AI-Ready: An Evaluation Guide for Health and Human Services Agencies

Center for Public Sector AI
US Digital Response

Version 1, updated November 5, 2024

Contents

The Evaluation Guide is a practical tool designed to help government agencies make informed, responsible decisions about adopting AI technologies.

It consolidates best practices in technical feasibility, product delivery, and federal guidance into a clear structure for evaluating the risks and potential of early-stage AI proposals.

The sections include:

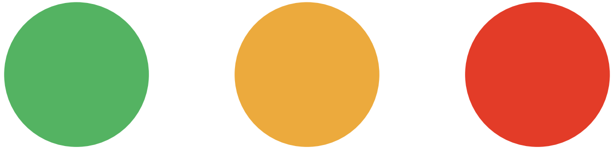
- Value Proposition
- Technical Feasibility
- Mitigating Bias & Ethical Considerations
- Responsible Use & Deployment

The Evaluation Guide is a living resource, meant to evolve and improve over time. While it provides a structured framework, it does not replace the essential need for ongoing monitoring and evaluation of AI technologies themselves.

Developed by the Center for Public Sector AI in collaboration with U.S. Digital Response, the AI Evaluation Guide supported our 2024 open call evaluation process. We thank our partners, including the Aspen Institute Financial Security Program and Humane Intelligence, and invite your feedback. For questions or input, please reach out to info@cpsai.org.

Summary Structure of the Guide

The Evaluation Guide is organized into four pillars, each with key areas of focus. For each focus area, clear guidance is provided on what "good" looks like, using a simple, color-coded traffic light system.



This tool can help you

Suggestions on how to use the tool (based on previous use cases):

- Facilitate discussions with your team to align key priorities for your agency and state, define what success looks like, and set clear expectations for solutions addressing your most urgent challenges.
- Formulate questions for your vendors that directly reflect these priorities and needs.
- Help your team compare options and make informed decisions about the potential risks and opportunities of specific tools and/or use cases.

We have also formatted this guide into an excel document here to facilitate your working practices.

Key to Comparative Slides: Column Structure

Green

The proposal follows all best practices as laid out.

Yellow

The proposal follows best practices as laid out, but significant details are unclear or present risks.

Red

The proposal presents major flaws and omissions to best practice as laid out.

Collaborators

Thank you to all our partners who attended CPSAI's inaugural 2024 Flagship Gathering and beyond, sharing valuable feedback and insights that have enriched this guide. Your engagement and contributions are essential to supporting Health & Human Services.

Center for Public Sector AI

US DIGITAL RESPONSE

humane intelligence

FINANCIAL SECURITY PROGRAM
aspen institute

VALUE PROPOSITION

This section examines the value of the use case and argument for leveraging AI. The focus is on clear outcome metrics and a use case approach that maximizes benefits while minimizing potential harms.

Use Case

A clear problem is addressed that users experience in public sector services. It presents a specific and well-supported problem statement with relevant quantitative or qualitative data on those user needs. It convincingly demonstrates how the proposed solution or approach addresses a critical issue, with strong evidence of user engagement and feedback directly shaping the solution. It includes specific examples of how user input was gathered and utilized.

There is some user research included, but it lacks depth or fails to clearly influence the proposal's direction. There are some questions about its potential impact that need to be addressed before it can be considered further.

Examples of "yellow" use case gaps:

- *There is some user research, but it does not clearly point to the proposal as the best solution.*
- *Metrics for gauging whether a solution is successful are limited or missing.*

The use case is not clearly articulated use case, nor does it demonstrate how the proposed solution or approach addresses a critical issue in public sector services. It lacks specific details and does not provide a compelling rationale for its implementation. No evidence is provided of user interviews, feedback, or other engagement with users, failing to demonstrate that the proposed solution is aligned with actual user needs.

VALUE PROPOSITION

AI Applicability

There is a compelling case for why leveraging AI, whether through machine learning, deep learning, or other techniques, is superior to non-AI approaches. The proposal clearly defines the AI methodologies involved and justifies their use and investment based on the problem's complexity and desired outcomes.

There is a plausible case for why an AI solution could be more effective than a non-AI solution/approach but lacks supporting information or data. The proposal does not sufficiently clarify the specific AI techniques used, which raises questions about their appropriateness and effectiveness.

Examples of “yellow” AI applicability gaps:

- *The performance of simpler non-AI solutions has not been quantified.*
- *Performance improvements from AI have not been credibly estimated.*
- *Unclear if the amount of improvement achievable with AI is worth the additional cost and complexity.*

There is a failure to articulate why leveraging AI for the use case is better than a non-AI solution/approach. It also lacks clarity on the specific AI technologies used, making it difficult to assess the solution's suitability and potential effectiveness.

Risk & Potential Unintended Consequences

The proposed use case is either low risk - i.e., not “rights impacting” or the proposal clearly identifies and articulates the risks and how it would minimize them. A well-developed mitigation strategy is provided. It includes a human fallback in all decision-making and provides transparency to the user regarding how decisions and recommendations were made. All parties understand the ways in which the project could fail or fall short of expectations. Responsibilities for each party have been clearly established, and each has plans to mitigate risks.

For further information on “rights impacting” can be found here from the United Nations.

The proposed use case is high risk - i.e., “rights impacting.” A fully developed risk mitigation strategy is not provided. While some risks are acknowledged, additional detail and planning are needed to ensure all concerns are adequately addressed.

Examples of “yellow” risk identification & mitigation gaps:

- There is a human fallback mechanism for decision-making, but the reasons underlying the decision are not visible to the user.
- The proposal does not sufficiently consider the costs and practical effectiveness of human fallbacks.
- Risks have been identified, but the proposal does not detail effective mechanisms for detecting negative impacts when they occur.

The proposed use case is high risk - i.e., “rights impacting.” It fails to identify or address significant risks and potential unintended consequences associated with the use case. It lacks a risk mitigation strategy, leaving major concerns unaddressed and posing a high likelihood of negative impacts.

VALUE PROPOSITION

Definition of Success & Failure

The proposed use case clearly defines what is considered success and failure, informed by comprehensive user research and partner input. It includes a plan for measuring success and failure, with clear breakpoints where decisions will be made to stop or continue.

What is considered success and failure are clearly defined but have not been developed in partnership with users and clients. There are no clear metrics for measuring success and failure, and there is no governance around stop/go decision-making.

What is considered as success and failure have not been defined or considered. The proposal lacks measures or plans to assess outcomes.

Economic Impacts

There is a clear delineation of the economic impacts of the proposed use case and the economic trade-offs involved. The proposal provides a robust cost-benefit analysis/business value assessment, incorporating future costs, such as those associated with the increasing cost of computing. It may include things like a detailed analysis of job creation or loss, cost savings, and long-term financial viability.

A moderately detailed analysis of economic impacts is provided but lacks an exploration of the trade-offs or a comprehensive cost-benefit analysis/business value assessment. Some economic factors, such as long-term viability or the impact on jobs, are mentioned but not thoroughly explored, leaving gaps in understanding the full economic implications.

There is no delineation of the economic impacts of the proposed use case or any mention of trade-offs. It lacks any job creation or loss analysis, cost savings, or long-term financial viability. It does not provide a cost-benefit analysis/business value assessment, making it difficult to assess the project's economic viability.

VALUE PROPOSITION

Ethical Disclosures

There is a clear delineation of ethical concerns, including conflicts of interest, user trust, and compliance with regulations. The proposal demonstrates a strong commitment to ethical transparency, focused on long-term viability and legal protection and the strategies for mitigating them.

There is an acknowledgment of ethical concerns, but it provides a moderately detailed analysis. Some ethical issues, such as conflicts of interest or technology sourcing, as well as legal protections, are mentioned but not thoroughly explored, leaving gaps in understanding the full ethical implications. The disclosure is present but lacks comprehensiveness.

There is a failure to delineate ethical concerns, providing no clear analysis or disclosure of potential conflicts of interest, user trust, or compliance with regulations. This lack of transparency raises significant concerns about the ethical responsibility of the project.

TECHNICAL FEASIBILITY

This section examines the technical feasibility, ensuring that major design considerations and technical challenges have been identified and convincingly addressed across all critical dimensions including data, system design, algorithms, and security.

Data

Data needed to train and evaluate systems is readily available. Volumes are sufficient, and quality is high (or can be cleaned and filtered to meet needs). Biases in data are well understood and can be mitigated. Processes such as data audits are suggested to ensure there is no private or harmful information in the dataset and that the volume of data meets the minimally viable standards without being excessive.

There is some consideration that the data will need some manipulation/cleaning to be usable by the product/approach. Biases are mentioned, and data coverage & quality appear adequate, but the risk is not fully thought through.

Examples of “yellow” data strategy gaps:

- *The volume of data is not ideal for building the service, but other sources of usable data may exist.*
- *Data needs substantial cleaning to be usable by AI systems.*
- *Not all biases in data are well understood, but overall data coverage and quality seem adequate.*
- *Data has some well-understood biases, but mitigations still need to be investigated.*

Data needed to train and evaluate systems is not readily available due to regulatory restrictions, privacy concerns, or lack of computerization. Volumes may be insufficient, or data quality may not be sufficient to meet release quality needs. The proposal lacks a comprehensive understanding of data biases, and no data audit has been conducted to address potential privacy or harmful information issues, raising significant concerns about the dataset's suitability for the project.

Algorithms

There are technically documented algorithmic techniques to generate the desired behaviors or insights from data. Required performance levels are explicitly understood and technically documented, ensuring that the algorithmic approach is replicable and can consistently achieve the desired outcomes at the necessary standards of effectiveness and reliability.

Algorithmic techniques exist, but performance levels may not meet user requirements. Creative user interface design, improvements to data, or algorithms could help, but success is not guaranteed. The proposal lacks sufficient documentation of the required performance levels and replicability, leading to uncertainty about the algorithm's ability to meet user needs consistently.

Algorithmic techniques to implement the proposed behavior are not mature enough to deliver results at the quality levels needed to be useful. The proposal fails to provide the technical documentation necessary for replicability. It does not clearly define the required performance levels, raising significant concerns about the reliability and effectiveness of the proposed algorithmic approach.

Evaluation

Metrics are defined for evaluating, improving, and monitoring system performance. The metrics are practical to implement and are aligned with overall project success.

The proposal mentions metrics for evaluating and improving system performance but lacks clarity on specific metrics or their implementation, with no link to project success or failure criteria.

Examples of “yellow” evaluation gaps:

- *The evaluation strategy has notable omissions and implementation challenges, though these could potentially be addressed.*
- *Evaluation metrics are defined, but it’s unclear how effectively they capture project success.*
- *Success criteria are outlined, but there’s no clear path to translating them into quantitative metrics for system improvement and monitoring.*

There are significant impediments to the quantitative evaluation of the project, making critical performance monitoring performance and iterative improvement difficult. Measuring what good looks like is not referred to.

System Design

It shows a clear understanding of all interfaces to users, administrators, and other automated systems and addresses these in its design.

Important user interface or system integration challenges have not been addressed. Costs and the practicality of solving them are uncertain.

For example, integrations with legacy systems that are different for each deployment, vastly different IT requirements across agencies, etc., may be theoretically solvable but cost-prohibitive in practice.

Significant barriers prevent the system integrations that are necessary to make the project successful.

Cybersecurity

Limiting access through role-based and least privilege principles, supported by multi-factor authentication and regular audits. Data should be encrypted and classified appropriately, with secure management and disposal practices. Continuous monitoring should be mentioned, along with a strong incident response plan and mention of regular employee training, and adherence to legal regulations.

There is mention of access control management and the importance of cybersecurity. However, there is little depth of approach or support with training or understanding of specific compliance considerations.

There is no mention or capability of access rights & restrictions, neglecting any discussion of audits and not using multi-factor authentication where appropriate. There is no encryption of data nor secure disposal practices, potentially making it vulnerable to breaches. No mention of a response plan nor employee training was included, coupled with a lack of understanding of legal regulations specific to the situation.

User Experience

There is a strong user experience design, is accessible to W3C AA standard or above, and has been usability tested with a diverse group of users, including both public-facing and employee-facing contexts. It has been translated into multiple languages to ensure inclusivity. If applicable and required by the partner, there are ways to customize the look and feel at implementation, providing a consistent user experience across state systems. Accessibility and usability have been prioritized from the start, ensuring the tool is fully functional and intuitive for all user groups.

There are some user experience designs that have been usability tested, but only with a limited group of users. It is accessible to W3C A standard, with plans for further accessibility enhancements. The tool has been translated into one other language. If applicable, there are customization options to account for the client brand. While some accessibility and usability considerations are included, they may not fully meet the diverse needs of all users, including both public-facing and employee-facing groups.

There is a poor user experience design, is hard to use, and does not meet W3C accessibility standards. It has not been translated into any other languages other than its origin, and the interface cannot be customized. The lack of accessibility and usability considerations makes the tool unsuitable for both public-facing and employee-facing environments, potentially limiting its effectiveness and adoption.

Capacity Considerations

Staffing capacities required to successfully implement the project are clearly delineated, with specific attention given to the technical skills needed. The proposal ensures that the team has the necessary expertise and capacity to implement and sustain the project effectively. The delineation of technical skills allows for more effective workforce mapping and ensures that the project is feasible, given the current and projected staff capabilities.

Staffing capacities required to successfully implement the project are moderately delineated. While the proposal identifies some of the technical skills needed, it lacks a comprehensive mapping of the required expertise, which may pose challenges in project implementation. There is a risk that the existing staff may not fully possess the minimal skills needed, potentially impacting the project's feasibility.

Staffing capacities required to successfully implement the project are poorly delineated, with significant gaps in identifying the necessary technical skills. The proposal fails to mention or outline potential gaps in staff capacity or provide any guidance on support, making the project unlikely feasible without significant additional training or hiring.

Legal Audit

There appears to be full alignment with legal constraints, having undergone a comprehensive legal audit. This audit ensures compliance with all current laws and regulations, including those relevant to data privacy, AI ethics, and other applicable legal frameworks. The project is also prepared to adapt to future legal changes, demonstrating a proactive approach to legal compliance.

🟩

It is potentially aligned with legal constraints, but the legal audit is incomplete or lacks depth. While the project meets most current legal requirements, there are potential risks due to gaps in the audit or an incomplete understanding of rapidly changing legal contexts. Further legal review is needed to ensure full compliance and readiness for future legal developments.

🟨

It fails to meet legal requirements and has not undergone a sufficient legal audit. Significant legal risks exist due to non-compliance with current laws and regulations, and the project is unprepared for changes in the legal landscape. Without addressing these issues, the project cannot proceed.

🟥

Data

It clearly outlines a robust strategy for mitigating bias in data sourcing and handling. It demonstrates transparency about the data sets and models used, including data and models from 3rd parties. Specific measures to ensure robust representative data, data diversity, and fairness are documented. Overall, the proposal clearly articulates a proactive approach to managing systemic, statistical, computational, and human bias. Potential risks are clearly articulated and have been thoroughly considered in consultation with potentially affected users.

It acknowledges the importance of mitigating bias in data but provides limited detail on how it is achieved. Some steps are taken to ensure data diversity and fairness, but the approach is not comprehensive, leaving potential gaps in addressing bias.

Examples of "yellow" data gaps:

- *It uses 3rd party models or data (e.g., private sector large language models) without a clear understanding of what data has been used in their creation, how their performance has been evaluated, and the potential issues this may cause.*
- *It acknowledges potential data risks but does not have comprehensive plans for evaluating and mitigating bias.*

It does not address how data is sourced or handled to mitigate bias. It lacks transparency about the data sets used and fails to demonstrate any measures taken to ensure data diversity and fairness.

ETHICS AND SAFETY

This section examines the ethical implications and safety measures, ensuring that potential risks, biases, and societal impacts have been thoroughly identified and addressed. This includes ensuring that best practices safeguards are in place to promote fairness, accountability, and the well-being of affected individuals and communities.

Development & Deployment

AI vulnerabilities are thoroughly identified, reported, and incorporated into a detailed pre-deployment roadmap. The project includes a comprehensive assessment of potential risks during the development phase, ensuring all vulnerabilities are addressed before deployment. This proactive approach minimizes risks and enhances the reliability and security of the AI system.

AI vulnerabilities are identified and reported but have not yet been fully incorporated into the pre-deployment roadmap. While the project acknowledges potential risks, the assessment is incomplete, leaving some vulnerabilities unaddressed before deployment. To ensure the system's security and effectiveness, further work is needed to integrate these considerations into the development process.

No assessment of AI vulnerabilities leaving significant risks unaddressed. The lack of a vulnerability assessment during the development phase poses a high risk to the system's security and reliability, making the project unfit for deployment without substantial revisions.

Ongoing Monitoring & Maintenance

There is a well-developed and detailed plan for ongoing monitoring to detect and mitigate bias. It includes specific mechanisms and processes for regular reviews and product updates, demonstrating a proactive and thorough approach to maintaining ethical standards and addressing bias throughout the solution's lifecycle.

There is a plan for ongoing monitoring, but lacks detailed processes and comprehensive strategies. It recognizes the need for regular reviews but lacks clear methods for detecting and mitigating bias over time.

Examples of "yellow" ongoing monitoring gaps:

- *The proposal does not clearly assign responsibilities for ongoing monitoring and mitigation of biases.*
- *The proposal does not show a clear understanding of how data quality and distributions could evolve over time.*
- *Bias mitigation strategies used at launch may not be practical for regular data updates, such as relying on manual testing methods.*

There is no plan for ongoing monitoring to detect and mitigate bias over time. It does not provide any mechanisms or processes for regularly reviewing and addressing ethical considerations, leaving potential issues unmonitored.

Privacy & Security

A comprehensive and well-detailed approach to privacy and security is presented. It includes specific, robust measures and policies to protect sensitive data, demonstrating a strong commitment to maintaining user privacy and data security throughout the solution's lifecycle.

Privacy and security concerns are noted, but details on protective measures are limited. While some steps are taken to protect data, the approach lacks comprehensiveness and has strategic gaps.

Examples of "yellow" data protection gaps:

- *The proposal lacks a full understanding of systems handling sensitive data, its storage, and necessary privacy and security measures.*
- *The proposal overlooks the unique access needs of different user classes and lacks details on access control, granting, and logging.*
- *The proposal lacks strategies for providing vendors with realistic data during development, such as access to real, anonymized, or synthetic data.*

It fails to address privacy and security concerns adequately. It lacks clear policies or measures to protect sensitive data, posing significant risks to user privacy and data security.

Transparency, Third Party Evaluation, & Public Reporting

The proposed solution or approach allows for third-party evaluation and public reporting of results. It demonstrates a commitment to quickly addressing issues identified through these mechanisms. It

includes specific measures for independent assessments, regular reporting to the public, and clear mechanisms to ensure accountability and openness throughout the solution's lifecycle.

The proposal acknowledges the importance of third-party evaluation, transparency, and public reporting but provides limited or incomplete details on how these could be implemented. While there are some plans for external oversight and reporting, the approach is not fully developed or detailed.

Example of "yellow" evaluation & transparency gaps:

- *Proposal does not articulate a logging strategy that details what historical data is captured for reporting purposes, how long it's kept, and the kinds of reporting that it needs to support.*

The proposal does not allow for third-party evaluation, lacks transparency, and has no provisions for public reporting. There is no external oversight or mechanisms to ensure accountability, leaving the process opaque and unverified.

RESPONSIBLE USE & DEPLOYMENT

This section examines whether the project leverages best practice in its deployment strategy, with a focus on engaging users, commitment to incremental deployment to reduce risk, and a documented approach to data ownership and security infrastructure.

Testing

There is a well-developed and detailed testing plan, including specific methodologies and processes to thoroughly evaluate the solution before deployment. It demonstrates a strong commitment to ensuring the reliability, safety, and effectiveness of the AI system through comprehensive and rigorous testing procedures.

A basic testing plan is included but lacks detailed processes or comprehensive strategies. While it acknowledges the importance of testing, the approach is not fully developed, and there are gaps in ensuring a thorough evaluation of the solution's reliability, safety, and effectiveness.

Examples of "yellow" testing gaps:

- *Testing plans rely primarily on expensive manual effort, which could limit the scale, frequency, and consistency of testing.*
- *Broad testing goals are outlined, but the proposal lacks key details like the kinds of metrics or scenarios that will be used to test.*

It lacks a comprehensive testing plan, providing little to no detail on how the solution will be tested before deployment. It fails to demonstrate that adequate steps are taken to ensure the reliability, safety, and effectiveness of the AI system.

Deployment Approach

A comprehensive and well-structured iterative deployment approach is outlined, including detailed plans for sandboxes, small pilots, and other risk-reducing methods. It demonstrates a strong commitment to minimizing risks through gradual and controlled deployment phases, ensuring the solution is thoroughly tested and refined before full-scale implementation.

All parties understand and agree to ongoing maintenance requirements and responsibilities.

The need for an iterative deployment approach is acknowledged but provides limited details on how it will be implemented. While it includes some risk-reducing methods, such as small pilots or sandboxes, the plan is not fully developed or lacks comprehensiveness in its approach.

Examples of “yellow” deployment strategy gaps:

- *Deployment stages have been identified, but the goals and acceptance criteria for each stage (e.g., what risks to mitigate or what evidence is needed to proceed to the next stage) have not been established.*
- *The agency has not identified an owner or lacks in-house expertise to meet post-deployment maintenance needs.*

An iterative deployment approach is lacking and does not include plans for sandboxes, small pilots, or other risk-reducing methods. It proposes a full-scale deployment without adequate testing phases, increasing the risk of significant issues or failures.

User Engagement

A comprehensive and well-detailed plan for user engagement and research is presented. It includes specific methodologies for involving end-users throughout the development and deployment process, ensuring continuous feedback and insights to enhance the solution's relevance, effectiveness, and user satisfaction.

User engagement and research are acknowledged as important, but specific implementation details are limited. Plans for end-user involvement are underdeveloped, with gaps in ensuring continuous feedback and involvement from users.

Examples of “yellow” user engagement gaps:

- *Users are consulted on some stages of development (e.g., developing requirements) but are not engaged in others (e.g., evaluating prototypes).*
- *User feedback has representation gaps (e.g., language, experience level)*
- *User training needs and transition support have not been addressed.*

There are no plans for user engagement or research. It fails to involve end-users in the development and deployment process, leading to a lack of feedback and insights that could improve the solution's relevance and effectiveness.

Data Ownership & Use

A comprehensive data ownership and control approach is outlined to ensure efficient data use, with policies granting users clear rights and control over their data. The plan underscores a commitment to protecting user autonomy, data rights, transparency, and accountability in data management. Third-party data use is fully documented for alignment with these principles (e.g., through libraries, cloud services, hosted ML models, etc.), and users can opt out of data sharing or request human review of automated decisions impacting their data.

Data ownership and control are recognized as important, but details on efficient management are limited. While it includes some policies that address data rights and user control, the approach lacks development, potentially impacting user autonomy and data protection. For example, the proposal offers users control but lacks specifics on implementation, particularly when data spans multiple systems, complicating rights like data deletion and export on legacy systems.

Examples of “yellow” data ownership & use gaps:

- *The proposal promises user data control but lacks specific implementation details.*
- *Proposed user data rights overlook challenges when data is spread across multiple systems, making deletion or export difficult on legacy platforms.*

Data ownership and control is not addressed, providing no clear policies on who owns the data or how it can be efficiently used. It raises significant concerns about data rights and user autonomy, potentially leading to misuse or unauthorized access. The lack of provisions for user opt-out or human review further exacerbates the risks to user data.

System & Infrastructure Security

A clear and detailed roadmap and system and infrastructure security plan is available. This plan comprehensively addresses all security aspects, including physical, network, and application-level protections. The roadmap ensures that security measures are proactive, regularly updated, and aligned with best practices to protect against emerging threats. The plan also includes specific protocols for responding to security incidents and maintaining the integrity of both data and infrastructure.

The roadmap and plan for system and infrastructure security is ambiguous. While there is some recognition of the need for security measures, the plan lacks clarity and detail, potentially leaving certain areas vulnerable. The approach may include basic protections, but it does not fully address all layers of security or provide a comprehensive strategy for maintaining and updating security measures over time.

There is no roadmap and plan for system and infrastructure security, leaving significant gaps in protection against potential threats. The absence of a structured security strategy poses high risks to the system's integrity and the safety of the data it handles. Without a clear plan, the project is vulnerable to breaches and other security failures, making it unsuitable for deployment.

CPSAI is a nonpartisan, non-profit organization that aligns government leaders and partners to minimize the potential risks associated with AI while maximizing the transformative opportunities it presents. Our mission is to equip government leaders with the tools, knowledge, and expertise they need to make thoughtful decisions about where and how to deploy AI and other emerging technologies to address their most pressing service delivery challenges.

Center for Public Sector AI

cpsai.org