Policy Brief

A Common Language for Responsible AI

Evolving and Defining DOD Terms for Implementation

Author
Emelia S. Probasco

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

October 2022

Executive Summary

The deputy secretary of defense's memorandum entitled "Implementing Responsible Artificial Intelligence in the Department of Defense" articulates five ethical principles for artificial intelligence systems: responsible, equitable, traceable, reliable, and governable.[1] Those guiding principles have evolved the Department of Defense's thinking on responsible AI, but they are not sufficient for implementing responsible AI principles across everything from development to acquisition to operations. One foundational task toward implementing these guidelines, as laid out in the DOD memorandum "Responsible Artificial Intelligence Strategy and Implementation Pathway," is the standardization of language and definitions relating to the characteristics of responsible AI.[2]

Policymakers, engineers, researchers, program managers, and operators all need the bedrock of clear and well-defined terms that are appropriate to their particular tasks in developing and operationalizing responsible AI systems. Creating those standard terms and definitions requires input from all the communities involved in realizing responsible AI, both internal and external to the DOD. Thankfully, a community-defined taxonomy for responsible AI has already been started by the National Institute for Standards and Technology as a part of its draft AI Risk Management Framework (AI RMF), and the DOD could benefit by leveraging the work NIST has already done.

The rough alignment of NIST's trustworthy characteristics and the DOD's ethical principles provides a basis for both to work together and across industry to develop and deploy responsible AI. That broad consensus should also be the starting point for agreement on specific terms and definitions, where appropriate, that will help clearly guide developers, managers, and operators. By adopting or adapting to NIST's community-defined terms and definitions, the DOD stands to gain in three ways:

- first, by reducing misunderstandings that can lead to friction among the parties developing AI in the DOD, government, and industry;
- second, by having singularly focused and precise terms to guide developers, testers, policymakers, and operators in their efforts to develop, acquire, and operationalize responsible AI; and,
- third, by joining NIST and the rest of the U.S. government in projecting a consistent set of terms and definitions as norms of responsible AI are discussed and debated internationally.

This paper argues that the DOD could adopt or otherwise adapt to NIST's draft taxonomy as the standardized language for responsible AI in the DOD, excepting only two guiding principles which are truly unique to the DOD's context and mission: "responsible" and "traceable," as shown in Table 1.

Table 1: Summary of Term/Definition Evolution Recommendations

Keep DOD-unique terms and definitions	• Responsible • Traceable
Adapt DOD terms and definitions to NIST's terms and definitions	• Reliable → Reliability • Governable → Safe • Equitable → Managing Bias
Adopt terms and definitions from NIST	• Accuracy • Robustness • Transparency • Explainability • Interpretability • Privacy-Enhanced

Table of Contents

Executive Summary 1

Introduction 4

Background 6

Comparing the DOD's Terminology to NIST's Taxonomy 8

Characteristics Worth Adopting from NIST's Taxonomy 10

- Accuracy 10
- Robustness 11
- Safety 12
- Explainability and Interpretability 12
- Privacy-Enhanced 13

Characteristics That Can Be Adapted from NIST's Taxonomy 14

- Reliability vs. Reliable 14
- Managing Bias vs. Equitable 16

Characteristics Unique to the DOD 17

- Responsible 17
- Traceable 19

Key Takeaways 21

Conclusion 22

Author 23

Acknowledgments 23

Endnotes 24

Introduction

In May 2021, the deputy secretary of defense issued a memorandum entitled "Implementing Responsible Artificial Intelligence in the Department of Defense."[3] In it, she reinforced the five ethical principles the DOD had adopted in February 2020, which stated that all AI systems should be responsible, equitable, traceable, reliable, and governable. Those guiding principles have evolved the DOD's thinking on responsible AI, but they are not quite sufficient for implementing responsible AI principles across everything from development to acquisition to operations.

This June, the DOD took the next step toward more specific implementation guidance with the release of the "Responsible Artificial Intelligence Strategy and Implementation Pathway," a document that assigned specific lines of effort for evolving responsible AI guidelines to offices within the DOD.* The "Implementation Pathway" document is an important and exciting step forward because of its clear direction, but the list of LOEs makes clear the work still lying ahead.[4]

One foundational task for the DOD is the standardization of language and definitions relating to the characteristics of responsible AI. That standard language is needed by policymakers who are seeking to clearly communicate their expectations to developers and operators, as directed by LOE 5.1 and 5.2. It is also needed by program managers and engineers so that they might appropriately comply with policy directives through contracting, requirements setting, and risk management processes, as described in LOE 3.1. And it is even needed by policy and international leaders who are looking to establish commonly understood international norms around responsible AI for defense, as referenced in LOE 5.3.

Each of these different audiences—policymakers, engineers, researchers, program managers, and operators—needs the bedrock of not just a common language but also clear and well-defined terms that are appropriate to their particular tasks in developing and operationalizing responsible AI systems. For example, engineers and program managers in particular will need focused terms that avoid the commingling of multiple concepts as they go about negotiating technical tradeoffs in the development process, including setting contract requirements, or establishing risk management processes.

---

* Admittedly, there will be tradeoffs between characteristics when setting requirements and assessing risk, as NIST makes clear in its draft AI RMF, however, each characteristic must be individually understood and assessed before considering a potential tradeoff between characteristics.

Operators too will need terms they can consistently use as they go about learning the capabilities and limitations of new systems, implementing battle orders to control for operational risks, and standing up periodic maintenance regimes. And equally, policymakers have an interest in focused terms and definitions so that their guidance is followed and so that they can communicate clearly and consistently with international partners. A shared language and understanding will ultimately reduce—though not entirely eliminate—friction and uncertainty among this collection of professionals who must collaborate to realize a future with responsible AI for the DOD.

Choosing and defining the terms for responsible AI is not necessarily easy. It requires input from all the communities involved in realizing responsible AI, both internal and external to the DOD. Thankfully, a community-defined taxonomy for responsible AI has already been started by the Department of Commerce, where NIST has drafted a risk management framework along with a taxonomy of trust characteristics. The draft NIST "AI Risk Management Framework" defines and delineates key characteristics of responsible AI in terms that are equally useful to engineers and policymakers. The terms largely align with the DOD's five ethical principles of responsible AI but also improve upon the DOD's guidance by adding specificity that engineers, operators, and policy makers all need to do the work of implementing responsible AI.

Given the community engagement NIST has already done around its characteristics, as well as NIST's role in leading standards for the United States, this paper argues that the DOD could largely adopt NIST's draft taxonomy as the standardized language for responsible AI in the DOD, excepting only two guiding principles that are truly unique to the DOD's context and mission: "responsible" and "traceable."

Background

NIST first began drafting the AI Risk Management Framework in July 2021 as part of an extensive engagement plan that included academic, technical, and policy communities around the world.[5] The draft AI RMF's contributions are threefold: First, it moves the conversation from a more binary construct (is AI trustworthy or not?) to the more practical and systems engineering–based approach of risk management (is AI trustworthy enough, in "X" situation, if I do "Y"?). Similar to the DOD's recently articulated "Implementation Pathway," this context-dependent approach allows stakeholders to engage in creative and focused analyses on the likelihood and consequence of risks for a specific use, as well as the variety of technical and non-technical ways in which risks can be avoided or otherwise controlled either in the development or the deployment of a system.

Second, and importantly for the conversations ahead, the NIST framework contains a taxonomy of key characteristics that must be considered to achieve a level of "trustworthy" AI. In an early draft, NIST identified 12 distinct characteristics (see Table 3). In its August 2022 draft, those 12 distinct characteristics are still included with definitions, however several are listed under a single header, resulting in a total of seven more complex and less singularly focused characteristics (see Table 2). While the change in the August draft highlights the complex relationships between some of the characteristics, all 12 component characteristics and their definitions remain in the document to guide a standard language.

Third, the draft AI RMF helps to set the foundation for U.S. government-led approaches to the establishment of trustworthy AI across industries and even internationally. Albeit a non-regulatory agency, since 1901 NIST has successfully marshaled industry standards through the work of its technical staff and its industry engagement processes. For the AI RMF, it has been engaging with industry, academia, and international governments across multiple events and written exchanges over the course of a planned 18-month-long public dialogue. The consensus NIST shepherds will shape technical and policy conversations within the United States and internationally.

NIST's intention to engage in the international conversation around AI ethics and norms is demonstrated in part by the mapping of its characteristics to three major international AI policy documents: the Organization of Economic Co-operation and Development (OECD)'s Recommendation of the Council on Artificial Intelligence, the

proposed European Union AI Act, and Executive Order 13960.[6] As displayed in NIST's table (Table 2) the characteristics or guidelines from each policy document roughly correspond to NIST's taxonomy, though the alignment of precise definitions for each characteristic is still evolving along with the international conversations about trustworthy and responsible AI.

Table 2: NIST Mapping of AI RMF Taxonomy to AI Policy Documents

AI RMF	OECD AI Recommendation	EU AI Act (Proposed)	EO 13960
Valid and reliable	Robustness	Technical robustness	Purposeful and performance driven Accurate, reliable, and effective Regularly monitored
Safe	Safety	Safety	Safe
Fair and bias is managed	Human-centered values and fairness	Non-discrimination Diversity and fairness Data governance	Lawful and respectful of our Nation's values
Secure and resilient	Security	Security & resilience	Secure and resilient
Transparent and accountable	Transparency and responsible disclosure Accountability	Transparency Accountability Human agency and oversight	Transparent Accountable Lawful and respectful of our Nation's values Responsible and traceable Regularly monitored
Explainable and interpretable	Explainability		Understandable by subject matter experts, users, and others, as appropriate
Privacy-enhanced	Human values; Respect for human rights	Privacy Data governance	Lawful and respectful of our Nation's values

Source: National Institute for Standards and Technology, "AI Risk Management Framework: Second Draft," August 18, 2022.

While still in its draft form, NIST has announced plans to finalize the AI RMF toward the end of 2022. Given NIST's technical expertise, broad engagement, and institutional mandate, the DOD should consider now how it might adopt or otherwise adapt to the coming NIST guidance. Since the DOD's AI ethical principles are largely reflected in NIST's draft already, adopting its taxonomy—with just a few militarily relevant exceptions—could enable more efficient development and deployment of responsible AI.

Comparing the DOD's Terminology to NIST's Taxonomy

Like the EU AI Act or the OECD AI recommendations referenced in Table 2, the DOD's five ethical principles from the deputy secretary's 2021 memorandum may also be mapped to the NIST taxonomy (Table 3).

Table 3: DOD AI Ethical Principles Mapped to NIST AI RMF Taxonomy

DOD AI Ethical Principles	NIST AI RMF Draft Taxonomy

• Reliable
• Governable
• Equitable
• Responsible
• Traceable
	• Valid and Reliable (includes:
	  • accuracy,
	  • reliability, and
	  • robustness)
	• Safe
	• Fair—and Bias is Managed (includes:
	  • fairness and
	  • managing bias)
	• Secure and Resilient (includes:
	  • security and
	  • resilience)
	• Transparent and Accountable (includes:
	  • transparency, and
	  • accountability)
	• Explainable and Interpretable (includes:
	  • explainability and
	  • interpretability)
	• Privacy-Enhanced

Note: Solid lines indicate explicit connection, dashed lines indicate implicit connection.
Source: Kathleen Hicks, Implementing Responsible Artificial Intelligence in the Department of Defense, and DOD Responsible AI Working Council, "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway"

The rough alignment of NIST's trustworthy characteristics and the DOD's ethical principles provides a basis for both to work together and across industry to develop and deploy responsible AI. That broad consensus should also be the starting point for agreement on specific terms and definitions, where appropriate, that will help clearly

guide developers, managers, and operators as they do the work of creating responsible AI products.

As will be discussed below, there is much the DOD could adopt or otherwise adapt from NIST's trustworthy taxonomy. For example, there are at least five important NIST characteristics—accuracy, robustness, interpretability, explainability, and privacy—that, while made explicit by NIST, are at best implicit in the DOD's current guidance. Additionally, complex terms, such as the DOD's "reliable" and "governable," could be covered by several singular terms embedded within NIST's taxonomy, and NIST's capture of the issue of "managing bias" could be a small improvement to the DOD's "equitable" term and definition.

Two terms, however, are notably different for the DOD and must be preserved as unique to its standard language: "responsible" and "traceable." Given the military's unique norms, values, and culture, these words and their definitions are both distinct and important to realizing responsible AI for the DOD.

Characteristics Worth Adopting from NIST's Taxonomy

There are several key terms NIST emphasizes in the AI RMF which are only implied, or even absent, from DOD guidance to date. Two of the implied but not explicit terms in the DOD's guidance, "security" and "resilience," may not bear separate mention given the DOD's long-standing focus on these issues. Though some may argue that AI systems are different enough to warrant special attention to security and resilience, the NIST term and definition align well enough with the DOD's existing processes that it will not be discussed here.* Other terms, such as accuracy, robustness, safety, interpretability, and privacy, should be considered for explicit inclusion in AI-specific policy documents and standard language so that policymakers, operators, and engineers are well aligned on unique expectations for AI-enabled systems.

Accuracy

"Closeness of results of observations, computations, or estimates to the true values or the values accepted as being true"

-ISO/IEC TS 5723:2022, quoted in NIST AI RMF Second Draft

In the enthusiasm to adopt AI systems to accelerate operations or close gaps in staffing, recent cases have demonstrated that the accuracy of those systems sometimes escapes close scrutiny prior to deployment. For example, fielded facial recognition systems have falsely identified innocent citizens as criminals, AI-enabled background check systems have incorrectly screened out individuals in need of public services, and medical image screening systems enabled by computer vision have not demonstrated a clear advantage over human radiologists in detecting cancer.[7] While the DOD's acquisition and requirements review process is thorough and focused on "effectiveness," there is a tension and tradeoff that can exist between the accuracy of a system and its speed of response or potential to compress the observe-orient-decide-act (OODA) loop. That tradeoff will be highly context dependent, based on the capabilities of the system (e.g., is it a lethal system or logistics recommendation system with a 1 percent error rate?) as well as the situation in which that system is placed (e.g., is it a logistics system for

* For example, NIST's terms for security and resilience are closely related to DOD guidance on system survivability key performance parameters (KPPs) in the Joint Capabilities Integration and Development System Manual.

peacetime base support operations or a logistics system for combat operations?). Calling out accuracy, as NIST does within its "valid and reliable" characteristic, will help DOD engineers and program managers work with operators to ensure that the proper balance is struck for a responsible AI-enabled system. Additionally, the term and definition would lend greater clarity to a component of the DOD's "governable" principle, which includes a requirement that AI systems "fulfill their intended functions."

Robustness

"Ability of an AI system to maintain its level of performance under a variety of circumstances"

-ISO/IEC TS 5723:2022, quoted in NIST AI RMF Second Draft

A term missing from DOD guidance but elevated by NIST is the concept of "robustness." While robustness might be considered one of the "-ilities" in systems engineering, along with reliability, availability, and maintainability for example, it is not common in DOD parlance.[8] Robustness has a special significance for AI-enabled systems, as these systems are still widely acknowledged as being "brittle." In other words, while some AI systems may be efficient under controlled conditions with known inputs, they fail to generalize or otherwise adapt to uncontrollable factors outside an initial set of narrow assumptions.[9] Given the military's core saying that "no plan survives first contact," let alone contact with an enemy explicitly focused on exploiting the vulnerabilities of AI systems, it would seem particularly important that DOD AI-enabled systems are developed with an eye to the appropriate level of robustness required not just for real-world operations in general but also combat operations in particular. Adopting the robustness term and definition would help the DOD focus on this critical issue for its AI-enabled systems and may even inspire ideas for exploiting vulnerabilities in adversary AI-enabled systems.

Safety

"...approaches for AI safety often relate to rigorous simulation and in-domain testing, real-time monitoring, and the ability to shut down or modify systems that deviate from intended or expected functionality."

-NIST AI RMF Second Draft

Safety is normally an important concern for the DOD, and there are numerous standards and requirements around the concept of safety, especially when it comes to the safety of lethal systems. There is a special concern when it comes to AI, however, that the DOD makes clear in its definition of "governable," and which NIST includes in its definition of "safety," in that an AI system operator must have, according to the DOD's ethical principles, "the ability to disengage or deactivate deployed systems that demonstrate unintended behavior." Given the overlap between NIST's definition of "safety" and the DOD's concern for an ability to deactivate errant systems, it may benefit the DOD to adopt NIST's expanded notion of safety for AI systems and eliminate "governable," the concepts of which are addressed by NIST's "safe" and "accuracy" terms.

Explainability and Interpretability

"Explainability refers to a representation of the mechanisms underlying an algorithm's operation, whereas interpretability refers to the meaning of AI systems' output in the context of its designed functional purpose."

-NIST AI RMF Second Draft

NIST makes a subtle but important distinction about "interpretability," which receives only brief mention in the 2021 DOD memo and Defense Innovation Unit's (DIU) Responsible AI Guidelines.[10] In NIST's definition, the key difference is between explainability's "representation of the mechanisms underlying an algorithm's operation" and interpretability's role in communicating the meaning of its output. These two concepts serve two different purposes and often two different audiences. For example, a junior operator in the field must be able to properly interpret the recommendation of an AI system through a human-machine

interface, often under unique stress. An explanation of how the AI arrived at its recommendation at that point in time may be interesting but unhelpful to the operator in the heat of the moment. By contrast, decision support systems in less time-sensitive instances might warrant more explainability so that leaders can question embedded assumptions or biases in algorithmic recommendations.

The need for interpretability and explainability differs depending on the audience and the context for action. Both should be duly considered by developers and emphasized as standard language and requirements for the DOD's AI-enabled systems.

Privacy-Enhanced

"Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation)."

-NIST AI RMF Second Draft

While privacy is covered in the NIST AI RMF, as well as the DIU Responsible AI Guidelines, it is notably absent from the topline DOD memo. Privacy is important to the defense department for many reasons, beyond just security concerns. As major data losses like the 2014 and 2015 U.S. Office of Personnel Management breaches demonstrate, the government will be held to public account or subject to litigation for the loss or theft of Personally Identifiable Information (PII).[11] More than that, however, without proper management, data-hungry AI developers who do not study privacy issues or do not understand certain nuances may inadvertently overlook or underestimate the importance of privacy as a core value in liberal democracies and an important element of America's example to the world. Given the importance of privacy to safeguarding the PII of public servants, and because privacy is a key differentiator for liberal democracies, enumerating and codifying privacy as a differentiator for DOD systems is both sensible and strategic.

Characteristics That Can Be Adapted from NIST's Taxonomy

The NIST AI RMF and DOD terminology currently share four similar words but ascribe slightly different definitions: "reliability" and "reliable" as well as "fair and bias is managed" and "equitable." In both of these instances, NIST's definition provides an advantage because of its clarity but may also be useful because of NIST's role in rallying technology standards.

Reliability vs. Reliable

Term	NIST definition	DOD definition
NIST: Reliability The DOD: Reliable	"ability of an item to perform as required, without failure, for a given time interval, under given conditions" (from ISO/IEC TS 5723:2022)	"… explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across AI capabilities' entire life-cycle"

To NIST, "reliability" is a question of consistency. When tested, an AI-enabled system should produce a statistical error rate as a value that remains consistent over a period of use. For example, over the course of thousands of image classifications, the performance of a land-based mobile missile carrier recognition algorithm will correctly recognize mobile missile carriers 99.99 percent of the time. "Reliability" means that, for example, this required 99.99 percent correct recognition rate will not deteriorate over time such that it would eventually only correctly identify a mobile missile carrier 98 percent of the time (and the other 2 percent of the time it mistakes a missile carrier for non-combatant vehicle such as a passenger bus). Any amount of drift in the performance of a system over time can potentially pose a serious threat to military operations and, as a consequence, reliability is frequently an "-ility" used by systems engineers developing DOD systems.

By adopting "reliable" as an umbrella term that includes "safety, security, and effectiveness," the DOD faces two issues: first, that it misses the opportunity to identify the importance of consistent performance over time and second, that it combines safety, security, and effectiveness (or better yet, accuracy, as discussed above) when, in fact, these are three separate ideas that must be verified individually by engineers

during the development process.[12] Additionally, researchers, developers, and program managers will inevitably have to break out reliability from safety, security, and effectiveness in their risk management processes as well as system documentation and training for operators.

Each of the characteristics nested under the DOD's "reliable" are assessed in different ways and for different purposes. For example, effectiveness is meant to reflect the accuracy essential to a useful AI system (e.g., if the missile does not strike or even damage the target, it is worse than useless), but reliability adds a time component that ensures the system will be effective consistently, not just at the beginning or periodically. This is very different from a requirement for security, which may involve both cyber and machine learning expertise to protect the system from hacking or adversarial attack, for example; and different yet from safety, which may concern how the system shall behave should it lose communications, for example.

Policymakers and AI-system engineers should have an equal interest in delineating these unique requirements and debating the appropriate standards, testing protocols, and tradeoffs for each. These delineations will also be important to communicating with the research community, which has distinct lines of investigation into reliability, safety, and security standards and techniques for AI.

NIST offers a more succinct and clear definition of reliability, while the DOD's definition, albeit inclusive, is more difficult to implement consistently across a diverse set of responsible AI developers and operators. As such, the DOD would be best served to adopt NIST's "reliability" into its standard language and guidance, and satisfy the other components of "reliable" through "safety" and "security" terms and requirements.

Managing Bias vs. Equitable

Term	NIST	The DOD
Fair–and Bias is Managed/Equitable	"Fairness in AI includes concerns for equality and equity by addressing issues such as bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Systems in which biases are mitigated are not necessarily fair."	"The Department will take deliberate steps to minimize unintended bias in AI capabilities."

A nuance worth noting in the NIST framework is its emphasis on managing bias, versus the DOD's stated intent to be "equitable" and "minimize unintended bias." NIST's inclusion of the concept of management, as well as its categorization of distinct categories of bias (systemic, computational, and human) indicates an ongoing need to monitor and adjust to the influence of bias over time. The DOD's "equitable" may "minimize unintended bias," but perhaps only at the point of inception. The DOD could benefit from having a requirement that results in not just a system that manages bias at inception but one that enables the monitoring and management of bias throughout a system's life cycle. For example, a system could be correctly biased toward recognizing a certain type of woodland camouflage in one region of the world but that bias might need to be managed if the system is later deployed to a different region of the world that has different camouflage patterns. This could be especially important when such management needs to happen in the field—far away from engineers who are able to monitor and adjust for bias.

Characteristics Unique to the DOD

While migrating to NIST's terms and definitions in the above cases will help to satisfy the task in the DOD's responsible AI "Implementation Pathway," and help the DOD communicate better internally and with external partners, there are two important terms which are unique to the DOD: "responsible" and "traceable." Somewhat confusingly, these DOD terms are also linguistically similar to NIST terms, but there are crucial differences in the definitions that bear special focus by DOD leaders and partners. Maintaining the different words and different definitions for the DOD is important.

Responsible

Term	Definition
NIST: *Accountability*	"... expectations of the responsible party in the event that a risky outcome is realized. The shared responsibility of all AI actors should be considered when seeking to hold actors accountable for the outcomes of AI systems."
The DOD: *Responsible*	"DOD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities."

Of all the term comparisons, NIST's "accountability" and the DOD's "responsible" is the most challenging for three reasons: first, because the words are generally used in tandem; second, because "responsible AI" is an overarching term of art being used across the public and private sector to encompass all the key characteristics of trustworthy AI (NIST and others often use the phrase "trustworthy and responsible AI"); and third, because the DOD has a critically important and culturally unique understanding of both words. The DOD's definition of "responsible" could be considered more expansive than NIST's definition of "accountable" because it includes the mandate to exercise "appropriate levels of judgment and care, while remaining responsible for the development, deployment and use of AI." The notions of "judgment and care" and the specific reference to "deployment" are reminders of the operational situations in which AI may be employed by the DOD and in particular by military commanders.

The DOD definition implies that an individual's judgment when developing, deploying, and/or using an AI system matters. This standard speaks to the military's culture of command and delegation—that commanders bear responsibility for both their action and inaction in the absence of specific guidance from higher authority. As an example, a commander is responsible for their judgment to deploy an AI-enabled autonomous vehicle where such a system could inadvertently lead to conflict escalation, even if the system does not actually take any aggressive action. The "responsible" definition also implies an expectation of good judgment extends not just up the operational chain of command (in other words, to the commanders who might deploy a system) but also to the administrative chain of command, which includes those individuals managing the technical development of the AI system. The DOD is rightly holding developers and operators to a high standard given the nature of military operations, something it can do because it has authority over both developers and users.

The definitional differences on this point are critically important for DOD audiences—especially operational and administrative leaders—who need and want to clearly understand their responsibilities when it comes to AI-enabled systems. But to be "held responsible for" something implies accountability. In other words, there must be a clear record of the chain of events that has led to an outcome to hold a responsible party accountable for a given outcome. Accordingly, the DOD may wish to include "accountability" as a term for reliable AI systems, but make clear that an "accountability" requirement triggers a need for record keeping. What that record should include is outside the scope of this analysis, but deserves further investigation and debate.

Traceable

Term	Definition
NIST: *Transparency*	"… Transparency reflects the extent to which information is available to individuals about an AI system, if they are interacting – or even aware that they are interacting – with such a system…."
The DOD: *Traceable*	"The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of technology, development processes, and operational methods applicable to AI capabilities, including transparent and auditable methodologies, data sources, and design procedure and documentation."

The DOD's "traceable" definition includes NIST's word "transparent," but the DOD's "traceable" goes much further and the two terms should not be confused or conflated. Key to the DOD "traceable" guidance is that both AI development and deployment should ensure that "relevant personnel possess an appropriate understanding of technology, development processes, and operational methods applicable to AI capabilities." Since NIST's role is to advise developers and inform the public, they are unable to go as far as the DOD can in directing that users possess a level of understanding not just of the technology but of its development processes and operational methods.

The DOD has set a complex and high bar with its definition of traceable. Simplifying the term by breaking out several component parts as separate characteristics, as NIST did, might give AI-systems developers more concrete guidance on which to take action. For example, transparency is clearly an element of the DOD's "traceable" and could be separated as a key characteristic or requirement. Auditability or accountability (see above) may be another component characteristic that could be extracted and directly addressed. Finally, interpretability is another NIST component that could be drawn out to at least partially address the requirement to ensure "appropriate understanding" (also discussed above).

Even after separating transparency, auditability/accountability, and interpretability as component requirements for "traceable," there is still an important notion left to be addressed: the burden of "appropriate understanding" placed on the "relevant

personnel." Put another way, which service members will be qualified through what process before they can take an action using an AI-enabled system? This burden is mostly an issue of people, not of technology, which falls to policymakers responsible for developing operational concepts, combat orders, personnel policies, and, most importantly, training. Early connections between policymakers responsible for operational governance and engineers responsible for designing AI systems might help ensure the proper understanding is reached. Regardless, policymakers must further consider and define expectations for both "appropriate understanding" and vesting authority in "relevant personnel."

Key Takeaways

As the DOD sets out on its "Implementation Pathway," the department needs a clear set of well-defined and commonly understood terms to form the standard language for responsible AI for policymakers, engineers, program managers, and operators. Those groups of individuals will need the standard language to efficiently and effectively take the next steps in the development, acquisition, and eventual operation of responsible AI.

To meet this need, the DOD would benefit from largely adopting or adapting to NIST's evolving AI RMF taxonomy and terms, with the exception of two guiding principles unique to the DOD—responsible and traceable—which should be kept in the DOD's specific list of characteristics for responsible AI systems (see Table 1, reprinted below). Since these terms are unique to the DOD's mission and culture, policymakers, operators, and developers should make extra effort to ensure they are well understood. Altogether, these terms and the characteristics they describe could become the basis for new standard "-ilities" in the systems engineering and acquisition processes when it comes to requirements setting and risk management. Addressing these characteristics in the requirements process will not guarantee responsible AI systems, but it will force the development, training, and operational communities to reckon with issues and tradeoffs of particular concern to AI-enabled systems.

Table 1: Summary of Term/Definition Evolution Recommendations

Keep DOD-unique terms and definitions	• Responsible • Traceable
Adapt DOD terms and definitions to NIST's terms and definitions	• Reliable → Reliability • Governable → Safe • Equitable → Managing Bias
Adopt terms and definitions from NIST	• Accuracy • Robustness • Transparency • Explainability • Interpretability • Privacy-Enhanced

While adopting NIST's draft taxonomy and terms is most directly useful to the task of creating standard language for RFIs and RFPs laid out in LOE 3.1 of the "Implementation Pathway," it will also benefit efforts to coordinate across government (LOE 5.1 and 5.2) and engage in international norm development (LOE 5.3). By following NIST's lead in term selection and definition, the DOD will have addressed its own pressing need while also efficiently aligning the department with commercial sector partners and other parts of the government.

Finally, to build on former NASA Administrator Robert Frosch's point that "engineering is an art, not a technique," even the best-defined taxonomy cannot prevent disaster if requirements generation or risk management processes are treated as simple box-checking exercises. Program managers, developers, and users must all think creatively about how and when to accept, avoid, mitigate, or transfer potential risks against the backdrop of emerging societal norms as they collectively design and employ AI-enabled systems. Following the conversation NIST is leading on risk management is a necessary but insufficient first step.[13]

Conclusion

The DOD has repeatedly stated its commitment to creating responsible AI and being an international leader in AI ethics when it comes to national security. NIST's draft AI Risk Management Framework and the taxonomy and terms it has developed lend more specificity to the conversations the DOD cares about regarding the development of trustworthy or responsible AI. In most cases, the DOD could adopt NIST's terms and definitions to guide discussions of responsible AI. For the terms and concepts unique to the DOD's mission, "responsible" and "traceable," these two terms should be highlighted by policymakers to ensure the particularities for the DOD are understood so that they might properly guide development and implementation of new systems. Making a change now could help the DOD avoid confusion later internally and when working with external partners to create responsible AI-enabled systems.

Author

Emelia S. Probasco is a senior fellow at the Center for Security and Emerging Technology.

Acknowledgments

The author is especially grateful to Rita Konaev for her comprehensive reviews.

For feedback and assistance, the author would like to thank Ron Luman, Melissa Flagg, Igor Mikolic-Torreira, Dewey Murdick, Lynne Weil, Mina Narayanan, Husan Chahal, Heather Frase, John Bansemer, Matt Mahoney, Jess Shao, and Alex Friedland. The author is solely responsible for all errors.

Document Identifier: doi: 10.51593/20220028

Endnotes

[1] Kathleen Hicks, *Implementing Responsible Artificial Intelligence in the Department of Defense*, May 26, 2021, https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF.

[2] DOD Responsible AI Working Council, "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway," June 2022, https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF.

[3] Kathleen Hicks, *Implementing Responsible Artificial Intelligence in the Department of Defense*.

[4] DOD Responsible AI Working Council, "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway."

[5] National Institute of Standards and Technology, *AI Risk Management Framework: Second Draft*, August 18, 2022, https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

[6] "Recommendation of the Council on Artificial Intelligence," OECD Legal Instruments, May 21, 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; "The Artificial Intelligence Act," The AI Act, November 19, 2021, https://artificialintelligenceact.eu/; Executive Office of the President, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," Federal Register, December 8, 2020, https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.

[7] For a larger list of examples and a review of accuracy concerns with AI see Inioluwa Deborah Raji et al., "The Fallacy of AI Functionality," *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, (June 2022): 959–972, https://doi.org/10.1145/3531146.3533158; Karoline Freeman et al., "Use of Artificial Intelligence for Image Analysis in Breast Cancer Screening Programmes: Systematic Review of Test Accuracy," *BMJ* 2021; 374:n1872, September 2, 2021, https://doi.org/10.1136/bmj.n1872.

[8] For more about the relevance of "-ilities" in the systems engineering requirements and risk management process see parts 1 and 2 of Bill Kobran, "Considering the (Possib)-ilities (Part 1)," DAU Blog, July 7, 2020, https://www.dau.edu/training/career-development/logistics/blog/Considering-the-Possib-ilities-Part-1.

[9] Andrew J. Lohn, "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance," arXiv preprint arXiv:2009.00802 (2020), https://arxiv.org/abs/2009.00802.

[10] Jared Dunnmon et al., "Responsible AI Guidelines in Practice," Defense Innovation Unit, November 2021, https://www.diu.mil/responsible-ai-guidelines.

[11] John H. Jones, "OPM Hack Class Action Plaintiffs Win Initial Approval for $63M Payout," *FedScoop,* June 8, 2022, https://www.fedscoop.com/opm-hack-class-action-plaintiffs-win-initial-approval-for-63m-payout/.

[12] DOD Responsible AI Working Council, "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway," June 2022, https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF.

[13] Robert A. Frosch, "A Classic Look at Systems Engineering," in *Readings in Systems Engineering* (NASA, 1993): 1–7, https://spacese.spacegrant.org/Additional%20Readings/NASA%20Readings%20in%20SE.pdf, quoted in Michael D. Griffin, "How Do We Fix Systems Engineering?," *61st International Astronautical Congress Prague,* September 27–October 1, 2010, https://www.nasa.gov/sites/default/files/atoms/files/3_griffin_how_do_we_fix_systems_engineering.pdf.