Artificial Intelligence Review Framework for Trustworthy Artificial Intelligence Systems[1]

This document provides a framework for reviewing proposed artificial intelligence (AI) systems to determine if a tool or system is appropriate for use at the University. The table provides a list of key principles to support ethical AI use as well as points to consider in evaluating whether a tool meets university standards for trustworthy AI. The weight of each principle will vary given the context in which the tool is expected to be used as well as the maturity of the AI system (development vs testing vs implementation). Nonetheless, the general principles, even where aspirational in the case of AI development, should be considered for each AI system reviewed, whether provided under contract to the University or otherwise made available to the Yale community.

The framework is meant as a starting point to guide review, taking into account the context and particulars of a given AI system. The framework incorporates key aspects of current AI governance principles. Given the rapid evolution of the AI landscape, users may want to consider incorporating additional discipline specific frameworks as they emerge to facilitate a thoughtful review of AI systems, balancing the opportunities and risks of AI systems.

Lastly, in addition to assessing risks related to erroneous output, bias, privacy, and security as described in the framework, review of AI systems should consider their impact on human capability and judgment over time. Evaluators may consider whether the system is likely to support skill development, critical thinking, and independent decision-making, or whether it may inadvertently encourage over-reliance that diminishes user capacity. For educational, research, and clinical contexts especially, consideration should be given to preserving appropriate human expertise, maintaining opportunities for unassisted practice, and ensuring that AI augments rather than displaces the formation of professional judgment.

For the purpose of this document, the following definitions apply:

- AI System: a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI Systems vary in their levels of autonomy and adaptiveness after deployment.[2]
- Data Subject: an identified person whose Personal Information (data) is collected, stored, or processed by an AI System

[1] This framework does not supersede any applicable mandatory review by oversight committees such as the Yale Human Research Protection Program, YNHHS Enterprise Healthcare AI Governance Committee for clinical systems, Data Governance Executive Council, AI Governance Committee, etc.
[2] This document adopts the definition of AI Systems developed by the Organization for Economic Co-operation and Development. https://oecd.ai/en/ai-principles

- Personal Information: any information relating to an identified or identifiable Data Subject. Note that "identifiable" includes the ability to be identified both directly as well indirectly such that inclusion of sufficient information may render data without direct identifiers to be identifiable.

Principles	**Considerations**
Risk tolerance correlates to the potential consequences of harm arising from erroneous output	Has an impact assessment been performed to determine critical risk areas? What is the sensitivity and data classification[3] of the data to be input into the AI system? Is there the potential for harm if the input data is disclosed inappropriately by the AI system? How damaging would erroneous output from the AI system be to a data subject? Does the AI output impact consequential decisions that pose high risk (clinical care, hiring, admissions, etc.)? Are there any existing or pending state or local laws that apply to the AI system?
Free from bias and discrimination	Has the AI system been assessed for bias against subsets of data subjects for the proposed use case? Was the AI system developed using a training data set with appropriate representation for the proposed use case? Were the training data sets independent of test sets? Will there be human oversight to avoid discriminatory automated processing? Are there any demographic groups, including marginalized groups, that may be vulnerable to bias by the AI system and if so, how will this be addressed? Does the AI system meet the university's accessibility standards? How might system users impact the validity of the output and how might the use of the AI tool impact users' performance?
Transparent and explainable	Are the outputs of the AI System intelligible and explainable? Are those aspects of a process that involve an AI system disclosed to users including information about who may have access to their information and whether it is retained by the AI system?
Accountable	Is the system owner committed to adhering to good documentation practices for continuous testing and performance evaluation throughout the AI system's lifecycle (e.g., ALCOA+, model cards/datasheets)? What role will users play in assuring trustworthiness of output?

[3] See https://cybersecurity.yale.edu/know-your-risk

	Continuous performance evaluation: • What plans are in place and who will be responsible to assure that the AI system remains trustworthy over time and does not drift from initial risk assessments? • Will the users be informed about any corrective actions implemented as a result of the continuous performance evaluation? • Are there plans to assess user satisfaction and system success over time?
Reliability and Reproducibility	Does the AI system provide consistent output as intended in its design? Does the AI system have the potential to provide any false or misleading information and if so, what risks would that pose? Has the tool been validated for accuracy and assessed for frequency of hallucination? Has the vender (where applicable) provided information on their testing or has the tool been rated on credible AI leaderboards? If the model has a "human in the loop," was the testing focused on the performance of the human-AI team rather than just the performance of the AI in isolation? Does the AI system meet standards for research reproducibility relative to the disciplinary contexts of the use case proposed?
Fit for purpose	How has the AI system been assessed for efficacy in the specific use case at hand? Is the AI system being used for a purpose beyond what it was designed to do? Was the AI model trained on data appropriate for the proposed use and is it sufficiently representative to handle edge cases? What training and continuous support will be available to the intended AI users?
Safeguard data privacy	Is the data collected by the AI system or used to train the system provided with consent or with the awareness of the data subject? Are users informed that they are interacting with an AI system? Are data subjects made aware that their personal information will be analyzed by an AI system? Has the system been assessed to ensure that personal information will not be inappropriately disclosed? Is the Personally Identifiable Information (PII) collected for the AI system directly related to university functions or activities and is it necessary for that purpose?

	Are the mechanisms to allow a data subject to have their data removed, corrected, or otherwise exercise their privacy rights?
Purposeful	Does using the AI system advance the university mission? Is the AI system's proposed use consistent with university policies as well as university core social norms and values? Does the AI system pose risk of unintended consequences or misuse that threaten university values? If the AI system is provided by an external vendor, do their values align with the university's and could those values impact the trustworthiness of the AI system?
Data security	Does the AI system protect against external threats to its functionality or data including protection for access, corruption or deletion? Does the AI system comply with university minimum security standards based on the risk classification of the intended data inputs?
Sustainability	Are there sufficient resources (technical, training, financial, staffing, etc.) to implement and maintain the trustworthiness of the proposed AI system? Are there aspects of the AI system or associated data feeds that may create limitations on our ability to use the system over time? What evaluation has been done of the potential impacts of the AI system for its proposed lifecycle? What impact would it have if users were suddenly cut off from the AI system after adopting it?
Human oversight	Are any of the processing or decisions automated without review by the user? To what extent will the output be evaluated by a human in the loop? If inaccurate, could the output of the AI system cause significant harm that would necessitate continuous oversight?

Risk Based Framework Application:

AI systems are varied and the uses of these tools continue to evolve and expand. Consideration should be given to the risks associated with any given use case such that high risk activities undergo more rigorous review than those posing only minimal risk. High risk activities include those that have a consequential impact on individuals such as employment decisions, medical care, biometric identification for law enforcement, etc. In some cases, use of AI in high-risk activities would be at odds with university values and therefore pose an unacceptable risk.

It is expected that most AI systems would be considered low risk and the risk associated with a system that was not trustworthy would only have limited impact on data subjects. Low risk activities include AI-enabled spam filters and chat bots that facilitate access to publicly available information.

Between these extremes are systems that may not be perfect but may be acceptable to use if appropriate controls are implemented, such as providing notice to users about system limitations, or limiting the sensitivity of data that can be used in a given AI system. Ultimately, whether an AI system is appropriate requires balancing system trustworthiness as assessed using the principles described here with the usefulness for productivity, education, research, clinical care and other legitimate university needs.

Application of principles throughout the AI tool cycle

The following examples and case studies are illustrative. Appropriate application of the principles varies case-by-case, depending on the use, attendant risks, and other contextual factors.

Principle	Description	AI development	AI testing	AI Implementation	AI maintenance
Risk tolerance correlates to the potential consequences of harm arising from erroneous output	AI systems should be designed, deployed, and governed with a level of risk tolerance proportional to the severity and likelihood of harm that may result from incorrect, misleading, or unintended outputs. Lower risk tolerance is warranted for higher-risk use cases when erroneous outputs could result in significant harm, requiring stricter controls, validation, and oversight.	Identify potential risks and harm scenarios. Develop models, algorithms, and potential safeguards that minimize the possibility of high-impact errors. To determine the risk levels, consider intended use, affected stakeholders, and potential harm. Align system design choices (e.g., model complexity, safeguards) with the assessed risk level.	Rigorously test AI outputs under various potential scenarios to assess risk levels and the consequences of erroneous outputs (with higher-risk systems undergoing more rigorous validation and scenario testing). Use stress testing[4], edge-case analysis, and failure-mode evaluations proportionately to the risk.	Implement deployment controls (e.g., phased rollout, access restrictions, human review) and risk management strategies reflecting the acceptable risk tolerance for the use of the system.	Continuously monitor and update risk assessments, refining the AI system to reduce risk over time. Update mitigation measures when system behavior, data, or context changes.
Free from bias and discrimination	AI systems should be designed, developed, and operated to avoid inherent biases and prevent discriminatory outcomes, ensuring fair treatment for all individuals and groups. This principle emphasizes the need for algorithms and data to be scrutinized and adjusted to prevent perpetuating or amplifying societal biases and discrimination.	Ensure diverse and representative system training data. Incorporate fairness metrics[5] and bias mitigation techniques into the model selection and engineering.	Evaluate the system for biases and discriminatory outcomes through comprehensive testing with diverse datasets.	Implement continuous monitoring to identify and address biases in the AI system's decisions and outputs. Train users on spotting potential discriminatory use and on appropriate interpretation of outputs.	Conduct regular bias audits to detect drift or emerging inequities. Regularly update and retrain the AI with new data to ensure ongoing fairness and the minimization of biases.
Transparent and explainable	AI systems should be easily understandable to relevant stakeholders, with clear reasoning behind their decisions or outputs. Transparency involves openness about how the AI operates including its data sources and limitations, while explainability refers to the ability to	Develop algorithms that can provide explainable outputs. Document the system's purpose, decision-making logic, assumptions, and limitations, and the	Test for explainability by having non-experts (but intended stakeholders) understand the AI's decisions through generated explanations.	Prepare and provide users with clear documentation and interfaces that explain AI decisions in an understandable manner.	Determine frequency of reviewing existing training materials to ensure they are up to date. Update explanations and offer retraining based on new findings (e.g., system changes) and user feedback to

[4] For more information on stress-testing, see https://ankura.com/insights/stress-testing-ai-models-a-modern-imperative-for-model-risk-management .
[5] https://shelf.io/blog/fairness-metrics-in-ai/

	provide understandable insights into the decision-making processes of the AI.	system development process.			maintain transparency over time.
Accountable	The principle of accountability ensures that mechanisms are in place to hold individuals or organizations responsible for the AI system's design, deployment, operation, and impacts. This includes traceability and the ability to audit AI systems effectively.	Define roles and responsibilities for each part of the AI system including design, data sourcing, and model development. Ensure logging and traceability of actions and decisions.	Establish mechanisms to trace the source of errors and hold parties accountable. Establish and document process for validation, approval, and sign-off.	Assign operational ownership, including responsibility for monitoring, incident response, and compliance. Implement assurance plan - oversight mechanisms for accountability, such as regular audits and performance reviews.	Maintain detailed logs of AI operations and decisions for future reviews and accountability checks. Document updates and issue remediation and the responsible party.
Reliability and Reproducibility	AI systems should perform consistently, accurately, and dependably within defined parameters. This requires rigorous testing, validation, and monitoring to ensure the system meets performance standards and maintains reliability over time.	Design robust algorithms capable of performing well under varied conditions. Establish reliability requirements (e.g., accuracy thresholds, uptime expectations).	Conduct stress tests, edge case tests, and real-world scenario tests to ensure consistent performance and reliability.	Monitor the AI system continuously to ensure consistent performance and reliability and to promptly detect errors, degradation, or unexpected behavior.	Regularly update and revalidate the AI system to ensure it remains reliable despite changing conditions or data with retraining, recalibration, or fixes applied as needed.
Fit for purpose	AI systems should be appropriately designed and configured to meet clearly defined objectives, and their capabilities and limitations should align with the intended use case and context.	Clearly define the purpose and scope of the AI tool. Design it to meet these specific needs.	Ensure the AI system effectively addresses its intended purpose (and not just technical performance) through focused and realistic testing.	Deploy the AI system only in contexts and restrict use to approved purposes for which it is designed. Customize configurations if needed. Guide users on appropriate application and restrictions.	Adapt and update the AI system as necessary to continue meeting its original and evolving purposes. Continued alignment with purpose is reviewed, especially if organizational needs or external conditions change.
Safeguard data privacy	AI systems should be designed and deployed with respect to individual privacy rights by collecting, processing, and retaining personal data lawfully, minimally, and transparently, in accordance with applicable privacy regulations and standards and by protecting personal and sensitive information from unauthorized access, use, and disclosure.	Implement lawful data sourcing, data minimization, data anonymization, encryption, and access controls from the outset of development (privacy-by-design principle).	Test the robustness of privacy measures and data protection protocols.	Monitor and audit data access and usage to ensure privacy safeguards are adhered to.	Update security measures and privacy protocols to protect against emerging threats.
Purposeful	AI systems should be developed with clear objectives that deliver meaningful benefits	Define and adhere to clear ethical guidelines,	Ensure testing scenarios are aligned with the AI's	Deploy the AI system in a manner that maximizes its	Regularly review the AI's impact to ensure it continues

	to society. The purpose should align with ethical standards and the organizational values, mission, and goals.	organizational values and legal obligations during development.	intended beneficial purpose. Confirm that system behavior supports the stated purpose and does not introduce unintended uses.	intended benefits to users and society.	to serve its intended purpose, continued relevance and appropriateness.
Data security	This principle involves implementing robust measures to secure data against external and internal breaches, theft, and corruption and ensuring the integrity, availability, and confidentiality of data throughout the AI lifecycle.	Implement comprehensive data security practices, such as encryption and secure data storage, during development.	Identify and fix vulnerabilities through penetration testing and other security assessments.	Continuously monitor and protect the AI system and data from potential security threats. During production, enforce security measures such as encryption, authentication, and access management.	Regularly update security measures and protocols to address new threats and vulnerabilities.
Sustainability	Feasibility assesses whether an AI initiative can be practically and successfully implemented, considering technical, financial, operational, and organizational constraints. It involves evaluating the practicality and sustainability of AI solutions including whether there are sufficient resources, infrastructure, and expertise available to support effective and maintainable implementation	Before committing to system development, assess technical, financial, and logistical feasibility to ensure practicality.	Perform feasibility studies to see if the AI meets practical constraints and performs as expected under real-world conditions.	Implement the AI system in a way that aligns with practical constraints and feasible applications. Consider scalability, integration with existing systems, and operational readiness.	Continually assess cost, performance, and resource requirements over time and adapt the AI system to maintain feasibility as technology and operational contexts evolve.
Human Oversight	AI systems should operate under human supervision to ensure monitoring of ethical decision-making and the ability to intervene and override in critical situations.	Integrate mechanisms for human intervention and oversight into the AI design.	Ensure testing involves human oversight to catch errors the AI might not account for. Test not only the performance of the system itself but the human-in-the-loop pair.	Implement policies and systems for regular human review and supervision of AI operations allowing human the ability to intervene, override, or escalate issues when necessary.	Maintain systems for ongoing human oversight, enabling adjustments and interventions as needed.

Case studies

Scenarios	AI development	AI testing	AI Implementation	AI maintenance
Developing and implementing an AI tool for teaching at a University	• The university defines the tool's purpose as providing formative feedback and learning support, not grading or academic decisions. • Training data is sourced from approved curriculum materials and anonymized student examples. • Bias risks are assessed to ensure explanations are accessible to students with different language backgrounds. • Privacy-by-design is applied by minimizing use of personal student data. • Human oversight is designed in: instructors remain responsible for all academic judgments.	• The AI is tested using sample student assignments to verify accuracy, clarity, and reliability of feedback. • Bias testing checks whether feedback quality differs across writing styles or language proficiency. • Transparency is validated by reviewing whether explanations are understandable to students and instructors. • Privacy controls are tested in sandbox environments using de-identified data separate from the training data. • Faculty review outputs before approval for deployment.	• The AI tool is integrated into the learning management system with restricted access. • Students are informed about how the AI works, its purpose, and its limitations. • Instructors are trained to interpret AI-generated feedback and intervene when needed. • Use is limited to approved courses and contexts to ensure fit for purpose. • Ongoing monitoring dashboards are activated.	• System performance is monitored for accuracy and relevance as curricula evolve. • Bias audits are periodically conducted to detect drift. • Privacy and security controls are reviewed and updated. • Instructors provide feedback to refine system behavior. • The tool's purpose and feasibility are reassessed each academic year.
Designing and implementing a clinical AI tool	• The tool is designed to support clinicians, not replace diagnosis or treatment decisions. • Training data includes diverse patient populations to reduce bias. • Risk tolerance is set low due to potential patient harm, leading to conservative model design. • Explainability requirements are defined so clinicians can understand recommendations. • Data security and regulatory requirements are embedded from the outset. • IRB approval (and other institutionally required ancillary approvals) are obtained for human subjects research activities to develop the tool.	• Extensive validation is performed using retrospective clinical data. • IRB approval is obtained for retrospective chart reviews and testing of the tool. • Performance is evaluated across demographic groups to identify potential disparities. • Clinicians review outputs to ensure recommendations are clinically reasonable and explainable. • Failure modes and edge cases are tested to assess potential patient harm. • Formal sign-off is required before deployment.	• Institutional approvals are obtained to implement the system in clinical workflows. • The AI system is integrated into clinical workflows with mandatory human review. • Clinicians receive training on appropriate use and limitations. • Security controls (authentication, encryption) are enforced in production. • Deployment may be phased, starting with low-risk clinical settings. • Accountability is clearly assigned to clinical leadership and system owners.	• Continuous monitoring detects performance degradation or emerging bias. • Models are retrained only after formal validation and approval. • Security patches and compliance updates are applied promptly. • Clinical incidents or near-misses trigger immediate review. • Continued use is reassessed based on safety, effectiveness, and regulatory guidance.