b_0

[34] Hawkins, R. Paterson, C., Picardi, C., Jia, Y., Calinescu, R. & Habli, I (2021). *Guidance on the assurance of machine learning in autonomous systems (AMLAS)* (version 1.1, March 2021). York, UK: Assuring Autonomy International Program, University of York. https://www.york.ac.uk/media/assuring-autonomy/documents/AMLASv1.1.pdf

[35] IBM (2022). *Everyday ethics for artificial intelligence.* https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

[36] IEEE (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (Version 1).* IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://ethicsinaction.ieee.org/wp-content/uploads/ead1e.pdf

[37] Information Commissioner's Office (ICO) (2020). *Guidance on the AI auditing framework. Draft guidance for consultation (Version 1.0; 20200214)*. London UK: ICO / UK

Government. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

[38] International Standard on Assurance Engagements (ISAE, 2013). *ISAE 3000 (Revised), Assurance engagements other than audits or reviews of historical financial information: International framework for assurance engagements and related conforming amendments* (Final Procurement, December, 2013). International Federation of Accountants / International Auditing and Assurance Standards Board. https://www.iaasb.org/publications/international-standard-assurance-engagements-isae-3000-revised-assurance-engagements-other-audits-or

[39] Japanese Society for Artificial Intelligence (JSAI, 2022). *Ethical Guidelines.* https://www.ai-gakkai.or.jp/ai-elsi/wp-content/uploads/sites/19/2017/05/JSAI-Ethical-Guidelines-1.pdf

[40] Jobin, A., Ienca, M., & Vayena, E. (2019). *The global landscape of AI ethics guidelines.* Nature Machine Intelligence volume 1, pages 389–399. https://www.nature.com/articles/s42256-019-0088-2

[41] Johnson M. & Vera, A. H. (2019). No AI is an island: The case for teaming intelligence, *AI Magazine, 40*(1), 16-28. https://doi.org/10.1609/aimag.v40i1.2842

[42] Kop, M. (2021). *EU Artificial Intelligence Act: The European approach to AI.* Transatlantic Antitrust and IPR Developments. https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/

[43] Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.* The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

[44] McDermott, P. Dominguez, C., Kasdaglis, N., Ryan, M., Trahan, I. & Nelson, A. (2018). *Human Machine Teaming Systems Engineering Guide (MITRE Technical Report MP180941)*. McLean, VA: The MITRE Corporation. https://www.mitre.org/sites/default/files/2021-11/prs-17-4208-human-machine-teaming-systems-engineering-guide.pdf

[45] Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics, 3*, 869-877. https://doi.org/10.1007/s43681-022-00209-w

[46] National Academies of Sciences, Engineering, and Medicine (NASEM) (2021). *Human-AI teaming: State of the art and research needs.* Washington, DC: The National Academies Press. https://nap.nationalacademies.org/catalog/26355/human-ai-teaming-state-of-the-art-and-research-needs

[47] National Security Commission on Artificial Intelligence (NSCAI) (2021). *Final report*, *2021*. https://www.nscai.gov/2021-final-report/

[48] National Security Commission on Artificial Intelligence (NSCAI) (2020). *Key considerations as a paradigm for responsible development and fielding of artificial intelligence (Line of effort on ethics and responsible AI: Quarter 2 report, July 22*). https://arxiv.org/pdf/2108.12289.pdf

[49] New South Wales Government (NSWG, 2022). Artificial intelligence assurance framework. https://www.digital.nsw.gov.au/sites/default/files/2022-09/nsw-government-assurance-framework.pdf

[50] OECD (2019). *Artificial intelligence in society*. OECD Publishing, Paris, France, 2019 https://doi.org/10.1787/eedfee77-en.

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

[52] OECD (2022). *Recommendation of the council on artificial intelligence OECD/LEGAL/0449*. http://legalinstruments.oecd.org

[53] **ODNI (2020a). *Artificial intelligence ethics framework for the intelligence community* (V1.0), June, 2020.** https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community

[54] **ODNI (2020b). *Principles of artificial intelligence ethics for the intelligence community.*** https://www.intelligence.gov/principles-of-artificial-intelligence-ethics-for-the-intelligence-community

[55] Porter, P., McAnally, M., Bieber, C., Wojton, H., & Medlin, R. (2020). *Trustworthy autonomy: A roadmap to assurance, Part I: System effectiveness.* Alexandria, VA: Institute for Defense Analysis. https://apps.dtic.mil/sti/trecms/pdf/AD1131283.pdf

[56] PWC (2019). *A practical guide to responsible artificial intelligence.* https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf

[57] Rotner, J., Hodge, R., & Danley, L. (2020). *AI fails and how we can learn from them.* McLean, VA: The MITRE Corporation. https://www.mitre.org/sites/default/files/2021-10/pr-21-2414-ai-fails-how-we-can-learn-from-them.pdf

[58] Select Committee on Artificial Intelligence, National Science & Technology Council (SCAI-NSTC)(2019). *The National artificial intelligence research and development strategic plan: 2019 Update.* Washington, DC: Executive Office of the President of the United States. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

[59] **Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction, 12*(3), 109-124.** https://aisel.aisnet.org/thci/vol12/iss3/1/

[60] **Shneiderman, B. (2021). Human-centered AI. *Issues in Science and Technology*, *37*(2), 56-61.** https://issues.org/wp-content/uploads/2021/01/56–61-Shneiderman-Human-Centered-AI-Winter-2021.pdf

[61] **Shneiderman, B. (2022). *Human-centered AI.* Oxford, UK: Oxford University Press.** https://global.oup.com/academic/product/human-centered-ai-9780192845290?cc=us&lang=en&

[62] Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N. et al. (2016). Grand challenges for HCI researchers. *ACM Interactions*, Sept-Oct, 24-25. DOI: 10.1145/2977645

[63] Stanton, B. & Jensen, T. (2021). Trust and artificial intelligence. Draft NISTIR 8332. National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8332-draft.pdf

[64] **Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI.* Gaithersburg, MD: National Institute of Standards and Technology.** https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[65] Tate, D. M. (2021). *Trust, trustworthiness, and assurance of AI and autonomy.* Alexandria, VA: Institute for Defense Analysis. https://apps.dtic.mil/sti/trecms/pdf/AD1150274.pdf

[66] Tate, D. M. (2021). *T&E of AI and autonomy: An assurance case framework, version 2.0.* Alexandria, VA: Institute for Defense Analysis. https://nps.edu/documents/115559645/122225231/2021+Dist+A+IDA+T%26E+of+AI+and+Autonomy+Jun+2021.pdf/0f8eb2ac-d4d6-2db2-d137-3efacbdf887e?t=1631725132862

[67] Topcu, U., Bliss, N., Cooke, N., Cummings, M., Llorens, A., Shrobe, H., & Zuck, L. (2020). *Assured autonomy: Path toward living with autonomous systems we can trust. Computing* Community Consortium Workshop. https://arxiv.org/pdf/2010.14443

[68] UK Government (2021). *Ethics, transparency and accountability framework for automated decision-making.* London, UK: Department for Science, Innovation and Technology, Centre for Data Ethics and Innovation, Cabinet Office and Office for Artificial Intelligence. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making

[69] UNESCO (2022). *Recommendation on the ethics of artificial intelligence.* https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

[70] Ward, P., Stanley, J., Dominguez, C. & McDermott, P. (2023). *A vision for assuring human-centered AI (MITRE Technical Report)*. McLean, VA: The MITRE Corporation.

[71] White House Executive Office of the President (2020). *Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government.* Washington, DC: US Government (E.O. 13960; Document: 85 FR 78939). https://www.govinfo.gov/content/pkg/FR-2020-12-08/pdf/2020-27065.pdf

[72] White House Executive Office of the President (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* Washington, DC: US Government (E.O. 14110; Document: 88 FR 75191). https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[73] White House Office of Science and Technology (2022). *Blueprint for an AI Bill of Rights: Making automated systems work for the American People.* Washington, DC: US Government. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[74] World Health Organization (WHO, 2021). *Ethics and governance of artificial intelligence for health. WHO Health Ethics and Governance Team.* https://www.who.int/publications/i/item/9789240029200

Appendix C: AIA Landscape Glossary Additional Resources

The following resources were used to define terms in the AIA Landscape Glossary. Note that some of the resources used to develop the landscape were also used to define the assurance categories and concepts and so are listed in both resource lists. In some instances, framework definitions were either inconsistent with each other or with established definitions, or a definition was not provided in the original framework(s). In these instances, these resources were used to identify or generate an appropriate definition (for specific sources, see Appendix C & Appendix D).

[75] Bardon, A., Bonotti, M., Zech, S. & Ridge, W. (2023). Disaggregating civility: politeness, public-mindedness, and their connection. *British Journal of Political Science, 53* (1), p. 308-325. https://doi.org/10.1017/S000712342100065X

[76] Department of Defense (2024). DOD Accessibility Statement. *DOD Section 508 Connecting Individuals with Information.* https://dodcio.defense.gov/DoDSection508/Std_Stmt.aspx

[77] European Union (EU, 2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.* https://eur-lex.europa.eu/eli/reg/2016/679/oj

[78] FDA (2016). *Data Integrity and Compliance with CGMP: Guidance for Industry.* Office of Communication, Outreach & Development. https://www.fda.gov/files/drugs/published/Data-Integrity-and-Compliance-With-Current-Good-Manufacturing-Practice-Guidance-for-Industry.pdf

[79] General Accountability Office (GAO) (2021). *Artificial intelligence: An accountability framework for federal agencies and other entities.* GAO-21-519SP, June 2021. https://www.gao.gov/assets/gao-21-519sp.pdf

[80] ISO (2015). Quality management systems – Fundamentals and vocabulary. *ISO 9000:2015.* https://www.iso.org/standard/45481.html

[81] ISO (2022a). Information technology—artificial intelligence—Overview of ethical and societal concerns*. ISO/IEC TR 24368:2022*. https://www.iso.org/standard/78507.html

[82] ISO (2022b). Trustworthiness – Vocabulary. ISO/IEC TS 5723:2022. https://www.iso.org/standard/81608.html

[83] Johnson, M., Bradshaw, J., Feltovich, P. J., Jonker, C. M., van Riemsdijk, B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity*. Human Robot-Interaction, 3*(1), 43-69. https://doi.org/10.5898/JHRI.3.1.Johnson

[84] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

[85] McDermott, P. Dominguez, C., Kasdaglis, N., Ryan, M., Trahan, I. & Nelson, A. (2018). *Human Machine Teaming Systems Engineering Guide (MITRE Technical Report MP180941)*. McLean, VA: The MITRE Corporation. https://www.mitre.org/sites/default/files/2021-11/prs-17-4208-human-machine-teaming-systems-engineering-guide.pdf

[86] MITRE (2024). Creating equity through data driven systems thinking. *MITRE Social Justice Platform.* https://sjp.mitre.org

[87] Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). *Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI.* https://apps.dtic.mil/sti/pdfs/AD1073994.pdf

[88] Nielsen, J. (2012). *Usability 101: Introduction to Usability.* Nielsen Norman Group. https://www.nngroup.com/articles/usability-101-introduction-to-usability/

[89] NIST (2010). Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). *NIST Special Publication 800-122*. Gaithersburg, MD: National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-122.pdf

[90] NIST (2017). An Introduction to Information Security. *NIST Special Publication 800-12 Rev. 1.* Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-12r1

[91] NIST (2020). *The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management (v 1.0).* Gaithersburg, MD: National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf

[92] NIST (2021). Four Principles of Explainable Artificial Intelligence. NISTIR 8312. Gaithersburg, MD: National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf

[93] NIST (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. *NIST Special Publication 1270.* Gaithersburg, MD: National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

[94] NIST (2024a). Glossary. *NIST Information Technology Laboratory (Computer Security Resource Center)*. Gaithersburg, MD: National Institute of Standards and Technology. https://csrc.nist.gov/glossary

[95] NIST (2024b). Interoperability. *NIST Information Technology Laboratory (Voting*), Gaithersburg, MD: National Institute of Standards and Technology. https://www.nist.gov/itl/voting/interoperability

[96] OECD (2019). *Artificial intelligence in society*. OECD Publishing, Paris, France, 2019 https://doi.org/10.1787/eedfee77-en.

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

[98] OECD (2022). *Recommendation of the council on artificial intelligence OECD/LEGAL/0449*. http://legalinstruments.oecd.org

[99] ODNI (2020a). *Artificial intelligence ethics framework for the intelligence community* (V1.0), June, 2020. https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community

[100] ODNI (2020b). *Principles of artificial intelligence ethics for the intelligence community.* https://www.intelligence.gov/principles-of-artificial-intelligence-ethics-for-the-intelligence-community

[101] Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction, 12*(3), 109-124. https://aisel.aisnet.org/thci/vol12/iss3/1/

[102] Shneiderman, B. (2021). Human-centered AI. *Issues in Science and Technology*, *37*(2), 56-61. https://issues.org/wp-content/uploads/2021/01/56–61-Shneiderman-Human-Centered-AI-Winter-2021.pdf

[103] Shneiderman, B. (2022). *Human-centered AI.* Oxford, UK: Oxford University Press. https://global.oup.com/academic/product/human-centered-ai-9780192845290?cc=us&lang=en&

[104] Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI.* Gaithersburg, MD: National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[105] Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2021). A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable. In D. Calvaresi, A. Najjar, M. Winikoff, & K. Främling (Eds.), *Explainable and Transparent AI and Multi-Agent Systems: Third International Workshop, EXTRAAMAS 2021* (pp. 119-138). (Part of the Lecture Notes in Computer Science book series; Vol. 12688). Springer. https://doi.org/10.1007/978-3-030-82017-6_8

[106] Ward, P., Gore, J., Hutton, R., Conway, G., & Hoffman, R. (2018). Adaptive skill as the *conditio sine qua non* of expertise. *Journal of Applied Research in Memory and Cognition, 7*(1), 35-50. https://doi.org/10.1016/j.jarmac.2018.01.009

[107] Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & Society, 36*, pp.585–595. https://link.springer.com/article/10.1007/s00146-020-01066-z#author-information

[108] Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *The* 2nd *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Lille, France. https://doi.org/10.48550/arXiv.1505.05723

Appendix D: AIA Landscape Glossary

AI Assurance Category	Specific AI Assurance Need	*Definition*	*Source*
Integrity Enabled		***Satisfies expectations for technical and scientific integrity.***	Developed by MITRE
Integrity Enabled	Interoperable	*Allows various devices to communicate or work together to perform common processes and achieve shared goals.*	Adapted from NIST (2024a) ITL-CSRC Glossary & NIST (2024b) ITL-Voting
Integrity Enabled	Reliable	*The capability to perform as required or on demand, without failure, for a given time interval, under expected conditions.* ***Qualifier:*** *Produces repeatable processes and reproduceable outcomes.*	Adapted from NIST AI RMF (Tabassi, 2023) and ISO (2015) 9000:2015
Integrity Enabled	Support Data Integrity & Quality	*Supports the ability to assess and maintain completeness, consistency, accuracy, reliability, representativeness, and quality of data and data sources throughout its lifecycle, and in storage, during processing, and while in transit.*	Adapted from FDA (2016) & NIST (2017) SP 800-12 r1
Integrity Enabled	Support Model/System Integrity	*Supports the ability to assess and maintain the soundness of a model or system's architecture, operations, and/or outcomes across its lifecycle, such that it performs as intended, unimpaired, and free from unauthorized manipulation.*	Adapted from Leslie (2019) & NIST (2017) SP 800-12 r1
Integrity Enabled	Support Scientific & Engineering Integrity	*Enables those who build and implement AI systems to be guided by established professional and scientific values and practices*	Adapted from Fjeld et al. (2020)
Integrity Enabled	Sustainable	*Processes are in place to ensure that the system can persist and be adapted over time to meet the needs of the communities in which it is deployed.* ***Qualifier:*** *Ensuring that data and system integrity are maintained over time.*	Adapted from NIST AI RMF (Tabassi, 2023) & ISO (2022a) ISO/IEC TR 24368:2022

Integrity Enabled	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use):* ***Qualifier:*** *Information needed to determine, test, and evaluate data, system/model, and scientific/engineering integrity is available as and when needed.*	Adapted from Fjeld et al. (2020)
Effective		***Achieves intent and/or desired outcomes.***	Developed by MITRE
Effective	Accurate	*The capability to maintain closeness of results of observations, computations, or estimates to the true values or the values accepted as being true.*	Adapted from ISO (2022b) ISO/IEC TS 5723:2022
Effective	Adaptive	*The capability to be responsive to change, including the ability to determine when current understanding, plans, or goals have deviated from expectations and/or the ability to achieve intent via alternative means.*	Adapted from Ward et al. (2018)
Effective	Goal-driven	*Supports the ability to achieve human goals, manage goal conflicts, and identify goal trade-offs and their respective impacts.* ***Qualifier:*** *Considers mission-relevant goals and is aligned with the organization's mission-relevant objectives in the context of risk tolerance levels and professional responsibility.*	Adapted from Fjeld et al. (2020)
Effective	Reliable	*The capability to perform as required or on demand, without failure, for a given time interval, under expected conditions.* ***Qualifier:*** *Consistently performs as expected.*	Adapted from NIST AI RMF (Tabassi, 2023) & ISO (2015) ISO/IEC 9000:2015
Effective	Resilient	*The capability to withstand perturbation (e.g., vulnerability, threat, unexpected event, or misuse) and return to normal function afterwards.* ***Qualifier:*** *Ability to stretch current capabilities and/or to degrade gracefully in a manner that permits normal function to continue.*	Adapted from NIST AI RMF (Tabassi, 2023)
Effective	Robust	*The capability of a system to maintain operations, performance, and/or expected impact under a variety of circumstances.*	Adapted from NIST AI RMF (Tabassi, 2023) & ISO

			(2022b) ISO/IEC TS 5723:2022
Effective	Support Collaboration & Communication	*Supports the ability of diverse agents to exchange information and work together.* ***Qualifier:*** *Supports interaction across and interdependencies between multiple internal and/or external entities to improve mission outcomes.*	Adapted from Johnson et al. (2014)
Secure		***Resistant to unauthorized activities***	Adapted from NIST AI RMF (Tabassi, 2023) & ISO (2022b) ISO/IEC TS 5723:2022
Secure	Ensure Availability	*The capability to ensure timely and reliable access to and use of information.*	Adapted from NIST (2017) SP 800-12 r1
Secure	Ensure Confidentiality	*The capability to preserve authorized restrictions on information access and disclosure.* ***Qualifier:*** *Including means for protecting proprietary information.*	Adapted from NIST (2017) SP 800-12 r1
Secure	Ensure Integrity	*The capability to guard against improper information modification or destruction and ensure information non-repudiation and authenticity.*	Adapted from NIST (2017) SP 800-12 r1
Secure	Reduce Threats & Vulnerabilities	*Incorporates protocols to avoid, protect, respond, and recover from system weaknesses and both adversarial and non-adversarial threats.*	Adapted from NIST AI RMF (Tabassi, 2023) & NIST (2017) SP 800-12 r1
Secure	Resilient	*The capability to withstand perturbation (e.g., vulnerability, threat, unexpected event, or misuse) and return to normal function afterwards:* ***Qualifier:*** *Ability to stretch current capabilities and/or to degrade gracefully in a manner that secures against and permits recovery from deliberate attacks, accidents, or naturally occurring threats or incidents.*	Adapted from NIST AI RMF (Tabassi, 2023) & NIST (2024a) ITL-CSRC Glossary

Governable		***Implements a framework of policies, rules, and processes for appropriate oversight within and across relevant organizations.***	Adapted from NIST 1270
Governable	Ensure Compliance	*Regulatory procedures are in place to prevent and address any divergence from standards and regulations.*	Developed by MITRE
Governable	Legally Responsible	*Regulatory procedures are in place to identify individuals or entities at fault for harm caused by the system or other legal breaches*	Adapted from Fjeld et al. (2020)
Governable	Provide Oversight & Regulation	*Regulatory procedures are in place to ensure that a diverse body of stakeholders identifies standards and regularly assesses system operations against them*	Adapted from Fjeld et al. (2020) & NIST AI RMF (Tabassi, 2023)
Governable	Protect System Assets	*Regulatory procedures are in place to identify parties responsible for guarding and overseeing internal and external system (including third-party) assets and components.*	Adapted from NIST AI RMF (Tabassi, 2023)
Governable	Reduce Liability	*The capability to assess potential failures to prepare and minimize the need for legal recourse and compensation, and permit insurability.*	Adapted from Shneiderman (2020, 2021, 2022)
Governable	Transparent	*The capability to make functions, operations and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier:*** *Information needed to oversee the system's operation, and for external parties to assess the oversight of the system, is available when needed.*	Adapted from Fjeld et al. (2020)
Safe		***Does not lead to a state in which human life, health, property, or the environment is endangered.***	Adapted from NIST AI RMF (Tabassi, 2023) & ISO (2022b) ISO/IEC TS 5723:2022
Safe	Reduce Harm	*Built and tested to prevent misuse and avoid unintended harms of all types.*	Adapted from Fjeld et al. (2020)

Safe	Reliable	*The capability to perform as required or on demand, without failure, for a given time interval, under expected conditions.* ***Qualifier:*** *Consistently minimizes the potential for harm.*	Adapted from NIST AI RMF (Tabassi, 2023) & ISO (2015) ISO 9000:2015
Safe	Resilient	*The capability to withstand perturbation (e.g., vulnerability, threat, unexpected event, or misuse) and return to normal function afterwards.* ***Qualifier:*** *Ability to stretch current capabilities and/or to degrade gracefully in a manner that maintains operations within acceptable levels of safety.*	Adapted from NIST AI RMF (Tabassi, 2023)
Accountable		***Answerable to the Stakeholders it empowers and to those it impacts for its proper and appropriate functioning, and obligated to address identified deficiencies.***	Adapted from OECD (2019, 2021, 2022)
Accountable	Auditable	*The capability to periodically document, review, and evaluate the AI solution, assess its impacts, and provide on-demand access to information needed to determine the extent to which specified requirements are fulfilled.*	Adapted from NIST AI RMF (Tabassi, 2023), Fjeld et al. (2020), & GAO (2021) GAO-21-519
Accountable	Responsible	*Decisions about AI system development and use are aligned with intended aims and values, and recognize the unique influence they exert on people and society.*	Adapted from ISO/IEC TR 24368:2022 & NIST AI RMF (Tabassi, 2023)
Accountable	Support Feedback & Redress	*Provide the opportunity for all Stakeholders, including individuals who are potentially impacted, to provide feedback, address concerns, and engage in procedures designed to change aspects of the system in ways that improve, rectify, repair, and/or remedy impacts (e.g., reporting problems, appealing system outcomes, and opt out of system processes).*	Adapted NIST AI RMF (Tabassi, 2023)
Accountable	Traceable	*Processes and outcomes can be monitored and traced back to simple root causes, or in complex situations, traced to potentially multiple and non-linear causes.*	Adapted from Shneiderman (2020, 2021, 2022)

Accountable	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier:*** *Stakeholders including impacted communities have appropriate access to information about values, choices, and intentions behind the system.*	Adapted from OECD (2019, 2021, 2022) & Walmsley (2021)
Private		***Safeguards information collection and use to preserve autonomy and dignity.***	Adapted from NIST (2020) Privacy Framework
Private	Enable Confidentiality	*The capability to restrict data access to protect personal privacy and proprietary information.* ***Qualifier:*** *Including means for protecting personal privacy.*	Adapted from NIST (2010) SP 800-122
Private	Enable Consent	*Individuals must explicitly agree to the processing of personally relevant data and be informed of risks and options.*	Adapted from NIST (2024a) ITL-CSRC Glossary & EU (2016)
Private	Protect Data	*All parts of the system lifecycle are designed to protect the rights of data subjects.*	Adapted from Fjeld et al. (2020)
Private	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier:*** *Stakeholders are made aware of data processing practices and associated risks, and any personally relevant information processed by the system.*	Adapted from NIST (2020) Privacy Framework & MITRE (2024) Social Justice Platform
Interpretable		***Makes processes and outputs apparent and meaningful in the context of functional and anticipated purposes***	Adapted NIST AI RMF (Tabassi, 2023)
Interpretable	Clear	*The system presents its processes and outputs such that the human can readily incorporate them into the workflow.*	Adapted from Leslie (2019)
Interpretable	Comprehensible	*The capability to provide users with access on-request to sufficient contextual information (e.g., system goals, objectives, inputs, assumptions, expected operating conditions, constraints) to allow them to develop a meaningful and up-to-date mental*	Adapted from NIST (2021) NIST.IR 8312/8637 & Verhagen et al. (2021).

		model of the system in a way that permits evaluation of its operations and outputs and/or anticipation of its behavior.	
Interpretable	Explainable	*The capability to provide a description of how or why an output was produced that captures the reasoning process(es) and/or technical mechanism(s) that actually led to the outcome, along with supporting evidence.*	Adapted from NIST (2021) NIST.IR 8312 & Mueller et al. (2019)
Interpretable	Justifiable	*The capability to provide an adequate reason (e.g., moral rationale) for producing a particular outcome that is capable of withstanding scrutiny, without necessarily providing a causal explanation.*	Adapted from Leslie (2019)
Interpretable	Support Collaboration & Communication	*Supports the ability of diverse agents to exchange information and work together.* ***Qualifier:*** *Supports building common ground vertically (across echelons) and horizontally (across units) to permit understanding of the 'bigger picture.'*	Developed by MITRE
	Transparent	*The capability to make explicit the functions, operations and outcomes (incl. data, algorithms, and models in use).* ***Qualifier:*** *Stakeholders have appropriate access to required information about the AI system's processes and outputs*	Adapted NIST AI RMF (Tabassi, 2023)
Equitable		***Addresses disparities in use and outcomes across individuals and groups.***	Adapted from MITRE (2024) Social Justice Platform
Equitable	Accessible	*Supports comparable ease of use and access across all users.*	Adapted from DOD (2024)
Equitable	Inclusive	*Processes and methods are included that consider the demographic diversity and diverse user experiences of those communities for whom the system is designed.*	Developed by MITRE
Equitable	Non-discriminatory	*Processes are in place to ensure that individuals and groups with similar non-protected characteristics are assigned similar*	Adapted from NIST (2024) ITL-CSRC Glossary & Zliobaite (2015).

		outputs; differences in protected characteristics should not cause significant differences in outputs.	
Equitable	Participatory	*Processes are in place to support engaging, across the entire AI lifecycle, with Stakeholders that represent a broad range of perspectives, including those from potentially impacted communities.* ***Qualifier:*** *Ensure marginalized communities are included to reduce inequity.*	Adapted Fjeld et al. (2020) & NIST AI RMF (Tabassi, 2023)
Equitable	Transparent	*The capability to make functions, operations, and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier****: Any discrepancies in treatment among individuals and groups are clearly communicated.*	Adapted from MITRE (2024) Social Justice Platform
Equitable	Unbiased	*Any systematic preference for or against some group of impacted people due to data or models is identified and mitigated as much as possible.*	Adapted from NIST (2024a) ITL-CSRC Glossary
Human Empowered		***Leverages human capabilities and enables pursuit of human goals.***	Adapted from Shneiderman (2020, 2021, 2022)
Human Empowered	Goal-driven	*Supports the ability to achieve human goals, manage goal conflicts, and identify goal trade-offs and their respective impacts.* ***Qualifier:*** *Considers the operator's goals in the context of broader operational, strategic, and societal goals.*	Adapted from Shneiderman (2020, 2021, 2022)
Human Empowered	Responsive	*The capability to promptly probe and obtain answers from and about the AI system, including its development, intentions, operations, outputs, and associated explanations.*	Adapted from Leslie (2019)
Human Empowered	Support Collaboration & Communication	*Supports the ability of diverse agents to exchange information and work together.* ***Qualifier:*** *Facilitates shared understanding and workflows among diverse stakeholders.*	Developed by MITRE

Human Empowered	Support Human Awareness	*Humans know when they are interacting with or are affected by AI, and know which tasks an AI is performing where they are out of the loop.*	Adapted from Fjeld et al. (2020)
Human Empowered	Support Human Control	*Humans can direct resources, activities, and priorities as needed and, where necessary, can modify or take over an AI's decisions or actions.*	Adapted from McDermott et al. (2018), Shneiderman (2020, 2021, 2022), & Johnson et al. (2014)
Human Empowered	Support Human Judgment	*Humans are engaged in an AI's decision process(es) throughout the AI lifecycle, and especially during operations.*	Adapted from ODNI (2020a, 2020b)
Human Empowered	Support Human Machine Teaming	*Adaptive, bi-directional team interaction among humans and machines that augments human capabilities for improved mission outcomes.*	Developed by MITRE
Human Empowered	Usable	*User interfaces are easy to use, efficient, memorable, learnable, and minimize and permit recovery from error, and are considered satisfactory by those who need to interact with them.*	Adapted from Nielsen (2012)
Civil		***Designed and operates in accordance with social norms and the public good***	Adapted from Bardon et al. (2023)
Civil	Enable Workforce Development	*Supports human jobs, economies, and AI workers, and their development, without putting them at risk.*	Adapted from OECD (2019, 2021, 2022)
Civil	Environmentally Responsible	*Actively protects or, at least, does not represent a threat to the environment and/or broader ecosystem.*	Adapted from Fjeld et al. (2020)
Civil	Fair	*The system benefits society as a whole and does not contribute to or perpetuate social imbalances.*	Adapted from Fjeld et al. (2020)
Civil	Goal-driven	*Supports the ability to achieve human goals, manage goal conflicts, and identify goal trade-offs and their respective impacts.* ***Qualifier:*** *Considers the goals of communities in which the system is deployed.*	Adapted from Shneiderman (2020, 2021, 2022)

Civil	Participatory	*Processes are in place to support engaging with Stakeholders across the entire AI lifecycle that represent a broad range of perspectives, including those from potentially impacted communities.* ***Qualifier:*** *Impacted communities play a key role in developing and sustaining the system.*	Developed by MITRE
Civil	Promote Human Values, Rights, & Ethics	*The system works in humanity's best interests; supports, observes, and does not conflict with commonly held human values, ethics, rights, and societal norms.*	Adapted from Fjeld et al. (2020) & Shneiderman (2020, 2021, 2022)
Civil	Reduce Mis/dis/mal-information	*The capability to manage context and content to reduce risk of manipulation and polarization of opinions and beliefs.*	Adapted from OECD (2019, 2021, 2022)
Civil	Sustainable	*Processes are implemented to ensure that the system can persist and be adapted over time to meet the needs of the communities in which it is deployed.* ***Qualifier:*** *Ensures that the system continues to be accepted over time by the communities in which it is deployed.*	Developed by MITRE
Civil	Transparent	*The capability to make functions, operations and outcomes explicit (incl. data, algorithms, and models in use).* ***Qualifier****: Documents and communicates to respective parties expected and actual impacts on communities*	Developed by MITRE