The Department of Veterans Affairs Establishes a Trustworthy AI Framework

National Artificial Intelligence Institute

VA | U.S. Department of Veterans Affairs

Executive Summary

The increasing capabilities of AI pose new risks and vulnerabilities for organizations and decision makers. Several trustworthy AI frameworks have been created by U.S. federal agencies and international organizations to outline the principles to which AI systems must adhere for their use to be considered responsible.

Different trustworthy AI frameworks reflect the priorities and perspectives of different stakeholders, and no single framework is currently considered definitive. Taken together, however, these frameworks can offer a holistic perspective on trustworthy AI values, allowing federal agencies to create agency-specific trustworthy AI strategies that account for unique institutional needs and priorities.

The Department of Veterans Affairs (VA) is the largest civilian agency and has the largest integrated healthcare system in the country. It has established several big data repositories, including the largest genomic knowledge base in the world linked to healthcare information. It also trains the largest number of nurses and doctors in the United States. Given these factors, VA is uniquely positioned to advance AI research, development, and the population at-large. Nonetheless, that development and use must be informed by its mission to provide high-quality care and services to Veterans. The purpose of this paper is to present a set of guiding principles that will provide the foundation for VA to design, develop, acquire, and use AI systems in a manner that fosters Veteran trust and confidence in AI systems while meeting the requirements of established laws and regulations. Such a framework will ensure that VA continues to deliver superior services to Veterans by leveraging emerging technologies while adhering to the highest ethical standards, including protecting their privacy and civil rights.

I. Introduction

Background and Context

Artificial intelligence (AI) models and applications have become faster, more accurate, and better able to solve problems that are costly, complex, time-consuming, or otherwise prohibitive for humans. Such performance gains have led to implementation of AI tools in nearly every professional domain with positive effects on productivity and well-being.[1, 2]

For example, AI has the potential to significantly change the health care and benefits landscape by improving outcomes and increasing the productivity and efficiency of service delivery. It can reduce administrative and other burdens, allowing staff to spend more time directly assisting Veterans and potentially raising staff morale and retention. Additionally, AI can help get life-

[1] Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2021. The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13(1), 333-72.

[2] Makridis, Christos A., and Saurabh Mishra. 2022. Artificial Intelligence as a Service, Economic Growth, and Well-Being. Journal of Services Research, 25(4)

saving treatments to market faster. Finally, by automating specific tasks and analyzing large amounts of data, AI can help healthcare systems better manage the demands of an aging population and changing patient expectations.

Already, AI models can match or outperform physicians at diagnosing colorectal cancer,[3] mesothelioma,[4] and lung cancer.[5] One high-profile AI tool has been shown to reduce sepsis-related mortality (which is responsible for over 250,000 deaths each year in the U.S.) by 20% by identifying risks before the condition is diagnosed using current standards of care.[6] Another predicts over 90% of acute kidney injury cases (a condition that affects nearly 20% of inpatients in the U.S.) that require dialysis, allowing clinicians to initiate potentially life-saving treatment earlier than would be possible using current methods.[7] An AI tool trained on CT scans can correctly identify intra-cranial hemorrhaging with over 95% accuracy, decreasing clinical turnaround time by over 40%.[8]

These examples demonstrate the life-saving potential of AI tools to come as future models, data sets, and computational capabilities continue to improve, particularly in the areas of personalized and precision medicine, imaging analysis, surgical assistance, natural language processing, operational and administrative optimization, and drug discovery.

While there have been many perilous predictions of AI on the labor force highlighted in the media (e.g., as in the case of radiologists displaced

[3] Zhou, D., Tian, F., Tian, X. et al. Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. Nat Commun 11, 2961 (2020). https://doi.org/10.1038/s41467-020-16777-6

[4] Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat. Med. 25, 1519–1525 (2019)

[5] Huang P, Lin CT, Li Y, et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. Lancet Digit Health. 2019;1(7):e353-e362. doi:10.1016/S2589-7500(19)30159-1

[6] Henry, K.E., Adams, R., Parent, C. et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 28, 1447–1454 (2022). https://doi.org/10.1038/s41591-022-01895-z

[7] Tomašev, N., Glorot, X., Rae, J.W. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 572, 116–119 (2019). https://doi.org/10.1038/s41586-019-1390-1

[8] Wismüller, A. and Stockmaster, L. "A prospective randomized clinical trial for measuring radiology study reporting time on Artificial Intelligence-based detection of intracranial hemorrhage in emergent care head CT", Proc. SPIE 11317, Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, 113170M (28 February 2020); https://doi.org/10.1117/12.2552400

by machine learning),[9] coupled with other others that have been very optimistic,[10] a more balanced view is that AI has the potential to streamline, automate, and augment human labor in health care[11] and benefit delivery[12] but the design and implementation of AI tools must be carefully managed to realize these gains.

The Need for Trustworthy AI Principles

The increasing capabilities of AI pose new risks and vulnerabilities for organizations and decision makers. While poor judgment and miscalculation have always had negative consequences, the prospect of AI-driven systems substantially changes the scale of those consequences. In addition, the black box nature of many modern AI tools introduces additional needs to ensure that such applications are circumscribed by oversight and accountability systems that mitigate risks.

In one high-profile clinical example, an AI model performed reliably in a controlled training setting but failed to detect sepsis in 67% of patients in a

[9] R. Washington, "Why scan-reading artificial intelligence is bad news for radiologists," The Economist, 2017.

[10] A. Spatharou, S. Hieronimus, and J. Jenkins. "Transforming health care with AI: The impact on the workforce and organizations," McKinsey. Deloitte. "The socio-economic impact of AI on European health systems," 2020.

[11] E. Brynjolfsson and T. Mitchell, "What can machine learning do? Workforce implications," Science, vol. 358, no. 6370, pp. 1530-1534, 2017. A. Alabdulkareem, M. R. Frank, L. Sun, B. AlShebli, C. Hidalgo, and I. Rahwan, "Unpacking the polarization of workplace skills," Science Advances, vol. 4, no. 7, 2018. D. Acemoglu, and P. Restrepo. "Automation and New Tasks: How Technology Displaces and Reinstates Labor," Journal of Economic Perspectives, vol. 33, no. 2, 2019.

[12] "VA decreases mail processing time for claims intake." VA Office of Public and Intergovernmental Affairs, Aug. 31, 2020. VA decreases mail processing time for claims intake

hospital setting, leaving them vulnerable to serious health complications.[13] In other contexts, AI image recognition tools have exhibited differential performance based on skin color,[14] which would present problems if performance was similarly biased in health care settings. High-profile reporting has also flagged potential racial bias in algorithmic criminal sentencing and predictive law enforcement contexts,[15, 16] highlighting how, without attentive design, validation, monitoring, and oversight, the use of AI may pose threats to health, well-being, and civil liberties, perpetuating and exacerbating existing inequalities and inefficiencies. Irresponsible use of AI systems may in turn undermine trust in such technologies and introduce barriers to the development and adoption of beneficial tools.

Systems that rigorously assess and mitigate the unique risks associated with AI are sometimes referred to as "trustworthy," as their design and implementation are intended to satisfy the highest possible standards of protection for those affected by their use.

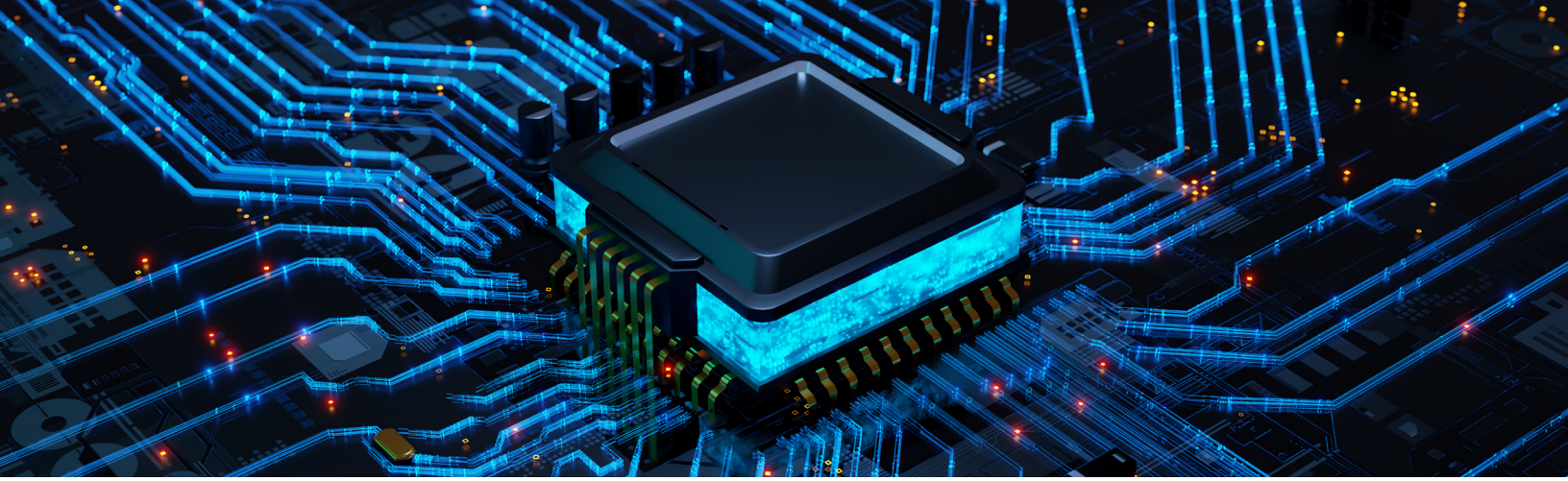
In this vein, several trustworthy AI frameworks have been created by U.S. federal agencies and international organizations to outline the principles that AI systems must adhere to for their use to be considered responsible. For example, Executive Order 13960: *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* states that "the ongoing adoption and acceptance of AI will depend significantly on public trust. Agencies must therefore design, develop, acquire, and use AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values." As such, it requires federal agencies to "design, develop, acquire, and use AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable laws..."

[13] Habib, A. R., Lin, A. L., and Grant, R. W. 2022. The Epic Sepsis Model Falls Short—The Importance of External Validation. JAMA Intern Med. 2021;181(8):1040-1041.

[14] Buolamwini, J. and Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of the 1st Conference on Fairness, Accountability and Transparency in Proceedings of Machine Learning Research, 81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html.

[15] "Machine Bias." ProPublica, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[16] "Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?" Smithsonian Magazine, March 5, 2018. https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/



Different trustworthy AI frameworks reflect the priorities and perspectives of different stakeholders, and no single framework is currently considered definitive. Taken together, however, these frameworks can offer a holistic perspective on trustworthy AI values.

The trustworthy AI frameworks most relevant to the mission and operations of VA include the following:

Executive Order 13960: This order establishes AI use and transparency requirements across Federal agencies. It lays out nine different trustworthy AI principles to which federal AI systems must conform but leaves the development of detailed compliance standards to other federal bodies, including agencies themselves.

VA Data Ethics Framework:[17] The VA Data Ethics Framework is a VA-specific set of requirements to ensure that use of Veteran data is safe, fair, and effective. The use of AI tools at VA will often, if not always, involve access to or production of Veteran-related data, so these standards are highly relevant to trustworthy AI activities.

Blueprint for an AI Bill of Rights: The White House Blueprint for an AI Bill of Rights is a nonbinding document released by the White House Office of Science and Technology Policy (OSTP) to guide the responsible use of AI in the United States. It lays out five principles focused on protecting the safety and civil liberties of those potentially affected by automated decisions.

[17] Also known as the VA Data Ethics Final Rule, see 38 CFR 0.605.

Executive Orders 13985 and 14091: Though not trustworthy AI frameworks per se, E.O. 13985: *Advancing Racial Equity and Support for Underserved Communities Through the Federal Government* directs agencies to embed fairness in decision-making processes, ensuring that programs and policies do not impose and perpetuate barriers to opportunities and benefits for historically underserved groups and E.O. 14091: *Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government* requires that "(w)hen designing, developing, acquiring, and using artificial intelligence and automated systems in the Federal Government, agencies shall do so, consistent with applicable law, in a manner that advances equity."

NIST AI Risk Management Framework: The National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF) describes a risk-based approach to developing, implementing, and overseeing AI systems. It lays out seven trustworthy AI principles across four core functions: Govern, Map, Measure, and Manage. The NIST AI RMF sources of risk across the AI lifecycle to support organizations in developing low-risk AI systems.

OECD AI Principles: The Organisation for Economic Co-operation and Development (OECD) AI Principles are a set of recommendations adopted by its member nations, which includes the United States, and several non-member signatories to ensure that the use of AI is trustworthy and respects human-centered and democratic values.

GAO AI Accountability Framework: The U.S. Government Accountability Office (GAO) AI Accountability Framework is a series of AI implementation guidelines. The document is organized differently from other frameworks; instead of a series of principles, it is a discussion of those principles at each step of implementation.

DoD AI Ethical Principles: The Department of Defense (DoD) AI Ethical Principles steer the design, development, procurement, and deployment of Responsible AI systems in the context of national defense.

While the existence of so many frameworks reflects the high degree of interest in trustworthy AI, the lack of uniformity between frameworks makes it difficult for practitioners to take theory to practice. To provide some clarity and structure to these varying frameworks, we compare frameworks according to similarities in the language they use to describe trustworthy AI principles, with the goal of constructing a VA-specific trustworthy AI framework.

Motivation for an Agency-Specific Trustworthy AI Framework

Executive Order 13960 directs Federal agencies to "design, develop, acquire, and use AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable law and the goals of Executive Order 13859 *(Maintaining American Leadership in Artificial Intelligence)*." Specifically, the order requires agencies to annually compile a complete inventory of their non-sensitive and non-classified AI use cases and ensure that all use cases are consistent with the nine trustworthy AI principles laid out in Section 3 of the order. The plans for achieving consistency must be approved by the VA's designated Responsible AI Official (RAIO), as authorized by Section 8 of the order, in coordination with the Data Governance Council (DGC) AI Working Group (AIWG) and other senior VA leaders.

As such, the proposed VA-specific trustworthy AI framework serves as:

1. A reference document for ensuring VA satisfies E.O. 13960 consistency requirements and strives to include other trustworthy AI frameworks impacting or informing VA's mission;

2. A foundation for implementation activities to ensure consistency with E.O. 13960 Section 8 as coordinated by the VA RAIO and VA Data Governance Council; and

3. An agency-wide consensus statement on VA's trustworthy AI values.

To fulfill E.O. 13960 requirements, VA needs a trustworthy AI framework tailored to the breadth and scale of its activities, the sensitivity of the data it handles, and the agency's responsibility to serve the needs of Veterans. A VA-specific trustworthy AI framework will position the agency to monitor AI activities effectively across VA as capabilities and circumstances evolve.

The principles outlined here can also inform guidance, as developed in a future VA Trustworthy AI Playbook, that is consistent with VA operations and strategic priorities.

Building VA-wide consensus on trustworthy AI principles will ensure consistency in standards across VA activities. In demonstrating its commitment to using AI in a responsible, transparent, effective, and equitable manner, VA can enhance its reputation and credibility as an organization. By designing and adopting an agency-specific trustworthy AI framework, VA also positions itself as a continued leader in trustworthy AI among federal agencies.

In the sections that follow, we discuss VA AI needs in the context of relevant trustworthy AI frameworks and align principles from those frameworks to inform a proposed trustworthy AI framework tailored to VA's mission and AI strategy. As other agencies possess different needs, priorities, and requirements, each framework will have different contents and emphases, but agencies may still benefit from adhering to a similar process and structure as described here.

VA AI Requirements and Priorities

As a Federal agency, VA is required to adhere to E.O. 13960 and the VA Data Ethics Framework, and the principles from these frameworks feature prominently in the proposed framework described in the following sections. However, our goal is not simply to ensure that VA trustworthy AI activities meet minimum requirements, nor is it to prioritize particular frameworks over others, but rather to build a framework that reflects the highest standards present in existing frameworks and provide a foundation for future practical guidance to practitioners on how to meet these standards.

First and foremost, VA use of AI tools must purposefully serve Veteran needs. Similar to the OECD AI Principles' human-centered approach to AI, the proposed VA Trustworthy AI Framework stresses the dignity and autonomy of Veterans by emphasizing that Veterans will not be exposed to undue risk or exploited as a convenient population for data-gathering purposes. Similarly, the VA framework aims to treat Veterans with the highest level of respect by providing transparency surrounding the use of AI systems. All frameworks examined here include principles related to transparency, illustrating its centrality to trustworthiness.

In addition to fulfilling E.O. requirements, VA must protect the privacy and security of Veteran data. The Veteran data that VA handles is particularly sensitive, including personally identifiable health information. High standards must be observed in the storage and use of this data, and these processes must be secure and accountable. Information on data use must be readily available in an easily accessible and understandable fashion. The VA Data Ethics Framework establishes requirements to ensure that use of Veteran data is safe, fair, and effective. The development and application of AI tools at VA will often, if not always, involve access to or production of Veteran-related data, so these standards will often be applicable to trustworthy AI activities.

The proposed VA Trustworthy AI Framework will not supersede or replace existing VA data use or research requirements. It will complement them to accommodate the specific concerns entailed by the use of AI technologies. It assumes that these technologies will also be in compliance with existing VA policies: that they will comply with the Data Ethics Framework, will, when appropriate, undergo IRB review for approval, and will obtain informed consent as appropriate.

Because of existing disparities in access to health care along racial, sex/gender, socioeconomic, age, ability, geographic, and other lines, risks related to bias and fairness are also particularly relevant to VA AI activities. Without careful attention to data provenance and AI system design, AI systems can perpetuate existing inequalities by generating predictions and recommendations that reflect underlying discriminatory processes and outcomes. In addition, underrepresentation of certain characteristics in datasets can lead to systematically skewed predictions or recommendations from AI-driven processes, which can lead to unfair and harmful outcomes. Bias and algorithmic discrimination are specifically addressed in several existing trustworthy AI frameworks, particularly in the Blueprint for an AI Bill of Rights. The proposed framework's bias management principle will guide bias identification and mitigation for the specific circumstances of VA AI systems and ensure that VA's use of AI is fair.

Finally, VA strives for excellence across its operations in order to maintain leadership and credibility in health care, research, and government. Achieving and maintaining operational excellence requires clear and effective guidance on AI roles and responsibilities, rigorous monitoring and oversight mechanisms, and comprehensive accountability mechanisms in the case of system failure. On these aspects, the GAO Accountability Framework and NIST AI RMF are particularly informative.

II. The VA Trustworthy AI Framework

Overview

The framework described here addresses the requirements of E.O. 13960 in addition to incorporating the perspectives of other relevant frameworks, namely: the VA Data Ethics Framework, the Blueprint for an AI Bill of Rights, the OECD AI Principles, NIST AI RMF, the GAO Accountability Framework, and the DoD AI Ethical Principles. The proposed VA Trustworthy AI Framework is the result of harmonizing these existing frameworks. Our goal with this proposed framework is twofold: to align with relevant trustworthy AI frameworks and standards that have impact on VA's mission, and to satisfy AI needs among the large and diverse group of VA stakeholders. Mission alignment and stakeholder fulfillment are essential features of any agency-specific trustworthy AI framework, so other efforts may consider replicating or modifying the process outlined in this section.

Figure 1. The VA Trustworthy AI Framework
The framework described below consists of six principles, which are illustrated in the outer hexagons in the figure to the left. These principles were selected and refined by examining relevant existing trustworthy AI frameworks and aligning elements to the mission and values of VA. Details on the construction of this framework can be found in the Supplemental Appendix.

In the following sections we provide descriptions of VA-specific trustworthy AI principles, their relationship to existing frameworks, and VA stakeholders with a particular interest in each principle.

VA Trustworthy AI Framework (PREDECISIONAL)	Purposeful	Effective & Safe		Secure & Private		Fair & Equitable	Transparent & Explainable		Accountable & Monitored	
	AI technologies are used to provide clear benefits to Veterans with minimal risks	VA AI systems are designed and monitored for robustness, accuracy, and reliability.	VA AI systems are rigorously tested and continuously monitored to ensure safety and well-being of Veterans	VA AI models are resilient against vulnerabilities and malicious exploitation	Stewardship of Veteran data is maintained in accordance with laws and VA's data ethics principles	VA manages and monitors AI systems for potential bias and algorithmic discrimination	Veterans expect to know when AI systems are used and what data is used by those systems	VA provides straightforward information on how AI systems work and are used to make healthcare decisions	VA promotes a culture of responsibility and learning across the AI lifecycle	VA uses logging, analytics, and automation to minimize uncertainty about AI operations
EO 13960	3 (b) Purposeful & performance driven	3 (c) Accurate, reliable and effective	3 (d) Safe, secure, and resilient	3 (d) Safe, secure, and resilient	3 (a) Lawful and respectful of our Nation's values (including privacy)	3 (a) Lawful and respectful of our Nation's values (including civil rights and liberties)	3 (h) Transparent	3 (e) Understandable	3 (f) Responsible & traceable	3 (g) Regularly monitored 3 (i) Accountable
VA Data Ethics Framework (38 CFR 0.605)	1. For the good of Veterans 6. Reciprocal obligation to Veterans	7. Obligation to ensure data security, quality, and integrity	7. Obligation to ensure data security, quality, and integrity	7. Obligation to ensure data security, quality, and integrity 5. Principled de-identification	7. Obligation to ensure data security, quality, and integrity 5. Principled de-identification	2. Equity 6. Reciprocal obligation to Veterans	3. Meaningful choice 4. Transparency	8. Veteran access to their own information 9. Veteran right to request amendment to their own information	6. Reciprocal obligation to Veterans	
EO 13985 & EO 14091	Provide equal opportunity and benefits; identify underserved communities; design policies to advance equity					Consistent and systematic fair, just, and impartial treatment of all individuals Advances equity				
White House Blueprint for an AI Bill of Rights		1. Safe and effective systems	1. Safe and effective systems	1. Safe and effective systems (security in context of safety)	3. Data Privacy	2. Freedom from algorithmic discrimination.	4. Notice and Explanation	4. Notice and Explanation	5. Human Alternatives, Consideration, and Fallback	
NIST AI RMF		4.1 Valid and reliable	4.2 Safe	4.4 Secure and resilient	4.7 Privacy-enhanced	4.3 Fair – and bias is managed	4.5 Transparent and Accountable	4.6 Explainable and interpretable	4.5 Transparent and Accountable	4.5 Transparent and Accountable
OECD AI Principles		1.4 Robustness, security and safety	1.4 Robustness, security and safety	1.4 Robustness, security and safety	1.2 Human-centered values and fairness	1.2 Human-centered values and fairness	1.3 Transparency and explainability	1.3 Transparency and explainability		1.5 Accountability
GAO AI Accountability Framework	3.1 – 3.7 Produce results that are consistent with program objectives	3.1 – 3.7 Results are consistent with program objectives 2.2 Reliable data used to develop models	1.6 Risk management	2.8 Security and Privacy	2.8 Security and Privacy	3.8 Bias: identify potential biases resulting from the AI system	1.9 Promote transparency by enabling external stakeholders to access information	1.9 Promote transparency by enabling external stakeholders to access information		3.9 Human supervision 4.1 – 4.5 Monitoring
DOD AI Ethical Principles	4. Explicit well-defined uses	4. Effectiveness subject to lifecycle assurance	4. Safety subject to lifecycle assurance	4. Security subject to lifecycle assurance 5. Detect and avoid unintended consequences		2. Take deliberate steps to minimize unintended bias	3. Possess transparent methodologies; data sources; and design procedures and documentation	3. Possess auditable methodologies; data sources; and design procedures and documentation	1. DoD personnel responsible for development, deployment, and use of AI capabilities	3. Auditable processes 4. Testing and assurance across lifecycles

Figure 2: The VA Trustworthy AI Framework principles mapped back to principles in existing frameworks. *The Proposed VA Trustworthy AI Framework principles mapped back to principles in existing frameworks. The table above illustrates the relationship between VA trustworthy AI principles (top row in gray) and principles in existing frameworks (lower rows). Brief definitions of VA trustworthy AI principles are provided in the top row and detailed descriptions can be found in the sections below. Further details about how this mapping was conducted can be found in the Supplemental Appendix.*

Purposeful

AI technologies are used to provide clear benefits to Veterans with minimal risks.

E.O. 13960 and the VA Data Ethics Framework both stipulate that AI should be used for a clear purpose. That purpose, as required by the VA Data Ethics Framework, is to provide a clear benefit to Veterans. The GAO AI Accountability Framework has a similar requirement.

In line with the OECD's principle of "Inclusive growth, sustainable development and well-being", all VA AI applications should be accompanied by concrete metrics describing their proposed benefits against which they can be measured after deployment. These outcomes should prioritize Veteran well-being while taking into account AI's potential to reduce disparities among disadvantaged populations.

Summary

- In accordance with existing VA policy on the use of Veteran data, AI that utilizes Veteran data should convey a clear benefit to Veterans.
- AI used in administrative contexts not involving Veteran data should convey a clear benefit to VA.
- AI that is the subject of VA research should aim to address a clear need in one of the above areas.

Stakeholders

- VHA Office of Research Oversight
- VHA Office of Integrity and Compliance
- OIT Office of Information Security
- VBA Automated Benefit Delivery
- VA Center for Minority Veterans
- VA Center for Women Veterans

Corresponding Sections in Existing Frameworks

E.O. 13960	3(b) Purposeful & performance driven
VA Data Ethics Framework (38 CFR 0.605)	(c)(1) For the good of Veterans (c)(6) Reciprocal obligation to Veterans
White House Blueprint for an AI Bill of Rights	N/A
OECD AI Principles	N/A
NIST AI RMF	N/A
GAO AI Accountability Framework	(3.1-3.7) Produce results that are consistent with program objectives
DoD AI Ethical Principles	(4) Reliable: Explicit well-defined uses

Effective & Safe

VA AI systems are designed and monitored for accuracy, reliability, and robustness. Risks are proactively identified and managed to ensure the safety and well-being of Veterans.

Efficacy and safety principles are present in all trustworthy AI frameworks considered here. Efficacy includes reliability, robustness, and accuracy across a system's lifespan. Safe AI systems should not cause physical or psychological harm, nor endanger human life, health, or property.[18]

In health care settings, efficacy of systems is vital to ensuring safety. Systems that provide diagnostics or other health care services cannot be safe if they are not accurate and reliable, since mistakes may directly affect a patient's health. To reflect this, we have combined safety and efficacy into a single principle.

AI tools need to be thoroughly tested and supported by rigorous statistical (and, where appropriate, causal) evidence, particularly for clinicians who may consider the use of AI to augment decision-making or assist with procedures. While many AI models will not yield a single causal variable, there should be sufficient scrutiny to specify causal sets of variables and rule out the possibility that unobserved correlates are driving the phenomena of interest. The methodology that is employed to advocate for specific applications of AI must be carefully tested and investigated before deployment to avoid the loss of life or unintended, adverse consequences.

AI systems should produce results that are appropriate and accurate by existing standards relevant to their use, do so reliably and consistently,

[18] See under "safety" at ISO/IEC TS 5723:2022(en), Trustworthiness — Vocabulary

and be able to continue functioning accurately and reliably under the conditions that may be reasonably expected in the context in which they are situated. Under unexpected conditions, AI systems should robustly minimize risk, falling back to human decisionmakers, shutting down, or pausing as appropriate. This should be true throughout the lifetime of the system, and VA will monitor systems to ensure they meet these criteria.

Sources of risk should be identified, removed when feasible, and carefully moderated and monitored when complete elimination is not possible. Risks change over time responsive to changing circumstances and technologies, so changes should be taken into account in the monitoring process. The safety and well-being of Veterans should always be the primary aim of work done with AI at VA.

Summary

- Systems should be effective based on intended use
- Systems should function accurately, reliably, and robustly across their lifespans
- Systems should function safely across their lifespans
- Risks are proactively identified and mitigated
- Safety should be monitored with an eye to changing circumstances and technologies

Corresponding Sections in Existing Frameworks

E.O. 13960	3(c) Accurate, reliable, and effective 3(d) Safe, secure, and resilient
VA Data Ethics Framework (38 CFR 0.605)	(c)(7) Ensure data quality, security, and integrity
White House Blueprint for an AI Bill of Rights	(1) Safe and effective systems
OECD AI Principles	(1.4) Robustness, security, and safety
NIST AI RMF	(4.1) Valid and reliable (4.2) Safe
GAO AI Accountability Framework	(1.6) Risk Management (3.1-3.7) Results consistent with objectives (2.2) Assess reliability of model development data
DoD AI Ethical Principles	(4) Safety and effectiveness subject to lifecycle assurance

Stakeholders

- VHA Office of Research Oversight
- VHA Office of Quality and Patient Safety
- VBA Office of Policy and Oversight
- VA Center for Women Veterans
- VBA Office of Automated Benefit Delivery
- VBA Office of Performance Analysis and Integrity
- VA Center for Minority Veterans

Secure & Private

VA AI models are resilient against vulnerabilities and malicious exploitation. Veteran data is maintained in accordance with laws and VA data ethics principles to preserve privacy.

Security and privacy are closely linked, especially in a health care setting. In fact, the healthcare sector has the greatest number of data breaches, according to the Privacy Rights Clearinghouse, elevating the importance of data security over sensitive information.[19] Security protects data and systems from threats. Privacy ensures that the collection and use of data does not lead to exposure of sensitive information that jeopardizes the VA or its stakeholders, especially Veterans.

Responsibilities for the protection of privacy in health care are already established in existing law (e.g., The Common Rule, HIPAA) and the VA Data Ethics Framework, so we defer to these sources for a more detailed discussion.

Summary

- Systems should be designed to function securely across their lifespans
- Systems should be resilient in the face of realized risks and changing circumstances
- Handled in alignment with existing VA Data Ethics Framework
- Use of systems should remain consistent with Constitution and privacy law
- Usage of privacy-preserving methods, such as the possible use of synthetic data and zero-knowledge proofs
- Data not used beyond intended purpose

[19] https://privacyrights.org/data-breaches

Corresponding Sections in Existing Frameworks

E.O. 13960	3(d) Safe, secure, and resilient 3(a) Lawful and respectful of our Nation's values (including privacy)
VA Data Ethics Framework (38 CFR 0.605)	(c)(7) Ensure data security, quality, and integrity (c)(5) Principled de-identification
White House Blueprint for an AI Bill of Rights	(1) Safe and effective systems (security in context of safety) (3) Data Privacy
OECD AI Principles	(1.4) Robustness, security and safety (1.2) Human-centered values and fairness
NIST AI RMF	(4.4) Secure and resilient (4.7) Privacy-enhanced
GAO AI Accountability Framework	(2.8) Security and Privacy: Assess data security and privacy for the AI system
DoD AI Ethical Principles	(4) Security subject to lifecycle assurance (5) Governable: Detect and avoid unintended consequences (Privacy not addressed.)

Stakeholders

- OIT Office of Information Security
- OIT Privacy Office
- VHA Oversight, Risk and Ethics
- VHA Office of Research Oversight
- VHA Office of Health Information Governance (within Office of Health Informatics)
- VA Center for Minority Veterans
- VA Center for Women Veterans

Fair & Equitable

VA manages and monitors AI systems for potential sources of bias and algorithmic discrimination.

We define bias in the context of AI following the statistical literature: instances where the expected value of the results differs from the true underlying parameter of interest. Such systematic deviations may vary in ways that are correlated with relevant data features, ranging from gender to socioeconomic status to geography. Although there are many competing definitions of fairness, we define it according to Dwork et al. who introduce the concept of "individual fairness," referring to phenomenon where similar inputs among different people yield similar outputs.[20]

Additionally, we use the definition of equity as provided by E.O. 14091: "...the consistent and systematic treatment of all individuals in a fair, just, and impartial manner, including individuals who belong to communities that often have been denied such treatment."

If left unchecked, bias may lead to algorithmic discrimination. Algorithmic discrimination is defined by the White House AI Bill of Rights as "when automated systems lead to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law."

Bias should be actively identified, evaluated, eliminated when possible and closely managed and monitored when elimination is not possible. Bias identification and management should occur throughout the lifecycle of the AI and in all stages of its use, from datasets to implementation of results. Bias may exist in any dataset, but sophisticated statistical effort and quality data should be used to correct and establish boundaries for this bias, especially as it relates to variables of interest. This is in line with the NIST AI RMF, which requires that use of AI be fair and bias-managed.

Bias and disparities in health care are well documented, such as the underdiagnosis of heart attacks in women, and the decreased access to pain management for Black patients.[21] Lack of diversity in clinical trials has likewise been a long-running issue, on which NIH is now taking action.[22] Without attention to the root causes behind existing variation in the data, AI models may learn inaccurate associations between certain characteristics and health outcomes, propagating inequalities.

A 2019 study provides a now-seminal illustration of how model misspecification can perpetuate existing disparities. In this case, a large-scale application of AI deprioritized Black patients for delivery of health care services because the AI was trained to predict future health care costs rather than health needs and outcomes. Due to existing disparities in access to care, this approach improperly conflated data describing

[20] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. Fairness Through Awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 12), 2012, 214-226.

[21] Starke, G., De Clercq, E. & Elger, B.S. Towards a pragmatist dealing with algorithmic bias in medical machine learning. Med Health Care and Philos 24, 341–349 (2021). https://doi.org/10.1007/s11019-021-10008-5

[22] Shamoo, A. E., & Resnik, D. B. (2009). Responsible conduct of research. Oxford University Press.

ability to pay with relevant diagnostic information.[23] This result underscores the importance of checking not only the quality of the data, but also the underlying variables that are used to train AI systems.

In line with E.O. 13985, VA AI activities should be conducted equitably, justly, and impartially, with an eye to correcting historical underserving and marginalization of affected groups and affirmatively advance equity, civil rights, racial justice, and equal opportunity.

Summary

- AI activities should be lawful and respectful of our Nation's values, including Constitutional rights and civil rights laws
- Bias should be identified, assessed, and managed throughout the lifecycle of the technology
- Stakeholder consultation encouraged; diversity of input is vital
- Follow E.O. 13985 requirements for advancing equity

Stakeholders

- National Center for Ethics in Health Care
- VHA Office of Quality and Patient Safety
- VHA Oversight, Risk, and Ethics
- VHA Office of Research Oversight
- VA Center for Women Veterans
- VHA Office of Quality and Patient Safety
- VHA Oversight, Risk, and Ethics
- VBA Office of Administrative Review
- VA Center for Minority Veterans

Corresponding Sections in Existing Frameworks

E.O. 13960	3(a) Lawful and respectful of our Nation's values (including civil rights and liberties)
VA Data Ethics Framework (38 CFR 0.605)	(c)(2) Equity (c)(6) Reciprocal obligation to Veterans
White House Blueprint for an AI Bill of Rights	(2) Freedom from algorithmic discrimination.
OECD AI Principles	(1.2) Human-centered values and fairness
NIST AI RMF	(4.3) Fair – and bias is managed
GAO AI Accountability Framework	(3.8) Bias: identify potential biases, inequities, and other societal concerns resulting from the AI system
DoD AI Ethical Principles	(2) Equitable: Take deliberate steps to minimize unintended bias

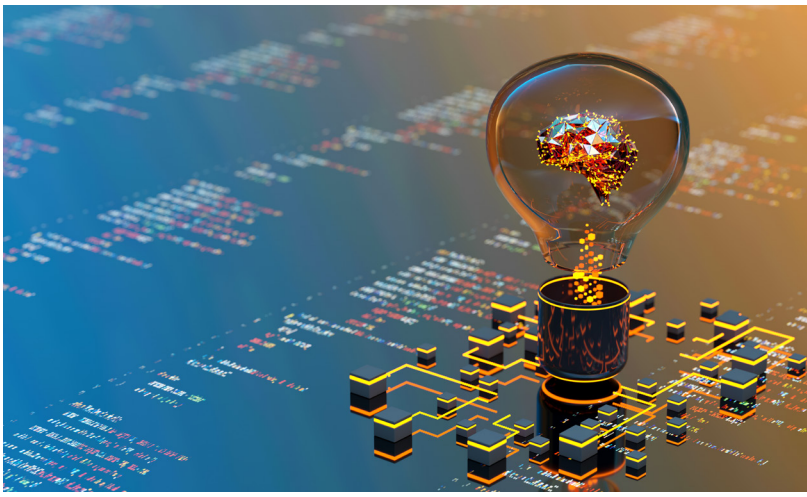
[23] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464): 447-453.

Transparent & Explainable

Veterans expect to know when AI systems are used and what data is used by those systems. VA provides straightforward information on how AI systems work and are used to make decisions.

Transparency is the ease with which relevant parties can see how and why AI is being used. To build trust, stakeholders should understand when, why, and how AI is being used, and this information should be communicated in ways that are broadly accessible for stakeholders from different backgrounds. Information on how AI systems are monitored and corrected should be made available.

Explainability refers to the accessibility and ease of understanding the output of AI models. A common critique of AI is that the underlying mechanisms that generate AI system outputs are a "black box." That is, the AI tool operates in ways that are not well-understood by humans because of the scale and complexity of the computational activities being performed. Explainability is especially important for clinical research because of the special relationship between clinicians and patients; if patients do not understand why they should adhere to a recommendation, trust is undermined. Likewise, if clinicians do not understand the logic behind an AI-driven recommendation, they are less likely to convey the information to patients and trust the recommendation – and rightly so. The NIST AI RMF points out that explainable systems are more easily debugged, audited, monitored and governed, and the OECD AI principles note that explainability fosters greater trust in AI systems.



Summary

- Informed consent is vital to ethical research, and consent cannot be informed if the patient does not understand the materials provided. Therefore, information should be made available in an understandable format.
- Users should be informed of the reason for use of the system and the way in which the system operates.
- Stakeholders have access to information about the system in use, including monitoring and correction.
- Information should be presented in an accessible manner.
- All relevant parties should understand what data is being used, and how it is being used.

Stakeholders

- VHA Office of Research Oversight
- VHA Office of Quality and Patient Safety
- VHA Office of Health Informatics
- Veteran Experience Office
- VBA Office of Automated Benefit Delivery
- VBA Office of Administrative Review
- VA Center for Minority Veterans
- VA Center for Women Veterans

Corresponding Sections in Existing Frameworks

E.O. 13960	3(e) Understandable 3(h) Transparent
VA Data Ethics Framework (38 CFR 0.605)	(c)(3) Meaningful choice (c)(4) Transparency (c)(8) Veteran access to their own information (c)(9) Veteran right to request amendment to their own information
White House Blueprint for an AI Bill of Rights	(4) Notice and Explanation
OECD AI Principles	(1.3) Transparency and explainability
NIST AI RMF	(4.5) Transparent and accountable (4.6) Explainable and interpretable
GAO AI Accountability Framework	(1.9) Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.
DoD AI Ethical Principles	(3) Traceable: Possess transparent and auditable methodologies, data sources, and design procedures and documentation.

Accountable & Monitored

VA promotes a culture of responsibility and learning across the AI lifecycle. VA uses logging, analytics, and automation to minimize uncertainty about AI operations. AI is used in line with existing VA frameworks, such as IRB requirements for informed consent, and the existing VA Data Ethics Framework. Human fallbacks and monitoring are provided, where appropriate.

Accountability is emphasized in frameworks such as the NIST AI RMF, the GAO AI Accountability Framework, and the HHS Trustworthy AI Principles. That means not only clearly designating the accountable parties, but also proactively monitoring and evaluating inputs and outcomes and addressing concerns with the appropriate parties to ensure continued improvement.

In order to establish accountability, AI use must be monitored. Monitoring appears in E.O. 13960 and the GAO AI Accountability Framework, both of which concern the application of AI in the federal government. Though it only appears explicitly in two documents, monitoring is implicit in the Blueprint for an AI Bill of Rights requirement for human fallback, and other requirements for transparency and explainability. This process ensures that AI applications are routinely tested and feedback is incorporated into the system to avoid risks such as model drift.

The VA Data Ethics Framework require Veterans to be given meaningful choice about the use of their data, and the Blueprint for an AI Bill of Rights recommends that people should be able to opt out of AI usage. When AI is using data that has the potential to compromise the subject's safety or is involved in a decision with impacts on health, wellbeing, or safety, consent to its use is vital.

For research and health care, this principle is constructed with the understanding that AI utilization will adhere to the already established requirements at VA for informed consent, whether they be IRB or patient care requirements. As noted in Transparency & Explainability, these processes are expected to be presented to research participants and patients in a clearly understandable format, and alternatives should be presented where appropriate. The enforcement of informed consent procedures rests with the established entities, but this trustworthy AI framework recognizes that they are vital in protecting the interests of VA employees and the Veterans we serve, and vital to the successful implementation of AI in a trustworthy and ethical manner at VA.

Summary

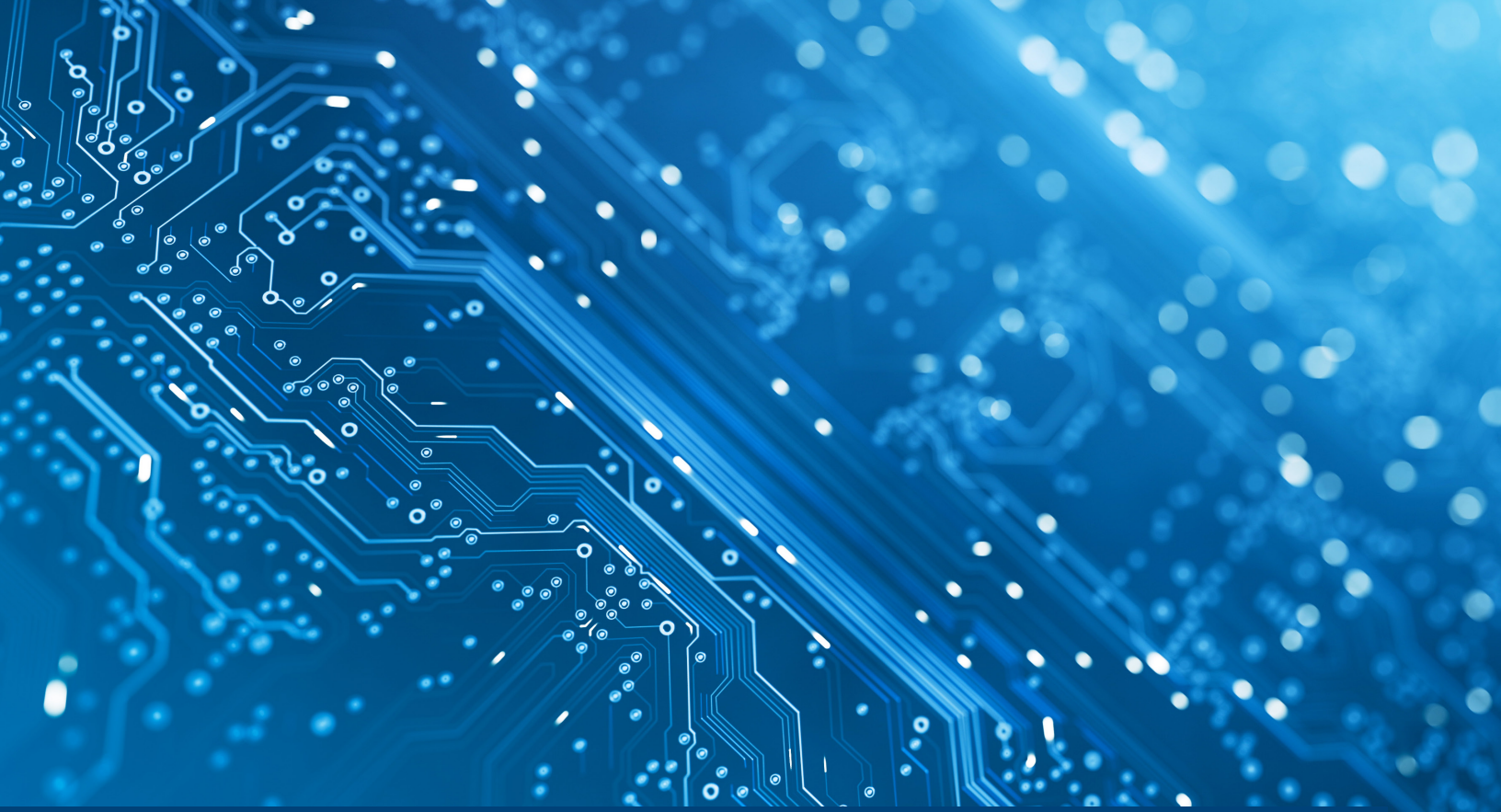
- Systems should be regularly monitored.
- Clear lines of accountability should be established for all AI programs used by VA.
- AI use should adhere to existing rules, regulations, and law as appropriate, especially regarding informed consent for treatment and medical research.

Stakeholders

- VHA Office of Research Oversight
- VHA Optimizing Health Care Value Program
- Veterans Experience Office
- VA Center for Minority Veterans
- VA Center for Women Veterans

Corresponding Sections in Existing Frameworks

E.O. 13960	3(f) Responsible & traceable 3(g) Regularly monitored 3(i) Accountable
VA Data Ethics Framework (38 CFR 0.605)	(c)(6) Reciprocal obligation to Veterans (c)(8) Veteran access to their own information (c)(9) Veteran right to request amendment to their own information
White House Blueprint for an AI Bill of Rights	N/A
OECD AI Principles	(1.5) Accountability
NIST AI RMF	(4.5) Transparent and Accountable
GAO AI Accountability Framework	(3.9) Human supervision: Define and develop procedures for human supervision of the AI system to ensure accountability (4.1-4.5) Monitoring: Ensure reliability and relevance over time
DoD AI Ethical Principles	(1) Responsible: DoD personnel responsible for development, deployment, and use of AI capabilities. (3) Traceable: Auditable processes (4) Reliable: Testing and assurance across lifecycles



III. Recommendations and Next Steps

For VA to transition towards Trustworthy AI Excellence, it must provide concrete actionable and measurable guidance to VA staff who design, develop, acquire, and manage AI systems. Also, VA must provide a means for measuring maturity and progress of principle implementations to meet various requirements, so the VA Trustworthy AI Framework must connect to measurable actions and outcomes.

To satisfy the above aims, the VA Trustworthy AI Framework is designed to serve as the foundation for a future playbook and assessment process to guide and measure VA-wide implementation of trustworthy AI principles.

While this framework is contextualized to the VA, we believe that the overarching principles and work that we have done in harmonizing other international and domestic frameworks is a useful blueprint for other federal agencies too, tailoring the implementation as they see fit.

Upon VA concurrence of this framework as the VA trustworthy AI standard, we recommend the following activities under the leadership of the National AI Institute (NAII) and oversight by the VA Data Governance Council (DGC) in alignment with stakeholders across VA.

- Develop a playbook that provides implementation guidance for each VA trustworthy AI principle. The playbook chapters must provide specific guidance that is actionable to VA staff and decision makers involved in design, development, acquisition, and use of AI systems at VA. The guidance must also be measurable to provide VA AI system owners and leadership with reliable and repeatable gauges of consistency of VA's AI systems with E.O. 13960 requirements, as well as track maturity of trustworthy AI systems across VA.
- Develop a risk-based model for assessing consistency of VA AI systems with E.O. 13960 requirements and adherence to other trustworthy AI frameworks of importance to VA. Although desirable, it is unrealistic to expect AI systems can achieve a perfect or absolute compliance with each trustworthy AI principle. Instead, VA must develop models and processes that inform risk-based decision making for AI system consistency to each principle. This provides a practical real-world assessment of consistency with trustworthy AI principles as well as an approach to developing plans of action and milestones (POAMs) to AI system owners to improve consistency with principles. This is consistent to the approach taken by organization to assess and implement cybersecurity controls and processes.
- Pilot the trustworthy AI framework implementation guidance across VA. A successful pilot plan will provide several phases that are progressively more complex in location scale and AI application area. Initial pilots will be conducted at smaller scale location with AI systems focused on specific health care and benefits domains that are well-understood by practitioners and leadership. Pilots will scale to larger AI systems with increasingly complex requirements, such as telesurgery and multimodal clinical applications. Successful pilot of VA's trustworthy AI framework implementation will test VA's risk-based approach for assessing consistency with the framework's goals and intent as well as provide AI system owners with a baseline of their specific AI systems level of trustworthiness.

IV. Supplemental Appendix

Building a Trustworthy AI Framework for VA

Identifying Common Principles

Having enumerated a selected sample of notable trustworthy AI frameworks, we now discuss how we created a harmonized framework for trustworthy AI. We identified the commonalities across the frameworks we selected. To do this, we first isolated the principles within each framework and examined the way in which each framework defined them. We then used qualitative methods (constant comparative analysis and thematic analysis) to compare principles across frameworks, using common elements in their definitions, as shown in Supplemental Figure 1.

In this process, we identified substantial overlap in the priorities of all the selected frameworks. At a basic level, most frameworks contained similar principles. For example, "safety and security" occurs as a principle in some form throughout all the frameworks reviewed. So does "privacy" and variations on "fairness" (such as "freedom from algorithmic bias" and "bias is managed"). There were only two principles that appeared in one framework that did not have

an equivalent in any other framework: one principle from the OECD and one from the White House AI Bill of Rights.

Though principles broadly overlapped with one another, they were often composed of different elements, or differed in their emphasis of certain elements. These differences reflect the context from which each of these frameworks arose. For example, the White House AI Bill of Rights includes security of AI systems as a safety issue, while E.O. 13960 and the NIST AI RMF dedicate independent sections to it. The management of bias and preventing algorithmic discrimination feature prominently in the White House AI Bill of Rights, the NIST AI RMF, and the GAO AI Accountability framework, but E.O. 13960 covers the issue simply with "Lawful and respectful of our Nation's values." While it is up to individual stakeholders to decide the level of emphasis required for each principle based on their organizational activities, we simply looked to harmonize these differences so there exists a common taxonomy.

The strongest points of overlap between principles were pinpointed for inclusion in our framework. But differences, too, were informative, especially when the context was closest to that of VA. One such example of this is the HHS Trustworthy AI Principles, which have very specific requirements for managing bias, involving feedback from a diverse group of stakeholders. The similar context (health care) with its particular history of bias, makes this recommendation a relevant one for VA, and led to the suggestion of stakeholder involvement in our own proposed framework.

	Purposeful	Effective	Safe	Secure	Private	Fair & Equitable	Transparent	Explainable	Accountable	Monitored
EO 13960	Purposeful & performance driven; should be used when benefits outweigh risks	Accuracy, reliability, efficacy are vital	Agencies ensure safety	Agencies ensure security & resiliency	Consistent with constitution & privacy law	Lawful & respectful, including civil rights & liberties;	Transparency in disclosing information to stakeholders	Operations & outcomes understandable to relevant parties	Agencies implement & enforce appropriate safeguards	Regularly monitored
VA Data Ethics Framework	For the good of Veterans Reciprocal obligation to Veterans	Obligation to ensure data security, quality, and integrity	Obligation to ensure data security, quality, and integrity	Obligation to ensure data security, quality, and integrity; Principled de-identification	Obligation to ensure data security, quality, and integrity; Principled de-identification	Equity; Reciprocal obligation to Veterans	Meaningful choice; Transparency	Veteran access to their own information; Veteran right to request amendment to their own information	Reciprocal obligation to Veterans	Veteran access to their own information; Veteran right to request amendment to their own information
White House Blueprint for an AI Bill of Rights	No	Systems should be effective based on intended use	Protects users from foreseeable possibilities of harm	Security tangentially under safety	Built in; users have agency over their data	Freedom from algorithmic discrimination	Inform of use of automated system	User should understand automated system's impact on them & how decision is made	No	No
NIST AI RMF	No	Valid and reliable	Safety to be monitored	Withstand adversarial attacks	Risks from processing of data and technical aspects should be considered	Fair, bias is managed	No	Explainable systems more easily debugged, audited, governed, documented, monitored	Expectations should be set of responsible party in case risks realized	No
OECD AI Principles	Inclusive growth, sustainable development and well-being	Function in robust way across lifetime	Function in safe way across lifetime	Function in secure way across lifetime	Human-centered values and fairness	Human-centered values and fairness	User should understand reason for engagement, have enough information to challenge	Meaningful information appropriate to context made available	Implementing entities should be held accountable for proper functioning	No
GAO AI Accountability Framework	Entities should have clear goals, establish tech. specifications to ensure intended purpose met	Reliability and accuracy should be tested across lifespan	Risks identified and mitigated	Risks identified and mitigated	Security and privacy assessed; address use of synthetic, imputed, augmented data	Identify and assess potential biases	External stakeholders have access to information; monitoring & correction documented	No	Define & develop procedures for human supervision & accountability	Regular monitoring
DoD AI Ethical Principles	Reliable: well defined use-cases	Reliable	Reliable	Reliable; Governable: detect & prevent unintended consequences	[Implicit in the context of national defense]	Equitable: will take steps to minimized unintended bias in AI capabilities	Traceable: possess transparent and auditable methodologies, data sources	Traceable: under "auditable"	Responsible: for development, deployment, and use of AI capabilities.	Traceable: auditable processes; Reliable: testing and assurance

Supplemental Figure 1: Principles across selected Trustworthy AI Frameworks. *Principles that differ slightly from interpretations in other frameworks are marked in light blue; principles which are absent are marked in dark blue.*

Selecting and Defining Principles

We proceeded to use these commonalities and differences to inform a framework that would fit the context of VA. In this process, we attempted to identify "fundamental" elements that composed the different principles we identified above. To do so, we closely inspected the definitions of each principle, noting when non-overlapping elements arose. For example, the White House AI Bill of Rights names "Safe and Effective" as a principle, which we then decomposed further into component parts (namely "safety" and "efficacy"). In this case, though safety and efficacy may be related in certain ways, they have different technical definitions as they relate to AI systems, so we chose to separate them.

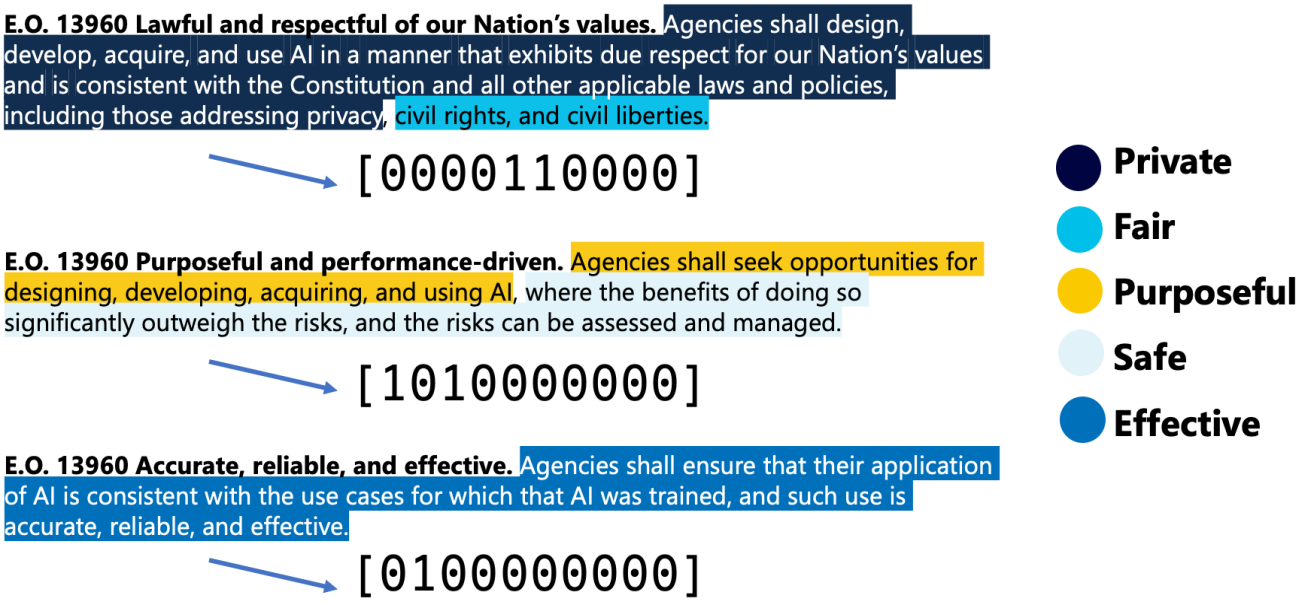
We carried out this exercise for each principle, allowing us to build definitions for non-overlapping elements. Once this process was complete, we manually coded each principle from every framework to indicate the presence or absence of common non-overlapping elements across frameworks. Doing so ensured that the set of elements we identified were sufficient to fully cover the existing frameworks and principles; that is, we were able to reconstruct each existing principle by combining different non-overlapping elements as appropriate.

With these elements and definitions in hand, we identified ten principles selected for inclusion in the VA Trustworthy AI Framework: Purposeful, Effective, Safe, Secure, Private, Fair, Transparent, Explainable, Accountable, and Monitored.

Refining and Grouping Principles

Internal review at NAII team meetings highlighted the need to condense these ten principles for ease of reference. After discussion about which principles were most complementary, and should be combined, six final principles were proposed: Effective & Safe, Secure & Private, Fair & Bias-Managed, Transparent & Explainable, Accountable & Monitored, and Purposeful. At this point, the definitions of these principles were drafted, as seen in the "Discussion of Trustworthy AI Principles" section above.

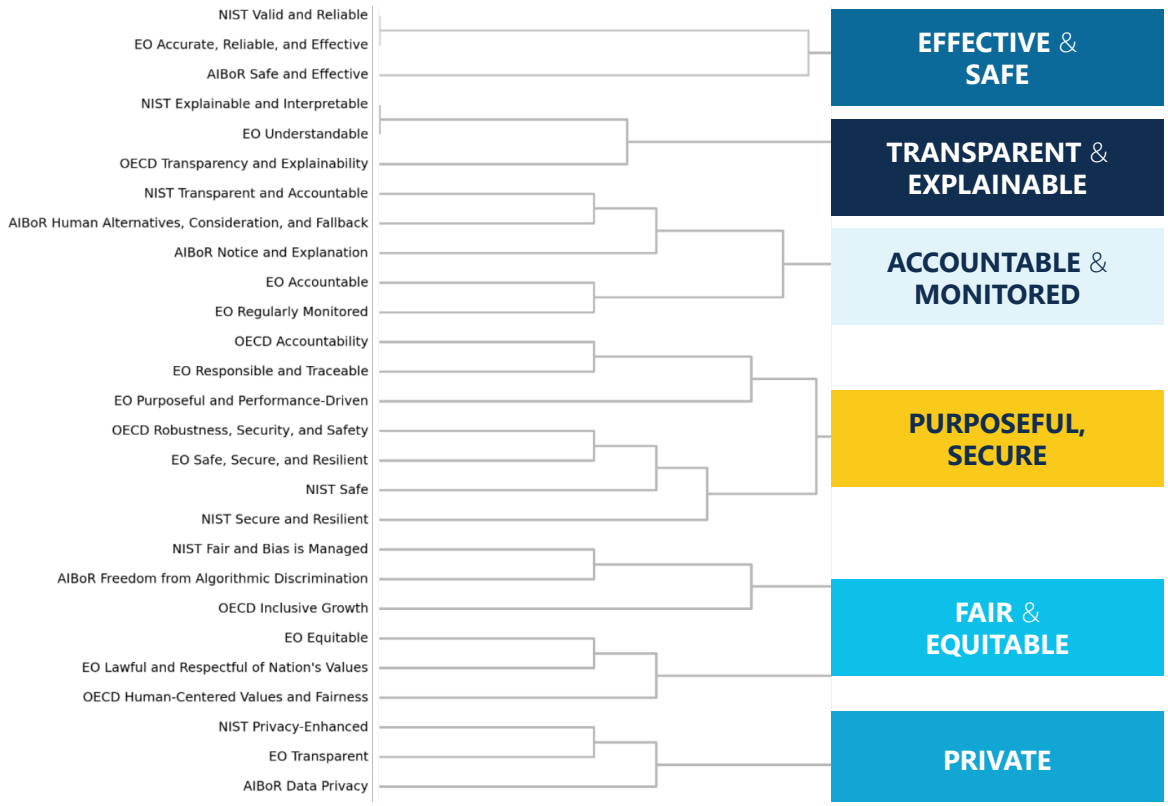
As a robustness check to determine whether these groupings were similar to groupings of principles in existing frameworks, we used the manual coding described above as an input for quantitative analysis. As shown in Figure 3, principles were transformed into binary vectors to indicate the presence or absence of each of the ten non-overlapping elements. For example, "Purpose" obtains the first spot in the vector, "Efficacy" the second spot, and so on. In this sense, the first example below in Supplemental Figure 2 has a sequence of four zeros followed by two ones and four zeros, meaning that the clause covers "Fairness" and "Privacy."



Supplemental Figure 2: Sample coding of trustworthy AI principles from E.O. 13960.
Comparative analysis identified common elements between frameworks, which were compiled into the set of elements shown in the box above. Principles from each framework were coded as illustrated on the left, where colored shading indicates the presence of an element. Principles were then coded as binary presence/absence vectors, where a "0" indicates the absence of an element and a "1" indicates its presence, allowing for quantitative comparison.

This vector encoding allowed us to quantitatively compare principles using hierarchical clustering on the 10-dimensional vectors. (The ten vector dimensions correspond to the ten trustworthy AI framework principles: purposeful, effective, safe, secure, private, fair and bias-managed, transparent, explainable, accountable, and monitored. The presence (or absence) of these elements within a particular framework principle was determined by manual coding.) This technique can reveal the relationship between different objects by grouping them according to similarity. It can be used to generate a dendrogram, which is a visual representation of relatedness similar to a family tree. The output of performing hierarchical clustering on the full set of selected principles is shown in Supplemental Figure 3.

For principles where short text summaries were available, the full text (approximately one paragraph) was used as the definition of the principle. For principles where summaries were not available, paragraph-length summaries were manually constructed from longer descriptions. We used the *scipy.cluster.hierarchy* functions *linkage* and *dendrogram* using Ward's linkage method with Euclidean distance for clustering.



Supplemental Figure 3: Hierarchical Clustering of Trustworthy AI Principles. *Principles from each framework were coded as binary vectors as illustrated in Figure 3, allowing comparison using hierarchical clustering. Shown is a visual representation of the output of the clustering algorithm, where elements that are most closely related are joined further to the left of the diagram. Application of clustering revealed that principles from different frameworks can be grouped thematically according to a scheme (illustrated using colored blocks on the right) very similar to what is shown in Figure 2. This high degree of agreement between qualitative and quantitative methods provides reason to believe that the Proposed VA Trustworthy AI Framework comprehensively and rigorously synthesizes principles and themes from other existing trustworthy AI frameworks.*

We then compared this quantitative analysis to the qualitative analysis represented in Figure 2, the thematic analysis of principles across frameworks. The two methods of analysis produced similarly overlapping sets of principles, increasing our confidence that our main efforts, which revolved around thematic analysis, stakeholder engagement, and team discussion, were theoretically sound.

Further Discussion of Informed Consent for AI at VA

Informed consent is one of the established requirements and procedures to which AI use at VA must adhere. Information on existing informed consent procedures at VA may be found in the VHA Handbook 1004.01(5) and VHA Directive 1058.03. The following is an overview of existing VA procedures with considerations for AI suggested.

Informed consent procedures at VA are differentiated into informed consent for treatment, and informed consent for research. These are two distinct processes and should be considered separately.

Treatment

Patients have a right to autonomous, informed participation in their health care decisions. The informed consent process at VA for health care enables the exercise of this right. Informed consent must be obtained for all treatment activities undertaken at VA.

During the informed consent process, information about the procedure should be provided in a clear and accessible way, so that the patient understands the situation, its potential consequences, what will happen in the procedure, and what treatment options there are. This will give the patient the information they need to make and communicate their choice. Some specific procedures require a signed consent; in the case of other, low-risk and routine procedures, oral consent is sufficient.

In the case of treatment involving an AI component, the following considerations should be observed:

- The patient should be informed of the use of AI
- The patient should be informed of how the AI makes its decision
- The patient should be informed of why the AI is being used instead of a human decisionmaker
- The patient should be informed of the differences between AI decisions and human decisions
- The patient should be informed of human fallback

Research

Informed consent is a cornerstone of the laws and regulations that govern research with human subjects in the United States. Informed consent for research differs from that in clinical treatment, and the procedures to obtain informed consent, as well as the information about the research activities to be provided to participants, must be reviewed by the Institutional Review Board (IRB). Further information on the process of review may be found in VHA Directive 1058.03.

AI, as a new technology, introduces new risks to medical research. These risks should be systematically evaluated and be well understood by all parties involved in research at the VA, including researchers, the IRB, and participants. In order to evaluate and address the specific risks that may arise in human subjects

research that uses AI, the NAII has piloted an AI IRB. This augments existing VA IRB human subjects protections by developing new questions that address AI-specific risks, for use when the IRB evaluates studies that use AI. This allows researchers and reviewers to assign a level of risk to the study, and decide on appropriate protections. The AI IRB is now being implemented in VA facilities in Washington, D.C.; Tampa, FL; Kansas City, MO; and Long Beach, CA. This effort is designed to pair with AI Model Cards, which describe key features of AI tools such as model architecture, data type, training process, performance metrics.

Identifying Stakeholders

Stakeholders were identified through an overview of the organizational landscape at VA using the 2020 VA Functional Organizational Manual. Stakeholders with topical focuses that dovetailed with the scope of each principle were selected. Stakeholder lists are open to expansion.


National Artificial Intelligence Institute
VA
U.S. Department of Veterans Affairs