Health Policy OPEN 7 (2024) 100128

Contents lists available at ScienceDirect

Health Policy OPEN

journal homepage: www.elsevier.com/locate/hpopen

From theory to practice: Harmonizing taxonomies of trustworthy AI

Christos A. Makridis [a,b,c,1,*], Joshua Mueller [a,1], Theo Tiffany [a], Andrew A. Borkowski [a], John Zachary [a], Gil Alterovitz [a,d]

[a] Department of Veterans Affairs, 810 Vermont Ave NW, Washington DC, 20001, United States
[b] University of Nicosia, Institute for the Future, AGC Towers, 28th October 24, Nicosia 2414, Cyprus
[c] Arizona State University, Business Administration, 300 E Lemon St, Tempe, AZ 85287, United States
[d] Brigham and Women's Hospital, Harvard Medical School, Center for Biomedical Informatics, Countway Lib, 10 Shattuck St Boston MA 02115, United States

ARTICLE INFO

Keywords:
Artificial intelligence
Ethics
Policy
Government
Trustworthy AI

ABSTRACT

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 there is no consensus on a single framework yet. We evaluate the leading frameworks and provide 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. We apply this approach to the Department of Veterans Affairs, an entity with largest health care system in US. Further, we contextualize our framework from the perspective of the federal government on how to leverage existing trustworthy AI frameworks to develop a set of guiding principles that can provide the foundation for an agency to design, develop, acquire, and use AI systems in a manner that simultaneously fosters trust and confidence and meets the requirements of established laws and regulations.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) models 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], particularly within the area of healthcare. Furthermore, the zeitgeist of large language models, most notably ChatGPT and all the subsequent plug-ins, has led to an explosion in new challenges and opportunities, particularly for health informaticists [3].

A large body of research suggests that AI can change the healthcare landscape by improving care outcomes, increasing the efficiency of care delivery, reduce administrative and other burdens for healthcare practitioners, manage demands of a changing and aging population, and even help get life-saving treatments to market faster [4]. Already, AI models can match or outperform physicians at diagnosing colorectal cancer [5], mesothelioma [6], and lung cancer [7]. 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 %, identifying risks before the condition is diagnosed using current standards of care [8]. 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 [9]. An AI tool trained on CT scans can correctly identify intra-cranial hemorrhaging with over 95 % accuracy, decreasing clinical turnaround time by over 40 % [10].

Our paper contributes to an emerging literature on practical approaches of mapping the broad aims of ethical AI to practice. Silcox et al. develop a "trust and value checklist" that helps clinicians understand and apply the recommendations from clinical decision support systems to patients [11]. Similarly, Dorr et al. suggest a code of conduct for AI in healthcare and Haupt and Marks caution against AI-generated medical advice and establishing standards where outcomes resulting from AI recommendations are benchmarked against existing outcomes [12]. Most closely related is Fjeld et al. who survey and analyze 36 documents relating to ethical AI, producing eight organizing themes from 47

* Corresponding author.
E-mail address: christos.makridis@va.gov (C.A. Makridis).
[1] Christos A. Makridis and Joshua Mueller are both first authors.

https://doi.org/10.1016/j.hpopen.2024.100128
Received 29 February 2024; Received in revised form 26 July 2024; Accepted 27 August 2024
Available online 5 September 2024

specific principles: privacy, accountability, safety & security, transparency & explainability, fairness & non-discrimination, human control of technology, professional responsibility, and promotion of human values [13].

Our paper builds on these prior contributions by producing a harmonized taxonomy of trustworthy AI principles from existing policy documents with applications to health informatics, particularly with regard to the ways that researchers, clinicians, and policymakers can manage risks associated with generative AI tools, like ChatGPT [14]. The federal sector, especially through the Department of Veterans Affairs, plays a large role in the U.S. healthcare system as the largest integrated medical provider, so our formalization and application of these principles will provide an important guide for practice. While our focus is on building trustworthy AI systems, our paper also builds upon an even larger literature in public health policy on the ethical changes that arise from predictive models [15].

We define AI according to the definition from the National Defense and Authorization Act (NDAA) of 2019, which says: "…artificial intelligence" includes the following: Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets. (1) An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action. (2) An artificial system designed to think or act like a human, including cognitive architectures and neural networks. (3) A set of techniques, including machine learning, that is designed to approximate a cognitive task. (4) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting."

2. Overview of Trustworthy AI Principles

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, that the results from many modern AI tools are not easily explainable introduces additional needs to ensure that such applications are circumscribed by oversight and accountability systems that mitigate risks [14,16]. While many decision-makers might be comfortable with known risks that can be managed, a major challenge with AI models is that they present a set of "unknown unknowns" – that is, a new set of risks that cannot be easily characterized and could threaten the efficacy and functionality of the entire enterprise [17].

For example, 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 hospital setting, leaving them vulnerable to serious health complications [18]. In other contexts, AI image recognition tools have exhibited differential performance based on skin color [19]. 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, although we recognize that there is a spectrum that trustworthy AI exists on. 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…"

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.

Several trustworthy AI frameworks are relevant to the mission and operations of federal agencies:

- 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.
- Blueprint for AI Bill of Rights: The 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.
- Executive Orders 13985 and 14091*: 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 [20]. 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.

The emergence of many frameworks and overarching principles has led to customization within various federal agencies. For example, the Department of Defense launched their own set of ethical principles to steer the design, development, procurement, and deployment of responsible AI systems in the context of national defense [21]. Similarly, the VA Data Ethics Framework, also known as the VA Data Ethics Final Rule (see 38 CFR 0.605) articulates a comparable set of requirements that help ensure the use of Veteran data is safe, fair, and effective. These agency-specific requirements are important, but the lack of uniformity makes it difficult for practitioners to take theory to practice. To provide

greater structure to these varying frameworks, we compare them according to similarities in the language they use to describe trustworthy AI principles to construct an agency-specific trustworthy AI framework, but our approach and results can also be abstracted to organizations more broadly outside the government.

To fulfill E.O. 13960 requirements, an agency 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 stakeholders. An agency-specific trustworthy AI framework will position the agency to monitor AI activities effectively across divisions and offices as capabilities and circumstances evolve.

As such, an agency-specific trustworthy AI framework serves as:

- A reference document for ensuring the agency satisfies E.O. 13960 consistency requirements, in addition to including other trustworthy AI frameworks impacting or informing the agency's mission;
- A foundation for implementation of E.O. 13960 Section 8 compliance activities as coordinated by the agency research and data governance bodies; and
- An agency-wide consensus statement on trustworthy AI values.

Our development of a streamlined taxonomy for trustworthy AI complements research work by MITRE who similarly argue that a shared methodological approach and taxonomy is required for consensus building [22]. In particular, they argue that AI models must be useful, safe, accountable and transparent, explainable and interpretable, fair, secure and resilient, privacy-enhanced, and underscore that the only way to achieve these aims is by creating greater standardization in the federal government to help define successful oversight and governance in the private sector more broadly.

3. Constructing an Agency-Specific Framework: The Proposed VA Trustworthy AI Framework

3.1. Overview

The framework described here addresses the requirements of E.O. 13960 (see Fig. 1), 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 (see Table 1). 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 the federal government'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.

Table 1
Corresponding sections in existing frameworks for purposeful principle.

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

The selected frameworks represent all federal agency Trustworthy AI Frameworks that were publicly available at the time of writing, federal government-wide documents that describe TAI principles (i.e., EOs and the Blueprint for an AI Bill of Rights), and one exemplar TAI framework from beyond the US (i.e., the OECD principles). The number of Trustworthy AI frameworks in existence beyond this set is large and constantly growing, so we chose to focus on a narrow set of frameworks with applicability specific to federal agencies (plus one additional non-federal framework for comparison).

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

3.2. Purposeful

3.2.1. AI technologies are used to provide clear benefits to the public 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, although the purpose applies more generally to the public, according to the GAO AI Accountability Framework. In line with the OECD's principle of "Inclusive growth, sustainable development and well-being", all AI applications should be accompanied by concrete metrics describing their proposed benefits against which they can be measured after deployment. These outcomes should prioritize well-being and consider AI's potential to reduce health disparities among disadvantaged populations. Recognizing that there are a range of interpretations behind "minimal risks," we encourage the forecasting of expected benefits and costs, including comparisons with non-AI alternatives.

Fig. 2 summarizes the relevant frameworks. Of note, AI that utilizes

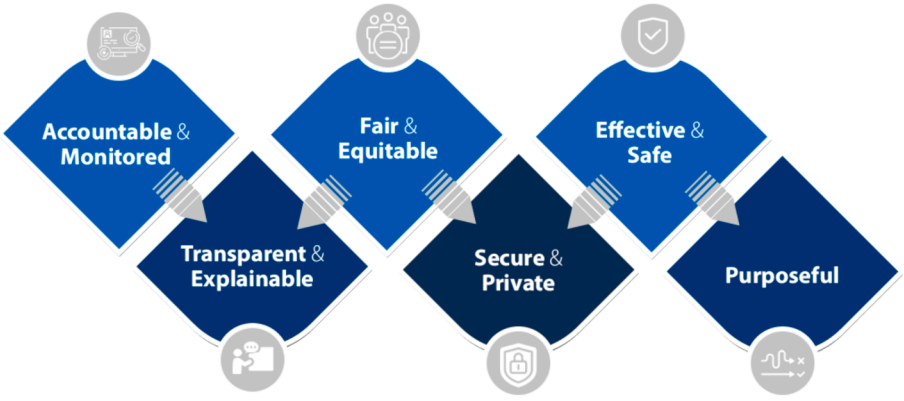

Fig. 1. The Trustworthy AI Framework. The framework described below consists of six principles, which are illustrated in the outer hexagons in the figure above. These principles were selected and refined by examining relevant existing trustworthy AI frameworks and aligning elements to the mission and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

VA Trustworthy AI Framework	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	
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	
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.				
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

Fig. 2. The Proposed Trustworthy AI Framework principles mapped back to principles in existing frameworks. The table above illustrates the relationship between trustworthy AI principles (top colored row) and principles in existing frameworks (lower rows). Brief definitions of 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.

public data should convey a clear benefit to the public, and AI that is the subject of federal research should address a clear need in one of the enumerated areas.

3.3. *Effective & Safe*

3.3.1. *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 the public*

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 [23]. In healthcare settings, efficacy of systems is vital to ensuring safety. Systems that provide diagnostics or other healthcare services cannot be safe if they are not accurate and reliable, since mistakes may directly affect a patient's health. To reflect this, we 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 administrators and leadership who may be considering the use of AI to augment decision-making or assist with procedures. Haupt and Marks argue that the recommendations need not have a transparent or causal interpretation, as long as they are reliable, but we suggest that even recommendations that are accurate in the short-run, but not transparent, may not stay accurate as the checks and balances on model evaluation weaken. One barrier has been that many AI models generally do not yield a single causal variable, instead generating predictions that are hard to interpret or explain. Although there has been a lot of recent work to obtain causal interpretations [24], there should be scrutiny to identify causal sets of variables and rule out the possibility that unobserved determinants of the outcome of interest are spuriously driving the observed phenomena. 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 – at least above current standards of practice – do so reliably and consistently, 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 [25]. 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 the federal government will monitor systems to ensure they meet these criteria (see Table 2).

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 considered in the monitoring process. Of note: (i) systems should be effective based on intended use, (ii) systems should function accurately, reliably, and robustly across their lifespans, (iii) function safely across their lifespans, (iv) risks are proactively identified and mitigated, and (v) safety should be monitored with an eye to changing circumstances and technologies (see Table 3).

Table 2
Corresponding sections in existing frameworks for effective & safe principle.

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

Table 3
Corresponding sections in existing frameworks for secure & private principle.

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

3.4. Secure & private

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

Security and privacy are closely linked, especially in a healthcare 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 [26]. Security protects data and systems from outside risks. Privacy ensures that the collection and use of data does not lead to exposure of sensitive information that jeopardizes the agency or its stakeholders, especially vulnerable populations.

Responsibilities for the protection of privacy in healthcare 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. Of note: (i) systems should be designed to function securely across their lifespans, (ii) systems should be resilient in the face of realized risks and changing circumstances, (iii) handled in alignment with existing VA Data Ethics Framework, (iv) use of systems should remain consistent with Constitution and privacy law, (v) usage of privacy-preserving methods, such as the possible use of synthetic data and zero-knowledge proofs, and (vi) data not used beyond intended purpose.

3.5. Fair & Equitable

3.5.1. Manage and monitor 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 [27]. Additionally, we follow equity as defined 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 healthcare are well documented, such as the underdiagnosis of heart attacks in women, and the decreased access to pain management for Black patients [28]. Lack of diversity in clinical trials has likewise been a long-running issue, on which NIH is now taking action [29]. Without attention to the root causes behind existing variation in the data, AI models will learn inaccurate associations between certain characteristics and health outcomes, propagating inequalities. Another, now seminal example is a large-scale application of AI deprioritized Black patients for delivery of healthcare services because the AI was trained to predict future healthcare costs rather than health needs and outcomes. Due to existing disparities in access to care, this approach improperly conflated data describing ability to pay with relevant diagnostic information [30]. 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. Of note: (i) AI activities should be lawful and respectful of our Nation's values, including Constitutional rights and civil rights laws, (ii) bias should be identified, assessed, and managed throughout the lifecycle of the technology, (iii) stakeholder consultation encouraged; diversity of input is vital, and (iv) follow E.O. 13985 requirements for advancing equity. Table 4 summarizes the main points below.

3.6. Transparent & explainable

3.6.1. Citizens should know when AI systems are used and what data is used by those systems. The government provides straightforward information on how AI systems work and are used to make healthcare decisions

Transparency is the ease with which relevant parties can see how and why AI is being used. To build trust, stakeholders should understand

Table 4
Corresponding sections in existing frameworks for fair & equitable principle.

E.O. 13960	Lawful and respectful of our Nation's values (including civil rights and liberties) [Section 3(a)]
E.O. 14091	When 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 [Section 4(b)] Ensure that their respective civil rights offices are consulted on decisions regarding the design, development, acquisition, and use of artificial intelligence and automated systems [Section 8(b)]
VA Data Ethics Framework (38 CFR 0.605)	Equity [Sect(c)(2)] Reciprocal obligation to Veterans [Section (c)(6)]
White House Blueprint Bill of Rights	Freedom from algorithmic discrimination [Section 2]
OECD AI Principles	Human-centered values and fairness [Section 1.2]
NIST AI RMF	Fair – and bias is managed [Section 4.3]
GAO AI Accountability Framework	Bias: identify potential biases, inequities, and other societal concerns resulting from the AI system [Section 3.8]
DoD AI Ethical Principles	Equitable: Take deliberate steps to minimize unintended bias [Section 2]

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 (see Table 5).

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 practice 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 same logic flows through to staff who are generally drivers behind the adoption (or lack thereof) of AI at a hospital wide level. 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.

Some policy frameworks have bundled explainability and interpretability together. We follow Rudin et al. in defining interpretability

Table 5
Corresponding Sections in Existing Frameworks for Transparent & Explainable Principle.

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

as follows: "An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain." [31] Of note: (i) informed consent is vital to ethical research and clinical practice, and consent cannot be informed if the patient does not understand the materials provided, so information should be made available in an understandable format, (ii) users should be informed of the reason for use of the system and the way in which the system operate, (iii) stakeholders have access to information about the system in use, including monitoring and correction, (iv) information should be presented in an accessible manner, and (v) all relevant parties should understand what data is being used, and how it is being used.

3.7. Accountable & monitored

3.7.1. *The federal government promotes a culture of responsibility and learning across the AI lifecycle, using logging, analytics, and automation to minimize uncertainty about AI operations. AI is deployed in line with existing USG frameworks, such as quality improvements and institutional review boards (IRBs) and informed consent, and the existing VA data Ethics framework. Human fallbacks and monitoring are provided, where appropriate*

Accountability is emphasized in several federal 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. Many AI related projects may be pursued under the banner of quality improvement within an agency, but that still requires monitoring and evaluation to ensure that it is having the intended effects or where pivots are required.

Although specific to the VA, the VA Data Ethics Framework requires Veterans be given a meaningful choice about the use of their data, and the Blueprint for an AI Bill of Rights recommends giving people an 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 healthcare, this principle is constructed with the understanding that AI utilization will adhere to the already established requirements at its host organization for informed consent, whether they be institutional review board (IRB) or patient care requirements. In fact, companion work of ours has introduced an IRB module for streamlining AI related projects and ensuring that they adhere to the trustworthy AI principles [32]. 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 federal agencies. Of note: (i) systems should be regularly monitored, (ii) clear lines of accountability should be established for all AI programs used by the federal government, and (iii) AI use should adhere to existing rules, regulations, and law as appropriate, especially regarding informed consent for treatment and medical research (see Table 6).

Table 6
Corresponding sections in existing frameworks for accountable & monitorable principle.

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

4. Application to large language models

There has been a surge of interest in large language models (LLMs), such as ChatGPT and DALL-E, which are large neural network models designed to process sequential data trained on a large corpus of data (e.g., images or text) [33]. Using Google Trends, we find that the search intensity of it spiked from zero to 100 (on a scale from 0 to 100) over the past year. To put it in perspective, search intensity for "gas price" averages around 20 and only spiked to 78 at the height of the energy price volatility in March/April of 2022. The surge in interest reflects a broader zeitgeist around sophisticated applications of AI that can resemble work traditionally done by humans, although economic estimates suggest that roughly "80 % of the U.S. workforce could have at least 10 % of their work tasks affected by the introduction of LLMs, while approximately 19 % of workers may see at least 50 % of their tasks impacted." [34].

Although there has been substantial research on the effects of automation on productivity, very little exists on generative AI. To our knowledge, the only, and very recent, study suggests that generative AI might have important productivity effects: Brynjolfsson et al. conduct a randomized experiment with a Fortune 500 firm where 5,179 customer service representatives were randomly assigned the use of large language models (LLMs) to aid performance with subsequent feedback from their managers [35]. They found that LLMs led to a 14 % improvement in cases resolved, but these effects were concentrated among the less experienced workers, suggesting that generative AI may have an augmentation effect.

We now apply our harmonized framework for trustworthy AI around a specific example, namely the use of generative AI tools in health informatics, like ChatGPT. The use-case we have in mind is one where an LLM is used to help field patient inquiries about their health before talking with a clinician. We enumerate each of the harmonized principles and apply to them to this generative AI setting.

- Purpose: Has the model been trained specifically for the set of possible patient inquiries, or is it a general-purpose tool trained off a larger dataset that may not be as informative for the population of interest? Has the scope of its application been well-defined, or is there a risk that users will want to rely on it for activities outside of its original scope?
- Effective & Safe: Has the tool's performance been benchmarked on its intended application, i.e. against human patient support systems? What common failure modes exist? What is the role of the human-in-the-loop when receiving outputs from an LLM, or is it fully automated?
- Secure & Private: How are patient prompts that are fed into the model stored? What sensitive data was used to train the model? Who has access to the model/where is the model hosted?
- Fair & Equitable: How does the model perform not only overall, but also separately by different partitions of the data, including race, gender, and socioeconomic status? Does the training data set resemble the population to which the model will be applied?
- Transparent & Explainable: Do those affected by the AI recommendations and diagnosis receive notice that an LLM was involved in the decision-making process? Is there a feedback process that allows a patient to talk with a human if they do not understand or trust the AI response?
- Accountable & Monitored: Is model performance continuously charted, and by whom? Is proper training provided to model users? How do users document their use of the model?

There are many other possible use-cases for LLMs and our example is meant to be illustrative, not exhaustive. For example, clinicians could also use an LLM as a diagnostician to help clinicians connect patient symptoms with possible diagnoses and recommendations. Similarly, LLMs could function as a clinical administrative assistant, such as scanning notes into a database or ordering tests for a patient. In sum, our trustworthy AI framework provides a process for managing risk around AI systems, particularly with the proliferation and surge of interest in generative AI tools within health informatics.

5. Conclusion

Numerous frameworks have emerged in recent years to promote the ethical use of AI. However, despite the abundance of these frameworks, there currently needs to be a common taxonomy for effectively relating these principles to practical behaviors. This poses a significant challenge, particularly within the federal government, where AI systems are extensively employed. The need for clear, actionable, and measurable guidance is even more pressing in the healthcare landscape of the Department of Veterans Affairs, given the potential for serious harm to millions of patients in the event of errors.

This paper addresses the gap by consolidating policy guidance on trustworthy AI and connecting each principle and specific actionable steps from the perspective of the federal government. Acknowledging the importance of contextualization and implementation of these principles within individual agencies, we offer a valuable blueprint for other domestic and international counterparts navigating the realm of trustworthy AI. By presenting a harmonized approach to policy guidance, this paper not only aims to support the development and deployment of ethical and reliable AI systems across different sectors and organizations but also to pave the way for a more promising future of AI implementation.

We nonetheless recognize that there are several limitations that make space for future work. First, there are many more frameworks than those that we have surveyed here, including nonprofit organizations and others in the private sector who have established taxonomies. Second, although there is recent evidence that industry has played an increasing role in AI [36], they are relatively less engaged on responsible AI research [36,37]. This elevates the importance of establishing a common taxonomy that fuels additional research and translation to practice. Third, we need a better understanding of the empirical consequences of practical implementation details. For instance, Makridis and Mueller et al. [32] explain the results of a pilot within the Department of Veterans Affairs where a supplement to the IRB module was created for AI use-cases. We leave these topics for future research.

CRediT authorship contribution statement

Joshua Mueller: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christos A. Makridis:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Theo Tiffany:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Andrew A. Borkowski:** Writing – review & editing, Supervision, Project administration, Conceptualization. **John Zachary:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Gil Alterovitz:** Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.hpopen.2024.100128.

References

[1] Brynjolfsson E, Rock D, Syverson C. The productivity J-curve: How intangibles complement general purpose technologies. Am Econ J Macroecon 2021;13(1):333–72.
[2] Makridis CA, Mishra S. Artificial intelligence as a service, economic growth, and well-being. J Serv Res 2022;25(4).
[3] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. N Engl J Med 2023;388:13.
[4] Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. Nat Med 2022;28:31–8. https://doi.org/10.1038/s41591-021-01614-0.
[5] Zhou D, Tian F, Tian X, et al. Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. Nat Commun 2020;11:2961. https://doi.org/10.1038/s41467-020-16777-6.
[6] Courtiol P, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med 2019;25:1519–25.
[7] 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):e353–62. https://doi.org/10.1016/S2589-7500(19)30159-1.
[8] Henry KE, 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 2022;28:1447–54. https://doi.org/10.1038/s41591-022-01895-z.
[9] Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572:116–9. https://doi.org/10.1038/s41586-019-1390-1.
[10] Wismüller A, 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. In: 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.
[11] Silcox S, Dentzer S, Bates DW. AI-enabled clinical decision support software: a "trust and value checklist" for clinicians. NEJM Catalyst 2020;1(6).
[12] Dorr DA, Adams L, Embi P. Harnessing the promise of artificial intelligence responsibly. J Am Med Assoc 2023.
[13] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman-Klein Center for Internet & Society.
[14] Lee et al. (2023), ibid.
[15] Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care. Health Aff 2014;33(7).
[16] The domain of explainable AI is advancing rapidly in real time, but currently not all AI models can be interrogated in sufficient detail to satisfy some notions of explainability/interpretability. As such, it is roughly accurate to note that explainability may not be attainable in all cases. However, in such cases, clear documentation and communication about the limits may be an appropriate mitigation for the lack of explainability, hence the pairing of Explainability with Transparency in the VA TAI Framework.
[17] This phrasing comes from a famous quote by former Secretary of the U.S. Department of Defense Donald Rumsfeld who remarked: "Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones." An analogous application exists with AI.
[18] 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.
[19] 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.
[20] U.S. federal agencies that design, develop, acquire, or deploy AI systems are responsible for all aspects of the system's performance across its lifecycle. To support AI system maintenance, the Government Accountability Office, an independent, non-partisan agency that works for Congress and provides federal agencies with objective, non-partisan, fact-based information to help the government save money and work more efficiently, developed this resource to identify key practices to help ensure accountability and responsible AI use by federal agencies.
[21] https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF.
[22] Mitre. Blueprint for trustworthy AI implementation guidance and assurance for healthcare. Coalition for Health AI, version 2023;1.
[23] See under "safety" at ISO/IEC TS 5723:2022(en), Trustworthiness — Vocabulary.
[24] Cui P, Athey S. Stable learning establishes some common ground between causal inference and machine learning. Nat Mach Intell 2022;4:110–5.
[25] There is an active debate about whether AI needs to exceed or achieve the same level of standard as humans. For example, AI could be required to exceed the top 75th percentile of human effort. For example, Ware et al. (2024) study this question from the perspective of robotics augmenting clinical capabilities for hernia repairs. They find that robotic assistance improves patient outcomes. In this sense, there is a "pareto improving" outcome where quality and scale of care for hernia repair improved. We do not take a stance on this broader question, but defer to individual sub-fields to determine what is most appropriate to them.
[26] https://privacyrights.org/data-breaches.
[27] 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.
[28] Starke G, De Clercq E, Elger BS. Towards a pragmatist dealing with algorithmic bias in medical machine learning. Med Health Care and Philos 2021;24:341–9. https://doi.org/10.1007/s11019-021-10008-5.
[29] Shamoo AE, Resnik DB. Responsible conduct of research. Oxford University Press; 2009.
[30] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447–53.
[31] Rudin, C. et al. 2021. Interpretable machine learning: Fundamental principles and 10 grand challenges. https://arxiv.org/pdf/2103.11251.pdf.
[32] Makridis CA, Boese A, Fricks R, Workman D, Klote M, Mueller J, et al. Informing the ethical review of human subjects research utilizing artificial intelligence. Front Comput Sci 2023;14(5).
[33] Bubeck et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://arxiv.org/abs/2303.12712.
[34] Eloundou T, Manning S, Mishkin P, Rock D. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. Science 2024;384(6702):1306–8.
[35] Brynjolfsson, E., Li, D., and Raymond, L. 2023. Generative AI at work. NBER working paper 31161.
[36] Ahmed N, Wahed M, Thompson N. The growing influence of industry in AI research. Science 2023;379(6635):884–6.
[37] Ahmed, N., Das, A., Martin, K., and Banerjee, K. (2024). "The Narrow Depth and Breadth of Corporate Responsible AI Research." https://arxiv.org/abs/2405.1219.