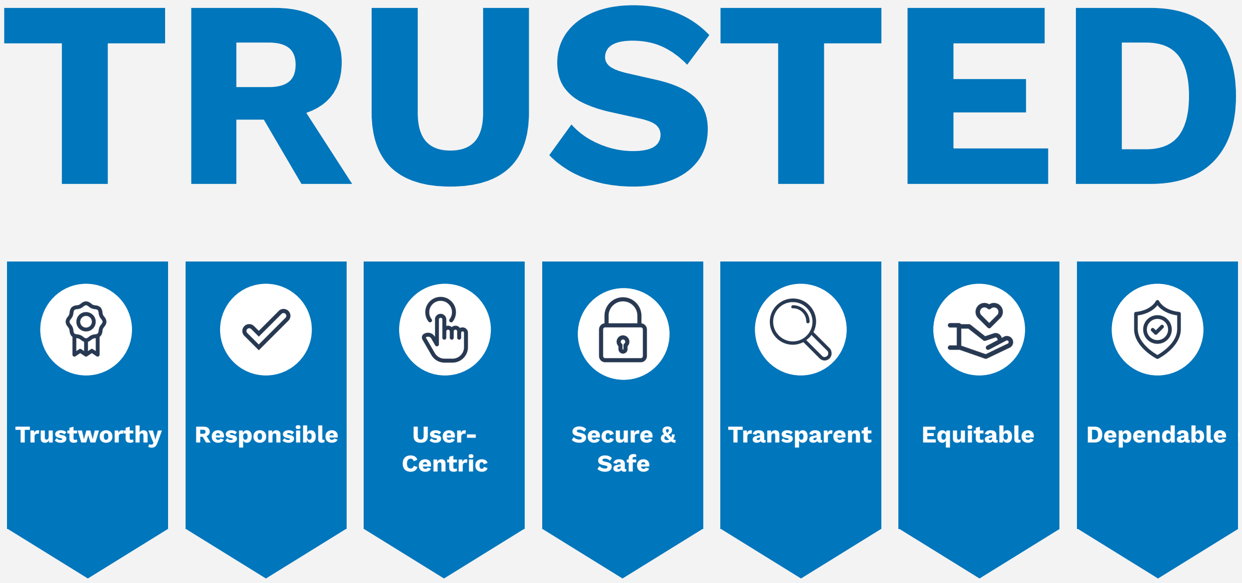
avia Generative AI Collaborative

TRUSTED AI Framework


TRUSTED

Trustworthy

Responsible

User-Centric

Secure & Safe

Transparent

Equitable

Dependable

Background: AI in healthcare has immense potential but also risks of adverse, unexpected, and/or negative outcomes, requiring a framework focused on health impact, fairness, ethics, and equity to ensure it benefits to all impacted stakeholders. Without guidelines, distrust of AI increases among providers and patients who cannot easily assess algorithm robustness or a health system's development process. A responsible governance approach is urgently needed to maximize AI's benefits in healthcare while proactively managing its risks.

With many frameworks out there, AVIA aimed to provide a synthesis of their most important parts in a digestible, translatable way.

Objective: This framework is meant for provider organizational governance bodies and committees to set a high-level guidance on the factors that may be informed and/or addressed to ensure safe, ethical, and effective use of AI in lieu of regulatory clarity.

How to use this framework: This framework should be considered throughout the development of AI capabilities and serve as a foundational mechanism to ensure health systems are addressing the key items and taking the appropriate steps toward responsible use of AI.

Inspired and influenced by the NIST AI Risk Management Framework, The White House's Blueprint for an AI Bill of Rights, the CHAI Blueprint for Trustworthy AI Implementation Guidance And Assurance For Healthcare, the CMS AI Playbook and the European Commission's Ethics Guidelines for Trustworthy AI

TRUSTED AI Framework

T - Trustworthy: Refers to the clarity and openness with which operations, algorithms, and data usage are communicated. For AI solutions, this would mean understanding the output of the AI (i.e., decision, recommendation, summary) what data, underlying algorithm or input it uses, and any potential biases in its training data

- **Validation[c,d] & Verification[a,b]:** Before deploying any AI system, health systems should conduct necessary, rigorous testing in real world scenarios (real use application vs. competency testing) to verify its reliability and validate its outcomes. Continuous monitoring should also be employed post-deployment to check for deviations and ensure any updates are deployed.

- **Stakeholder Feedback[a,b,d]:** AI tools hold the potential to alleviate staff burdens, not add to them. Regularly gathering feedback from patients, healthcare providers, and other stakeholders is key to understanding their trust levels to address and improve AI models.

- **Bias Detection & Correction[c,d]:** Implement systems to detect and correct biases in AI models, ensuring they provide equitable care to all patients. Analysis during development and regular audits may uncover biases such as systemic bias, computational and statistical bias or human-cognitive bias. Techniques like reweighting data or tweaking algorithms can help correct or reduce such biases.

R - Responsible: Entails taking ownership of the AI's actions, decisions, and any consequences (intended or unintended) resulting from its recommendations or analyses. This also includes providing appropriate human oversight.

- **Accountability[a,c,d]:** Health systems must establish a feedback system to hold responsible parties accountable for the decisions made by the AI, especially when errors occur. Key decisions should be documented along with comprehensive logs to enable auditing. Once audited, results and actions should be shared with responsible parties and leaders.

- **Ethical Development[a,c,d]:** Ensure that AI is developed with consideration to ethical implications, particularly regarding potential biases and fairness. To ensure equity, utilize multidisciplinary teams to review and approve use cases, test and train models appropriately and monitor model performance from cradle to grave. Ethical frameworks can help focus efforts on core principles and conduct impact assessments to guide the team to make responsible design choices and provide transparency to the public.

- **Human Oversight[a,b]:** Keep qualified healthcare professionals in the loop to review AI recommendations, override incorrect or unsafe decisions, and request explanations. Cultivate human-AI collaboration through workflows that empower clinicians to evaluate AI outputs based on their expertise before operationalizing.

[a] NIST AI Risk Management Framework
[b] The White House Blueprint for an AI Bill of Rights
[c] CHAI Blueprint for Trustworthy AI Implementation Guidance And Assurance For Healthcare
[d] CMS AI Playbook

TRUSTED AI Framework

U - User-Centric: Focuses on the ease of integration and use of AI solutions within existing systems, processes, and workflows. It ensures that the tool is intuitive and enhances, rather than disrupts, the user's operations.

- **User-Friendly Interfaces[c,d]:** While UX design is not AI-specific principle, due to their inherent complexity, AI systems should be designed to be easily understood and used by medical professionals and patients without requiring them to be AI experts. Focus on simplicity in design through iterative testing and build in contextual help features to assist users when needed. Leveraging intuitive visuals and natural language will facilitate seamless human-AI interaction.
- **Training & Education[a,b,d]:** For providers and staff, it is essential to provide proper training on how to use the AI tools and interpret their outputs. For patients, training and education should be kept at a minimum to enable utilization. Providing quick reference guides can support sustained usage by staff and patients, if needed.
- **Integration[a,c]:** Ensure AI solutions are seamlessly integrated into existing hospital systems and processes to increase adoption and avoid disruptions. Understanding workflows beforehand will improve change management as integrations may often redefine workflows altogether. and ease transitions and adoption by users.
- **Explainability & Interpretability[c,d]:** AI systems should be designed and developed with explainability in mind, enabling users to understand why certain outputs or predictions were made. Similar but different, systems should be interpretable, meaning their output should fit into the context of their designed functions.

S - Secure & Safe: Addresses the protective measures in place to guard against unauthorized access, data breaches, and other cyber threats. For AI in healthcare, this is of paramount importance given the sensitivity of health data.

- **Data Protection[b,c]:** To enable security, the organization must implement strong encryption, manage and monitor access controls and activity logging to protect data. Ensure robust encryption methods for data at rest and in transit by adhering to regulations like HIPAA or GDPR.
- **Anonymization:** When possible, health systems should remove personal identifiers from data for AI training and operations. Using secure multi-party computation will still enable analysis on anonymized data. Adding differential privacy noise can also reduce reidentification risks.
- **Privacy-Enhanced[b,c]:** Design AI systems to minimize the collection, storage and use of sensitive personal data. Limit data ingestion to only what is needed for the task. Build in data deletion schedules to discard unneeded data and use techniques like federated learning to train models without direct data access.

[a] NIST AI Risk Management Framework
[b] The White House Blueprint for an AI Bill of Rights
[c] CHAI Blueprint for Trustworthy AI Implementation Guidance And Assurance For Healthcare
[d] CMS AI Playbook

TRUSTED AI Framework

T - Transparent: Refers to the clarity and openness with which operations, algorithms, and data usage are communicated. For AI solutions, this would mean understanding how the AI makes decisions, what data it uses, and any potential biases in its training data

- **Algorithmic Transparency[c,d]:** Health systems should be clear about how AI models work, their decision-making processes, the data they were trained on, including the populations for which it may not be ideal. Additionally, organizations should provide clear documentation that explains the AI model features and weighting (i.e., model cards). Visualization tools and model descriptions can improve interpretability by providing insights into model predictions and describing outputs in understandable language.

- **Purpose Declaration[a,b]:** State intended benefits, appropriate use cases and limitations of AI systems upfront. Acknowledging risks in the patient consent process will set proper expectations and help manage inappropriate use by staff.

- **Open Communication[d]:** Establish a protocol for patients, healthcare providers, and other stakeholders to raise concerns and ask questions about the AI systems in use. Establishing feedback channels with developers/solutions will allow stakeholders to voice concerns. Responding to queries in a timely and substantive manner will foster engagement.

E - Equitable: Ensures that the AI solution works fairly for everyone, irrespective of race, gender, socioeconomic status, etc. It emphasizes that AI decisions should be made without unwarranted bias and that everyone should have equal access to its benefits.

- **Inclusive Data Sets[c,d]:** Ensure that the data sets used to train AI systems and models are robust, representative of wholistic characteristics of the targeted populations of use, considering factors such as race, gender, age, socioeconomic status, geographical area and more.

- **Bias and Fairness Assessment[b,c]:** During development, test AI systems with diverse subgroups to uncover biases proactively. After deployment, regularly evaluate AI models for biases that could lead to unequal treatment or outcomes for different patient groups. Conducting root cause analysis will pinpoint needed adjustments to address biases.

- **Accessibility[a,b,c]:** Ensure AI tools and technologies are accessible by all healthcare providers and patients, regardless of their location, physical abilities, preferred language or socioeconomic background. Involving diverse users in the design phase can help identify blind spots related to culture and background and prevent bias. Contextual testing will help spot offensive or alienating content before launch.

[a] NIST AI Risk Management Framework
[b] The White House Blueprint for an AI Bill of Rights
[c] CHAI Blueprint for Trustworthy AI Implementation Guidance And Assurance For Healthcare
[d] CMS AI Playbook

TRUSTED AI Framework

D - Dependable: Refers to the reliability, accuracy, and consistency with which AI systems perform their intended tasks over the life cycle. For health AI, being dependable means correctly interpreting patient data, maintaining robust performance over time, avoiding harmful errors, and operating safely even in edge cases across the duration of use.

- **Robust Testing[a,b,c,d]:** Regardless of built or bought algorithms, health systems should conduct extensive testing of AI systems using real-world clinical data to prevent and limit adverse and unexpected outcomes. Testing should cover normal and edge cases, verify accuracy across subgroups, and proactively identify potential failure modes or biases. For continuous improvement, document testing methodology, data characteristics, and results.

- **Fail-safes[a,b]:** Incorporate checks and overrides that default AI systems to known safe outcomes in case of failures, uncertainty, or suspicious outputs. For high-risk AI use cases, define minimum human validation requirements before any AI decision can influence downstream care. Build in appropriate caveats and alerts to notify human operators of anomalies.

- **Ongoing Monitoring[c] & Maintenance[a,b,d]:** Monitor AI performance on an ongoing basis as new data comes in, tracking key metrics like accuracy, fairness, and safety to ensure algorithms are up to date. Periodically re-validate systems using fresh real-world data to check for performance drifts. Retrain or refine models in a timely manner if monitoring reveals degraded performance. As accountability is best constructed when specific stakeholders are given particular responsibilities in a decision making context, health systems should consider defining clear ownership or delegation of responsibility for monitoring, amending, retraining, and/or decommissioning an algorithm.

[a] NIST AI Risk Management Framework
[b] The White House Blueprint for an AI Bill of Rights
[c] CHAI Blueprint for Trustworthy AI Implementation Guidance And Assurance For Healthcare
[d] CMS AI Playbook