

OPTIONS AND OPPORTUNITIES TO ADDRESS AND MITIGATE THE EXISTING AND POTENTIAL RISKS, AS WELL AS PROMOTE BENEFITS, ASSOCIATED WITH AI AND OTHER ADVANCED ANALYTIC METHODS

OPRE Report #2022-253

September 2022

OPTIONS AND OPPORTUNITIES TO ADDRESS AND MITIGATE THE EXISTING AND POTENTIAL RISKS, AS WELL AS PROMOTE BENEFITS, ASSOCIATED WITH AI AND OTHER ADVANCED ANALYTIC METHODS

Final Report
OPRE Report #2022-253
September 2022

Brian L. Zuckerman, James M. Karabin, Rachel A. Parker, William E. J. Doane, and Sharon R. Williams, IDA Science and Technology Policy Institute

Submitted to:
Joshua Williams (OPRE-COR), Project Officer
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Contract Number: NSFOIA-0408601/75ACF120F80001

Project Director: Brian L. Zuckerman
IDA Science and Technology Policy Institute
1701 Pennsylvania Ave., NW, Suite 500
Washington, DC 20006

This report is in the public domain. Permission to reproduce is not necessary. Suggested citation: Zuckerman, Brian L., James M. Karabin, Rachel A. Parker, William E. J. Doane, and Sharon R. Williams (2022). *Options and Opportunities to Address and Mitigate the Existing and Potential Risks, as well as Promote Benefits, Associated with AI and Other Advanced Analytic Methods*, OPRE Report #2022-253, Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Disclaimer

The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research, and Evaluation are available at www.acf.hhs.gov/opre.

Connect with OPRE



Executive Summary

Statement of Task and Analyses Conducted

In September 2020, the U.S. Department of Health and Human Services (HHS) Office of the Assistant Secretary for Planning and Evaluation (ASPE) and the Administration for Children and Families (ACF), Office of Planning, Research, and Evaluation (OPRE), contracted with the IDA Science and Technology Policy Institute (STPI) to conduct a study focused on emerging issues and needs associated with artificial intelligence (AI) in the health and human services sectors. The purpose of the study was to help inform HHS about the rapidly emerging standards and policies associated with AI research and development. The study has three primary objectives that served as the basis for STPI's analyses:

- Objective 1: Understand the AI guideline landscape across the Federal Government and how HHS could leverage and build on activity occurring in other departments
- Objective 2: Understand the existing and potential barriers, facilitators, risks, and benefits when using AI and other advanced analytic methods in health and human services, including for developing evidence in support of informed decision making
- Objective 3: Identify options and opportunities to address and mitigate the existing and potential risks, as well as promote benefits, associated with AI and other advanced analytic methods

This document is intended to fulfill Objective 3 by synthesizing results from internal, pre-decisional analyses by STPI for HHS under Objective 1 and Objective 2 of the task to identify potential options and opportunities to promote the benefits of the responsible development of AI and to mitigate existing and potential risks. It builds upon three analyses conducted by STPI: (1) a landscape analysis of 10 AI, machine learning (ML), and advanced analytics ethics documents published by Federal departments and agencies and released between 2018 and 2020; (2) a landscape analysis of AI activities across the Federal Government; and (3) a cross-case analysis of nine case studies with an analytical focus on understanding the existing and potential barriers, facilitators, risks, and benefits when using AI in health and human services.

This document begins by summarizing the benefits and challenges of AI identified through the cross-case analysis. It then considers a set of gaps that STPI identified in its landscape analysis of Federal guidance documents and the extent to which some of the considerations identified are already found in some HHS AI guidance documents (e.g., the Office of the Chief AI Officer (OCAIO)'s Trustworthy AI Playbook (*TAI Playbook*)). The document concludes with both broad and targeted considerations for HHS.

Benefits and Challenges of AI Identified

Benefits of AI identified through the health and human services-related case studies and literature reviews conducted as part of case study development fit into three categories:

1. ***Improving mission-oriented processes and services.*** AI offers a range of potential incremental improvements to service delivery such as increased utility (e.g., predictive capability, accuracy, sensitivity, specificity, consistency), task automation, cost savings, and more efficient use of resources (including scalability).
2. ***Enabling larger scale analyses.*** Using AI may lead to benefits related to analysis of “big data,” especially from diverse data sources. These analyses could lead to novel findings in healthcare, human services, and health sciences with the potential to advance HHS’s mission.
3. ***The potential for increased personalization and targeting.*** In health-related cases, personalization of care was identified as a potential benefit of AI. In human services domains, personalization of benefits could improve placement recommendations or improve matching users to services.

Seven categories of challenges were identified from the case studies and literature reviewed:

1. ***User confidence and trust.*** Trust in the reliability of AI models is an overarching challenge, as nearly every other challenge identified relates to some aspect of trust. Specific challenges primarily focused on the ability for decision makers to adequately interpret, evaluate, and act on information provided by a model (“justified trust”) and trust by citizens, including those who may not realize they are being affected by decisions based on a model or a model’s outputs (“public trust”).
2. ***Model performance.*** Challenges were identified related to the fit-to-use, accuracy, or robustness of predictive AI models related to putting a model into practice. These challenges relate to whether the model accomplishes what it was designed to do and issues related to addressing changes in the model or in the deployment setting over time.
3. ***Maintaining privacy.*** Weighing privacy risks against potential benefits can be a key challenge in deciding whether to pursue the development or use of AI. Maintaining privacy is seen as essential for people to maintain some degree of autonomy. Promoting and using strong privacy practices can help to build trust in AI. At the same time, strong privacy protections can have tradeoffs. For instance, prioritizing privacy may contribute to less transparency in the models and their underlying data, or can introduce limitations into models’ performance.

Challenges associated with privacy concerns were identified at all stages of the AI lifecycle and were related to data gathering practices, data protections and security (particularly for sensitive data), and model use.

4. ***Bias.*** The National Institute of Standards and Technology has identified three categories of bias in how AI is designed, developed, and used: 1) statistical and computational biases that occur when a model’s underlying data is not representative of the population the model is addressing; 2) systemic biases related to how data can capture or reflect historical and ongoing inequities; and 3) human bias in model development and results interpretation. Explicit examples of statistical and computational biases and systemic biases were identified in the case studies. Human bias was not explicitly identified in the case studies, but was implicit in some of them.
5. ***Data and dataset quality.*** Challenges related to data and dataset quality occur when data and datasets are incomplete, incorrect, nonrepresentative, or outdated. Poor data quality can contribute to insufficient model performance and bias.
6. ***Transparency and explainability.*** A lack of transparency and explainability can contribute to decreased trust, decreased capability in determining how a model can be used, or decreased capability for testing and evaluating models and identifying limitations. In the context of our case studies, “transparency” refers to the actions and operations surrounding a model and its outputs being visible to and understandable by desired parties. “Explainability” is sometimes used interchangeably, but is treated as a distinct concept that refers specifically to the ability to understand how a model arrives at a particular outcome given a certain input or set of inputs.
7. ***Capacity.*** Challenges related to capacity included the usability of systems, limitations on computational resources and other computing infrastructure necessary to implement AI, and limited expertise and workforce available to develop, use, and govern AI.

Gaps STPI Identified in Previous Federal AI Guidelines and Practices Relevant for HHS

STPI’s Objective 1 analysis of Federal AI guidelines identified a set of gaps in the 10 documents reviewed and of practices identified in individual documents that may be relevant for HHS, given its mission to promote health and human services. Considerations identified from that analysis have already been incorporated into some HHS AI guidance, such as the *TAI Playbook*:

- The importance of trust and trustworthiness of AI solutions.
- The need to address concerns regarding bias.
- The importance of involving stakeholders across the initiation, design, development, and operational stages.
- A need for a focus on accuracy and reliability, understandability, and transparency.
- The importance for guidelines to consider trade-offs between principles, especially with respect to increasing transparency while preserving privacy, increasing accuracy while preserving fairness, and increasing accuracy while preserving understandability.
- The need to ground guidance in relevant legal and regulatory authorities.
- The importance of providing specific implementation guidance on topics such as avoiding reproducing or enhancing societal inequities, identifying responsible decision makers, developing formal mechanisms for stakeholder communication, rights of redress, and model drift.
- The need to create AI governance structures that can guide Department-level policy and address gaps between guidance and application.

Options and Opportunities

HHS has made considerable strides in fostering the responsible use of AI as part of recent efforts. The use of AI in health and human services is evolving rapidly, as are HHS's policies and structures intended to foster that evolution. Our analysis suggests both a set of general considerations for HHS and a set of more targeted considerations that could be incorporated into future HHS guidance regarding responsible AI.

General Considerations

Based on the result of our analyses, STPI identified three general sets of options and opportunities for HHS to consider:

Adapt or adopt successful practices within HHS and across other Federal agencies.

HHS governance bodies with a stake in AI should identify existing best practices within HHS and across the Federal Government to determine potential successful practices that could be adapted or adopted more broadly across HHS. Examples such as an HHS Office of the Inspector General effort to create technological infrastructure, the Food and Drug Administration's work on regulation of AI embedded in medical devices, and the Department of Defense's attempts to create centralized structures for leveraging the appropriate use of AI could be considered.

Focus future improvements to existing guidance on the interactions between AI tools and the people who will use and benefit from them. Many of the opportunities for improvement identified involve interactions between stakeholders and AI tools rather than being technological in nature. Issues of trust, diversity, equity, and training of front-line staff who will be the primary users of AI models all require design, development, and sustainment processes that center the needs and values of users to complement technical precursors such as high-quality input data, technically competent development teams, and robust algorithms.

Strengthen centralized efforts. While HHS has made considerable strides in promoting responsible AI (e.g., through the activities of the OCAIO), our findings suggest that additional centralized efforts to foster AI capacity at HHS could be beneficial. Promoting organizational readiness, adapting best practices, and fostering an HHS-wide technical infrastructure and AI procurement approach would benefit from an organizational reach beyond the limited scope of organizations such as the OCAIO.

Targeted Considerations

Our analyses suggest that incremental improvements to existing practices have the potential to enhance the use of responsible AI across HHS and by their stakeholder communities. Examples of incremental options and opportunities that STPI's analyses identified include:

- Examine whether AI is an appropriate solution as a first step in program initiation checklists
- Explicitly describe the roles of social scientists, ethicists, and community outreach specialists in discussions of “diverse perspectives”
- Identify and mitigate barriers to trust-building as part of development processes
- Assess disparities in access to the internet and digital technologies explicitly during design decisions
- Provide more explicit guidance on whether to build AI solutions in-house or whether to contract for solutions, which may be proprietary technologies
- Incorporate the human dimension of safety and security into the considerations of AI system security alongside technical considerations
- Focus on training the staff users of AI models in addition to training the models themselves
- Explore new tools for preserving privacy
- HHS and tool developers should consider the potential willful misuse of AI as part of system design and development processes

- Operationalize existing guidance such as the *TAI Playbook* through documenting the creation, composition, collection process, preprocessing, cleaning, labeling, uses, distribution, and maintenance of datasets (“data sheets”) and documenting information about the model, intended uses, technical attributes, metrics on model performance, evaluation data, and training data (“model cards”)

Considerations for Catalyzing AI Capacity in HHS

Finally, STPI’s analyses also identified five broader considerations for increasing HHS’s capacity to catalyze responsible AI use:

1. Promoting organizational readiness;
2. Fostering HHS-wide AI infrastructure;
3. Promoting best practices and good governance approaches across HHS and across the Federal Government;
4. Building contracting platforms specific to responsible AI; and
5. Supporting social science research regarding responsible AI use in health and human services contexts.

Limitations

There are two noteworthy limitations of this document. First, AI is a rapidly-evolving field and governance of AI activities by Federal agencies is changing quickly as well. Our analyses are based primarily on documents published in 2020 and 2021. New approaches and considerations may have already emerged that could not be captured in our work. Second, our analyses of barriers, facilitators, risks, and benefits build upon a specific set of nine case studies related to particular health and services domains. There may be important challenges and opportunities associated with AI potentially relevant to health and human services domains that might have been identified were we to have used different starting points for our analyses, incorporated a broader set of domains, or conducted more case studies.

Contents

1.	Introduction	1
A.	Task Introduction	2
B.	Task Objectives and STPI’s Analyses in Support of Those Objectives.....	2
1.	Objective 1: Landscape Analysis	3
2.	Objective 2: Facilitators, Risks, and Benefits of AI.....	3
3.	Objective 3: Synthesis and Policy Implications	4
2.	Benefits and Challenges of AI Identified	5
3.	AI and AI Governance Relevant to HHS	9
A.	Comparison of the Maturity of AI-related Activities at HHS with Those of other Federal Agencies.....	9
B.	Evolution of AI Governance at HHS	10
C.	Gaps STPI Identified in Previous Federal AI Guidelines and Practices Relevant for HHS	11
4.	Options and Opportunities for HHS	13
A.	General Considerations	13
B.	Targeted Considerations.....	14
1.	Examine Whether AI Is an Appropriate Solution as a First Step.....	14
2.	Explicitly Describe the Roles of Social Scientists, Ethicists, and Community Outreach Specialists in Discussions of “Diverse Perspectives”	14
3.	Identify and Mitigate Barriers to Trust-building as Part of Development Processes.....	15
4.	Assess “Digital Divide” Concerns Explicitly during Design Decisions	15
5.	Provide More Explicit Guidance on “Buy versus Build” Decisions and the Role of Proprietary Technologies.....	16
6.	Incorporate the Human Dimension of Safety and Security into the Safe/Secure Principle	16
7.	Focus on Training the Staff Users in Addition to Training the Model	17
8.	Explore New Tools for Preserving Privacy.....	17
9.	Incorporate the Possibility of Willful Misuse into Design Considerations	18
10.	Operationalizing the <i>TAI Playbook</i> and Similar Strategies through Documentation Practices	18
C.	Considerations for Catalyzing AI Capacity in HHS	19
1.	Promote Organizational Readiness	20
2.	Foster an HHS-wide AI Infrastructure	20

3. Promote and Coordinate Best Practices and Good Governance Approaches across OPDIVs and across the Federal Government.....	21
4. Build Contracting Platforms Specific to Responsible AI.....	22
5. Support Social Science Research regarding AI Use in Health and Human Services Contexts	22
Appendix A. HHS Governance of AI to Date in Light of STPI’s Objective 1 Analyses	A-1
References.....	B-1
Abbreviations.....	C-1

1. Introduction

In 2019, two physicians at the National Cancer Institute imagined cancer clinical trials and oncology treatment 20 years in the future. In their article (Mittra and Moscow 2019), they assumed a world where a patient’s diagnostic information and health status would be analyzed using artificial intelligence (AI)-based algorithms to assign the patient optimally to clinical trials of new therapeutics or to personalized treatment regimens. Electronic medical records, test results, and information from wearable devices would all be seamlessly combined using the power of computational approaches to optimize treatments and maximize the likelihood of recovery.

In focusing on the future of cancer treatment and clinical trials, the authors did not mention the practical challenges of implementing these new technologies—difficulties in linking records effectively and meaningfully or in designing algorithms that are able to predict accurately how individual patients will respond to a combination of therapeutics. Nor did they consider whether such approaches might leave behind patient groups or communities—for example, those whose medical records were less likely to be digitized. The authors also did not consider the potential for ethical challenges, such as those associated with linking and reusing data from millions of people whose level of consent and understanding of the uses to which data might be put vary or the potential for malign actors to compromise personal health records underlying the AI models.

As futures such as these advance toward fruition, government agencies, private industry, nonprofits, and individual citizens will need to weigh the opportunities and trade-offs of AI while mitigating potential risks or harms to individuals and communities as such tools are increasingly deployed across health and human services domains.

A. Task Introduction

In September 2020, the U.S. Department of Health and Human Services (HHS) Office of the Assistant Secretary for Planning and Evaluation (ASPE) and the Administration for Children and Families (ACF), Office of Planning, Research, and Evaluation (OPRE), contracted with the IDA Science and Technology Policy Institute (STPI) to conduct a study focused on emerging issues and needs associated with AI in the health and human services sectors. The purpose of the study was to help inform HHS about the rapidly emerging standards and policies associated with AI research and development.

B. Task Objectives and STPI's Analyses in Support of Those Objectives

The study has three primary objectives that served as the basis for STPI's analyses:

- Objective 1: Understand the AI guideline landscape across the Federal Government and how HHS could leverage and build on activity occurring in other departments
- Objective 2: Understand the existing and potential barriers, facilitators, risks, and benefits when using AI and other advanced analytic methods in health and human services, including for developing evidence in support of informed decision making
- Objective 3: Identify options and opportunities to address and mitigate the existing and potential risks, as well as promote benefits, associated with AI and other advanced analytic methods

A DEFINITION OF "ARTIFICIAL INTELLIGENCE" USED BY SOME FEDERAL AGENCIES

- (1) 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 datasets.
- (2) 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.
- (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
- (4) A set of techniques, including machine learning, that is designed to approximate a cognitive task.
- (5) 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.

Source: John S. McCain National Defense Authorization Act for Fiscal Year 2019 (Public Law 115-232, Section 238(g))

Work under the first two objectives resulted in internal, pre-decisional analyses for HHS and the findings are summarized here.

1. Objective 1: Landscape Analysis

STPI undertook two analyses addressing Objective 1. In the first analysis, completed in June 2021, STPI conducted a landscape analysis of 10 AI ethics documents published by Federal departments and agencies and released between 2018 and 2020 (Pratico et al. 2021).

The analysis compares the contents of the documents to nine “Principles for AI Use in Government” that are outlined in Executive Order (EO) 13960, Section 3, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* (Executive Office of the President 2020). The analysis indicated whether and how the nine principles in the EO are discussed in the 10 agency AI guidelines and identifies additional concepts and topics in the guidelines outside of those included in the EO. STPI (1) assessed whether principles in EO 13960 are addressed in existing agency AI guidelines; (2) evaluated how the principles and related concepts in EO 13960 are discussed in existing AI guidelines; and (3) identified additional concepts or topics discussed in existing AI guidelines that are not included in EO 13960, but that would be useful to HHS in establishing AI ethics guidelines.

In a second analysis, completed in November 2021, STPI conducted a landscape analysis of AI activities across the Federal Government (Pratico et al., 2022). As part of this effort, STPI used data collected by the Administrative Conference of the United States (ACUS) on AI use cases across the Federal Government. HHS sponsors provided a separate catalog of AI use cases at the Department in March 2021. The two catalogs were analyzed to compare activities occurring across the Federal Government with HHS-specific use cases.

2. Objective 2: Facilitators, Risks, and Benefits of AI

In support of Objective 2, STPI completed a series of nine case studies with an analytical focus on understanding the existing and potential barriers, facilitators, risks, and benefits when using AI in health and human services (Karabin et al., 2022). Case studies were based either on interviews with HHS subject matter experts or literature reviews. They were chosen to capture a diversity of types and uses of AI at varying degrees of maturity. Case studies also reflected a diversity of HHS domains. After completing a cross-case analysis, the document further presented its findings in the context of HHS’s “Trustworthy AI Playbook” (*TAI Playbook*), released in September 2021 (HHS 2021b), by assessing the extent to which the implementation steps and checklists described in the *TAI Playbook* aligned with findings from the cross-case analysis.

STPI CASE STUDIES

- (1) Centers for Disease Control and Prevention and Georgia Tech AI Bias
- (2) Office of the Inspector General (OIG) AI Infrastructure
- (3) Centers for Medicare and Medicaid Services Fraud Detection
- (4) Agency for Healthcare Research and Quality Patient Selection Prediction
- (5) Cardiovascular health clinical decision support systems
- (6) Case Management
- (7) Child Welfare Services
- (8) COVID-19 Surveillance
- (9) Anti-Human Trafficking

Source: Karabin et al., 2022

3. Objective 3: Synthesis and Policy Implications

The intent of Objective 3 is to create materials that can inform decisions made by HHS regarding how to pursue use of AI at HHS, including the potential use of any standards related to AI or other technologies that may be pertinent to reducing bias. This document is intended to fulfill Objective 3 by providing a short synthesis of our previous work intended to provide considerations for future HHS policy-making efforts.

This document begins by summarizing the benefits and challenges of AI identified through the cross-case analysis. It then considers a set of gaps that STPI identified in its landscape analysis of Federal guidance documents and the extent to which some of the considerations identified are already found in some HHS AI guidance documents (e.g., the *TAI Playbook*). The document concludes with both broad and targeted considerations for HHS.

2. Benefits and Challenges of AI Identified

The results in this chapter are drawn from STPI's Objective 2 analyses. Benefits of AI identified through the case studies (Karabin et al., 2022) fit into three categories:

1. ***Improving mission-oriented processes and services.*** AI offers a range of potential incremental improvements to service delivery such as increased utility (e.g., predictive capability, accuracy, sensitivity, specificity, consistency), task automation, and cost savings and efficiency of resources (including scalability). Some examples drawn from the case studies are: (1) The cardiovascular decision support case suggests that AI has the potential to improve early diagnoses by providing greater accuracy (i.e., fewer errors) or identify patterns with diagnostic significance that go undetected by humans; (2) The human trafficking case is one where advanced analytics have been used to increase the efficiency of outreach by orders of magnitude relative to relying on personnel; and (3) In health care domains, there may be cost savings incurred by reducing incorrect referrals and automating record keeping through the use of advanced analytics.
2. ***Enabling new or larger analyses.*** AI may lead to benefits related to analysis of "big data," especially from heterogeneous data sources. The COVID-19 surveillance case, for example, identifies as a benefit of advanced analytics the ability to integrate data coming from different sources, including health data from local public health authorities, health data from academic and private sources, mobility data, and social media data.
3. ***The potential for increased personalization and targeting.*** In health-related cases such as the COVID-19 case, personalization of care was identified as a potential benefit of AI. In human services domains such as the child welfare case, advanced analytics offer the potential to improve placement recommendations and child-foster parent matching by more precisely targeting the specific needs of the child and the preferences of the caregivers.

Seven categories of challenges were identified:

1. ***User confidence and trust.*** Trust in the reliability of AI models is an overarching challenge, as nearly every other challenge identified relates to some aspect of trust. Specific challenges primarily focused on the ability for decision makers to adequately interpret, evaluate, and act on information provided by a model ("justified trust") and trust by citizens, including those who may not realize they are being affected by decisions based on a model or a model's outputs ("public trust").

2. **Model performance.** Challenges related to the fit-to-use, accuracy, or robustness of models related to putting a model into practice. These challenges relate to whether the model accomplishes what it was designed to do and issues related to addressing changes in the model or in the deployment setting over time.
3. **Maintaining privacy.** Weighing privacy risks against potential benefits can be a key challenge in deciding whether to pursue the development or use of AI. Maintaining privacy is seen as essential for people to maintain some degree of autonomy. Promoting and using strong privacy practices can help to build trust in AI. At the same time, strong privacy protections can have tradeoffs; for instance, prioritizing privacy may contribute to less transparency in the models and their underlying data, or can introduce limitations into models' performance. Challenges associated with privacy concerns were identified at all stages of the AI lifecycle and were related to data gathering practices, data protections and security (particularly for sensitive data), and model use.
4. **Bias.** The National Institute of Standards and Technology has identified three categories of bias in how AI is designed, developed, and used: 1) Statistical and computational biases that occur when a model's underlying data is not representative of the population the model is addressing; 2) systemic biases related to how data can capture or reflect historical and ongoing inequities; and 3) human bias in model development and results interpretation. Explicit examples of statistical and computational biases and systemic biases were identified in the case studies. Human bias was not explicitly identified in the case studies, but was implicit in some of them.
5. **Data and dataset quality.** Challenges related to data and dataset quality occur when data and datasets are incomplete, incorrect, nonrepresentative, or outdated. Poor data quality can contribute to insufficient model performance and bias.
6. **Transparency and explainability.** In the context of our case studies, "transparency" refers to the actions and operations surrounding a model and its outputs being visible to and understandable by desired parties. "Explainability" is sometimes used interchangeably, but is treated as a distinct concept that refers specifically to the ability to understand how a model arrives at a particular outcome given a certain input or set of inputs. A lack of transparency and explainability can contribute to decreased trust, decreased capability in determining how a model can be used, or decreased capability for testing and evaluating models and identifying limitations.

7. **Capacity.** Challenges related to capacity included the usability of systems, limitations on computational resources and other computing infrastructure necessary to implement AI, and limited expertise and workforce available to develop, use, and govern AI. Another capacity-related concern identified is that existing infrastructural inequities, such as disparities in access to the internet and digital technologies (commonly referred to as the “digital divide”), could be exacerbated as advanced analytics become more common.

3. AI and AI Governance Relevant to HHS

A. Comparison of the Maturity of AI-related Activities at HHS with Those of other Federal Agencies

Any consideration of future opportunities must begin with an understanding of the current state of health and human services-related use of AI. As part of our Objective 1 landscape analysis, STPI used data collected by ACUS on AI use cases across the Federal Government (Engstrom et al. 2020). These data were collected from public-facing websites by ACUS from January to August 2019 to understand better how AI models are currently implemented at executive departments and agencies, including on the stage of maturity of those activities. The ACUS-identified use cases suggested that the majority of HHS activities in the ACUS dataset were still in the planning stage at the time of data collection (Table 1). In contrast, the majority of non-HHS activities in the ACUS dataset were either partially or fully deployed. This finding correlated with information received by STPI from speaking with representatives from various HHS Operating Divisions (OPDIVs), who noted that many of the activities occurring at HHS are in early development or planning stages.

Table 1. Number of HHS and Non-HHS Cases in ACUS Dataset across Implementation Stages

Implementation Stage	Number of HHS Activities	Number of Non-HHS Activities
Fully Deployed	1 (5%)	52 (38%)
Piloting or Partially Deployed	8 (42%)	35 (26%)
Planning	10 (53%)	50 (36%)

Source: Pratico et al., 2022

Across the nine Objective 2 case studies, the use of AI in health domains appeared to be more advanced and more likely to be used operationally than in human services domains. An exception was in child welfare services, where AI-enabled applications have been in use for several years. Child welfare services is an area where legislation has incentivized the development of AI-enabled approaches because Congress has directed HHS to focus on incorporating innovative approaches to preventing mistreatment of children. HHS has used this authority to incentivize the development of AI-enabled risk assessments and preventative approaches as examples of new techniques.

B. Evolution of AI Governance at HHS

In January 2021, HHS released its *Artificial Intelligence (AI) Strategy*, an eight-page document that laid out high-level strategic principles for promoting the use of AI in health and human services domains (HHS 2021a). In March 2021, HHS created the Office of the Chief AI Officer (OCAIO), to “facilitate effective collaboration on AI efforts across HHS agencies and offices” (HHS n.d.). One initial effort of OCAIO has been to develop the *TAI Playbook*, released in September 2021 (HHS 2021b). The document was written in support of HHS’s efforts to promote the use of AI. To realize the benefits of AI, “we must maintain public trust by ensuring that our solutions are ethical, effective, and secure” (HHS 2021b, page 4). The 109-page document includes a discussion of six principles intended to promote the trustworthiness of HHS-developed AI solutions, internal considerations for HHS with respect to promoting trustworthiness throughout the AI development lifecycle, and external considerations such as regulatory issues.

The *TAI Playbook* is grounded in the principles laid out in EO 13960 (Executive Office of the President 2020) and Office of Management and Budget Memorandum M-21-06, “Guidance for Regulation of Artificial Intelligence Applications” (Executive Office of the President 2021). It is intended as an initial step toward HHS-wide policies and practices (e.g., acquisition policies) for governing AI. Its intended audience is senior leadership and program managers at HHS. The OCAIO has also created a Community of Practice across HHS AI practitioners and is working as of summer 2022 to launch an AI Council to implement the AI Strategy and to serve as a group of experts who can support HHS AI use (HHS n.d.).

TAI PLAYBOOK PRINCIPLES

- (1) Fair/Impartial: AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants
- (2) Transparent/Explainable: All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection
- (3) Responsible/Accountable: Policies should outline governance and who is held responsible for all aspects of the AI solution (e.g., initiation, development, outputs, decommissioning)
- (4) Safe/Secure: AI systems should be protected from risks (including Cyber) that may directly or indirectly cause physical and/or digital harm to any individual, group, or entity
- (5) Privacy: Individual, group, or entity privacy should be respected, and their data should not be used beyond its intended and stated use; data used has been approved by the data owner or steward
- (6) Robust/Reliable: AI systems should have the ability to learn from humans and other systems and produce accurate and reliable outputs consistent with the original design

Source: HHS 2021b, page 15

C. Gaps STPI Identified in Previous Federal AI Guidelines and Practices Relevant for HHS

STPI's Objective 1 analysis of Federal AI guidelines identified a set of gaps in the 10 documents reviewed and of practices identified in individual documents that may be relevant for HHS, given its mission to promote health and human services. Considerations identified from that analysis (Appendix A) have already been incorporated into some HHS AI guidance, such as the *TAI Playbook*:

- ***The importance of trust and trustworthiness of AI solutions.*** The *TAI Playbook* has “Trustworthy AI” in its title and the introductory message from the HHS Chief AI Officer focuses on the value of maintaining public trust as AI solutions are developed. The balance of the document attempts to further operationalize “trustworthy” AI and how to implement it in health and human services contexts.
- ***The need to address concerns regarding bias.*** The *TAI Playbook* primarily addresses concerns regarding bias as part of the operationalization of the Fair/Impartial principle. It identifies approaches to mitigate bias across all stages of the AI development lifecycle.
- ***The importance of involving stakeholders across the initiation, design, development, and operational stages.*** The *TAI Playbook* makes extensive reference to the importance of involving stakeholders across the initiation, design, development, and operational stages
- ***A need for a focus on accuracy and reliability, understandability, and transparency.*** A finding from the Objective 1 analysis is that many of the Federal guidance documents studied provided less detail related to three of the EO 13960 principles: (1) accuracy and reliability; (2) understandability; and (3) transparency. These three principles, however, are strongly incorporated into the *TAI Playbook*, embodied in the Transparent/Explainable and Robust/Reliable principles.
- ***The importance for guidelines to consider trade-offs between principles, especially with respect to increasing transparency while preserving privacy, increasing accuracy while preserving fairness, and increasing accuracy while preserving understandability.*** The *TAI Playbook* makes explicit reference to trade-offs between accuracy and fairness and between accuracy and understandability.

- ***The need to ground guidance in relevant legal and regulatory authorities.*** The *TAI Playbook* makes explicit reference to grounding AI development in relevant legal and regulatory requirements as part of both the Privacy and the Fair/Impartial principles. In addition, the document includes an appendix that identifies statutory authorities relevant to the implementation of AI solutions.
- ***The importance of providing specific implementation guidance on topics such as avoiding reproducing or enhancing societal inequities, identifying responsible decision makers, developing formal mechanisms for stakeholder communication, rights of redress, and model drift.*** Many of the specific implementation topics that STPI identified in guidance from the U.S. Agency for International Development are addressed in the *TAI Playbook* as well.
- ***The need to create AI governance structures that can guide Department-level policy and address gaps between guidance and application.*** HHS launched the OCAIO in March 2021. That office’s functions as of April 2022 are to: “drive implementation of the HHS AI strategy, stand up the HHS AI governance structure, coordinate the HHS response to AI-related federal mandates, and foster collaboration among HHS agencies and offices” (HHS n.d.). Other agencies’ governance structures for AI, such as those at DoD, have a broader coordination mission. These governance structures, as part of their coordination functions, should be designed to identify any gaps between Department-level guidance and the implementation of that guidance by sub-agencies.

4. Options and Opportunities for HHS

The use of AI in health and human services is evolving rapidly, as are HHS’s policies and structures intended to foster that evolution. HHS and its OCAIO have made considerable strides in fostering the responsible use of AI in their recent efforts. Our analyses suggest that improvements to existing practices—whether through updates to guidance documents such as the *TAI Playbook* or by fostering practices for operationalizing the principles described in them through documentation—have the potential to enhance the use of responsible AI by OPDIVs and their stakeholder communities. This chapter begins with general considerations that HHS could consider, followed by some more targeted considerations. The chapter concludes with considerations for catalyzing AI capacity across HHS.

A. General Considerations

HHS and its OCAIO have made considerable strides in fostering the responsible use of AI in their recent efforts. The specific options and opportunities STPI identified can be grouped into three higher-level categories:

Adapt or adopt successful practices within HHS and across other Federal agencies. One mission of central AI governance organizations within HHS should be to identify existing best practices within HHS OPDIVs and its Staff divisions (StaffDIVs) to determine potential successful practices that could be adapted or adopted Department-wide. Examples such as the OIG effort to create technological infrastructure and work by the Food and Drug Administration (FDA) on regulating AI/ML-enabled medical devices could be assessed with a view to implementing similar practices at the Department level. Opportunities to learn from other Federal agencies—specifically from DoD’s attempts to create centralized structures for promoting the use of AI, including infrastructure development and procurement practices—should also be pursued.

Strengthen centralized efforts. While HHS has made considerable strides in promoting responsible AI through the activities of the OCAIO, more rapid and sustained progress could be facilitated, however, through centralized efforts to foster AI capacity at HHS. Promoting organizational readiness, adapting best practices, and fostering an HHS-wide technical infrastructure and AI procurement approach require an organizational reach beyond the limited scope of organizations such as the OCAIO that are focused on coordination of existing efforts and providing high-level guidance.

Focus future improvements to existing guidance on the interactions between AI tools and the people who will use and benefit from them. Many of the opportunities for improvement of guidance documents such as the *TAI Playbook* involve interactions between stakeholders and AI tools rather than being technological in nature. Issues of trust, diversity, equity, and training of front-line staff who will be the primary users of AI models all require design and development processes that center the needs and values of stakeholders in addition to technical precursors such as high-quality input data, technically competent development teams, and robust algorithms. New research grounded in the social sciences (including interdisciplinary work that incorporates computing and data sciences as appropriate) on how best to responsibly use AI in health and human services contexts would address unresolved questions regarding future AI adoption.

B. Targeted Considerations

Our analyses suggest that incremental improvements to existing practices have the potential to enhance the use of responsible AI across HHS and by their stakeholder communities. Nine examples of potential considerations are presented below. It is noteworthy that the majority of these suggestions relate to the human dimensions of AI rather than updates with respect to these technologies themselves.

1. Examine Whether AI Is an Appropriate Solution as a First Step

The *TAI Playbook* is framed with the implicit assumption that an AI-enabled solution is going to be developed; the Playbook's purpose is to assist in the responsible development of the system. The Playbook contains a checklist to be followed that includes a set of questions intended to assess the need for system development. It includes reference to cost-benefit analyses and accuracy-explainability trade-offs,¹ but does not offer specific guidance as to when machine learning (ML)- or AI-enabled analyses are preferable to extant analytical tools. Providing criteria or at least examples of when AI-enabled analyses are more or less appropriate can assist OPDIVs in making informed choices regarding investing in AI-enabled tools.

2. Explicitly Describe the Roles of Social Scientists, Ethicists, and Community Outreach Specialists in Discussions of “Diverse Perspectives”

The *TAI Playbook* already emphasizes the need for stakeholder involvement and collective input in designing and implementing AI tools and the importance of designing AI responsibility. The document, however, is less specific regarding the expertise that will be required in order to design, develop, acquire, use, and sustain AI responsibly and to

¹ HHS 2021b page 30 (checklist). References to cost-benefit analyses (HHS 2021b, page 80), and accuracy-explainability trade-offs (HHS 2021b, page 86) are hyperlinked from the checklist page.

facilitate the stakeholder involvement required. The Playbook’s checklist for initiating an AI project responsibly asks for consideration of the following question: “Will the team include diverse perspectives and programmatic expertise?” (HHS 2021b, page 30). However, the document does not specify what “diverse” means in the context of project initiation and conduct; the Playbook’s discussion of “AI Project Team Roles” focuses on the programmers and data scientists involved in the technical development of the system (HHS 2021b, page 81) but does not include mention of ethicists, social scientists, or community outreach specialists who might be involved in stakeholder facilitation activities or incorporating responsible AI principles into model development and implementation. Future versions of documents such as the *TAI Playbook* should more explicitly identify these roles as part of responsible AI design and deployment, describe the kinds of non-computing expertise that would be valuable to include in AI-related projects, and specify that these efforts should be woven into the fabric of AI design, development, and deployment processes.

3. Identify and Mitigate Barriers to Trust-building as Part of Development Processes

For people to consider credible the claims about the trustworthiness of AI systems, implementation of the principles in the *TAI Playbook* are necessary but not sufficient. Transparency and trustworthiness are not purely algorithmic. HHS may also need to emphasize non-technological aspects of trust-building, such as:

- Contextual factors of peoples’ relationships with technology and broader socioeconomic structures;
- Fostering the development of AI-enabled tools that are viewed as credible and admissible in legal and regulatory processes;
- The broader ecosystem of trust in AI influencing public perceptions; and
- Institutional mechanisms to communicate the credibility of claims of trustworthiness.

As part of guidance regarding whether to initiate a new AI project (HHS 2021b, page 80), the next iteration of the *TAI Playbook* or other future AI guidance documents could include questions such as whether there may be specific challenges related to stakeholders’ perceptions of a particular implementing agency or the problem domain that might need to be overcome for the system to be accepted and trusted by stakeholders.

4. Assess “Digital Divide” Concerns Explicitly during Design Decisions

The *TAI Playbook*’s emphasis on fairness and bias includes ensuring that data are representative and not subject to undesirable historical or statistical bias.

The literature reviewed (Karabin et al., 2022) notes that a lack of access to tools by marginalized or older populations may also influence the use and acceptance of AI-enabled solutions. In use cases where affected populations may have limited access to or trust in digital technologies, strategies recommended by the *TAI Playbook* may require modification. The *TAI Playbook* recommends stakeholder involvement in developing and vetting models but, should affected stakeholders have varying levels of digital access and expertise, additional effort may be required to involve those with more limited exposure to AI tools. Similarly, users with differing levels of digital access and expertise may respond differently to privacy-transparency or accuracy-reliability trade-offs. Future versions of the *TAI Playbook* or other AI guidance documents would benefit from providing strategies for mitigating challenges that may result from mismatches between the technological sophistication of AI tools and the digital access and expertise of affected stakeholders.

5. Provide More Explicit Guidance on “Buy versus Build” Decisions and the Role of Proprietary Technologies

One concern identified in the case studies and the literature more generally is that making use of off-the-shelf algorithms—especially those that are closely held by their vendors—can contribute to concerns over transparency and explainability. The *TAI Playbook* discusses conditions under which project teams may wish to consider purchasing commercial tools as compared with considerations regarding the decision to build a tool internally (HHS 2021b, page 31). While the *TAI Playbook* identifies some considerations (e.g., “Organizations have less insight into how the vendor trained and tested the algorithm” and “Contracts should include provisions for appropriate access to data, design documentation, and test results”), the document does not define terms such as “appropriate” or specify conditions under which having insight into the training process is more or less appropriate. In future versions of documents such as the *TAI Playbook*, HHS may wish to delve further into the procurement process and to provide guidance to program staff about the conditions under which off-the-shelf procurement may be more (or less) appropriate.

6. Incorporate the Human Dimension of Safety and Security into the Safe/Secure Principle

The *TAI Playbook* incorporates safety and security at the principle level. We note, however, that the description of the “Safe/Secure” principle focuses on technical aspects more than on the human dimension even though safety and security are technically and socially intertwined. Designing technical security solutions requires considering the human element of system safety such as the threat of insider attacks or the role of inattention or human error in compromising security practices. In addition to the technical activities described already in the *TAI Playbook* (e.g., HHS 2021b, page 63), HHS may wish to include in future AI governance documents human-centered security practices (e.g., staff training, insider threat monitoring) in addition to the technical practices already incorporated into the *TAI Playbook*.

7. Focus on Training the Staff Users in Addition to Training the Model

One challenge identified in the case studies is that AI-enabled tools often were not used—or not used as intended—because they did not integrate well into the workflows of the staff who had to implement and make decisions guided by those analytics. STPI’s case study analyses (Karabin et al., 2022) found that often staff users of the tools were not trained in their proper use, leading users to misuse or ignore model outputs. The *TAI Playbook* describes in many locations and in great detail the importance of the proper use of data in training AI models, but devotes relatively little emphasis to ensuring that the users of an AI tool are trained in its proper use and interpretation. Similarly, in the Initiation and Concept phase checklist of the *TAI Playbook* (HHS 2021b, page 30) there is an “Evaluate Stakeholder Needs” section that asks who the users of a tool might be, but the checklist does not explicitly ask potential AI tool designers to assess whether or how a new analytical tool might upend existing workflows such that substantial training of staff might be required or that mitigation measures may need to be undertaken to ensure new analytical tools are properly used. Future HHS AI guidance documents should incorporate a focus on training the staff users.

8. Explore New Tools for Preserving Privacy

The *TAI Playbook* discusses the possibility of adversarial attacks, including attempts to breach the confidentiality of a model and the data underlying it to expose personal and protected information (i.e., inference attacks). The discussion of security risks and mechanisms to ameliorate them (HHS 2021b, page 93) includes Privacy Preserving Machine Learning (PPML) approaches such as Private Aggregation of Teacher Ensembles, which may be required to secure a system and its data. As PPML techniques continue to evolve, HHS may need to investigate how technical methods, such as differential privacy, satisfy compliance requirements and HHS’s needs to preserve privacy. The field of PPML is relatively new and there are still legal and policy ambiguities around questions such as what level of privacy protection is appropriate. Other agencies engaged in privacy protecting uses of AI and ML, such as the U.S. Census Bureau, may serve as models in this area (U.S. Census n.d.).

The *TAI Playbook* also references consent specifically in describing practices for securing data. As identified in the literature, however, up-front consent may not be sufficient to protect personal data or to ensure that participants have full knowledge of all of the potential re-uses of data that may occur. Large cohort studies such as the National Institutes of Health (NIH) *All of Us* study are making use of dynamic and ongoing consent techniques to ensure that study participants remain involved in consenting to the use of their data while allowing them to withdraw or modify their consent mid-study (NIH n.d.). Future iterations of documents such as the *TAI Playbook* may benefit from (1) describing the conditions under which initial informed consent may not be sufficient and (2) presenting potential opportunities for incorporating dynamic or ongoing consent techniques.

9. Incorporate the Possibility of Willful Misuse into Design Considerations

While most health and human services implementations of AI are unlikely to be willfully misused by adversaries (i.e. “dual-use” considerations), some may—as was identified in a 2022 publication where an AI intended for discovering chemotherapeutics also discovered chemical weapons (Urbina et al. 2022). While the *TAI Playbook* considers adversarial attacks and vulnerabilities, it does not explicitly explore dual-use considerations. Future iterations of AI guidance documents such as the *TAI Playbook* may benefit from incorporating questions into the early stages of model design (e.g., into the *TAI Playbook’s* project initiation checklists) such as those included in the European Commission’s guidelines, such as: “Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or [not] deploying the system)?” (European Commission 2019). The *TAI Playbook* already discusses analyzing potential system vulnerabilities to protect against hostile attacks (HHS 2021b, pages 43, 93)—this approach could potentially be expanded in future HHS AI guidance documents to analyze potential dual-use concerns.

10. Operationalizing the TAI Playbook and Similar Strategies through Documentation Practices

Although many organizations have

developed responsible AI principles, there are no universal methods or best practices that have been identified to date for operationalizing them; it is difficult to specify the operational meanings of terms such as “fairness” and the rapid evolution of AI tools raises the concern that it might not be feasible to embed those meanings into software development and management processes even if consensus definitions of these terms were created. The reviewed literature instead suggests embedding a definitions process into the design of AI-enabled systems; as stakeholders deliberate during the planning process for the development of a new tool or system, they can engage in ideation activities around how principles should be operationalized for that specific system. To examine privacy needs in a specific setting, for example, stakeholders might engage in discussion and deliberation on the actors, needed privacy protections harm that might occur, approaches to providing privacy, and the scope of privacy protections required (Wong and Mulligan 2018).

DATA SHEETS FOR DATASETS

Researchers have proposed that datasets such as those incorporated into AI-enabled tools be accompanied by “data sheets” (analogous to material safety data sheets for chemicals or data sheets for electronics components). The data sheets would include answers to common questions such as, “For what purpose was the dataset created?” or “Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?” (Gebru et al. 2021).

While the reviewed literature identifies challenges in operationalizing AI principles, there is more consensus that common processes for documenting AI systems can serve as a mechanism for promoting trustworthy and responsible uses of AI. Data sheets that document dataset characteristics or model cards that describe an algorithm’s technical attributions and training mechanisms can foster understandability and transparency of AI applications. Standardized practices for data sheets and model cards can also enable mechanisms of accountability. Documentation may assist in understanding whether and where, for example, testing for bias was conducted, what that testing consisted of, and how that testing informed the development and application of a model. This can be used to identify failure points and inform decisions of whether and how to use models based on the documentation of their underlying data and decisions made during the design and development processes.

C. Considerations for Catalyzing AI Capacity in HHS

In the year since the OCAIO was launched, it has promulgated an AI strategy and the *TAI Playbook* for beginning to operationalize that strategy, while working to develop structures to coordinate and harmonize AI efforts across HHS. Looking to the future, should HHS wish to continue prioritizing the introduction of AI-enabled tools across OPDIVs, there may be value to enhancing AI governance at HHS more broadly. Currently, the OCAIO is structured as a convening and coordinating body, with a relatively small staff and limited responsibility for AI implementation across HHS OPDIVs. The Department of Defense (DoD), on the other hand, has tasked its central AI coordinating office with building a department-wide AI infrastructure, fostering talent, and developing a procurement structure that is intended to facilitate rapid introduction of AI tools while promoting responsible use. Though HHS’s efforts are still maturing, monitoring them—as well as other efforts occurring at individual OPDIVs and StaffDIVs within HHS—to identify lessons learned appears to be warranted. Specific roles that the OCAIO or other bodies within HHS may need to foster to catalyze the responsible development of AI tools include:

- Promoting organizational readiness;
- Fostering an HHS-wide AI infrastructure;
- Promoting best practices and good governance approaches across OPDIVs and across the Federal Government;
- Building contracting platforms specific to responsible AI; and
- Supporting social science research regarding responsible AI use in health and human services contexts.

1. Promote Organizational Readiness

HHS could formally assess capacity and readiness for responsible AI use. Such a review could follow the model of a 2020 review by the RAND Corporation of DoD's AI posture (Tarraf et al. 2019).² That study identified barriers slowing DoD's progress, such as:

- A lack of metrics to assess progress;
- Insufficient authorities, visibility, and budget commitments provided to the DoD Joint Artificial Intelligence Center (JAIC)³ to achieve its mission;
- Insufficient Testing Evaluation Verification Validation (TEVV) capabilities to support the safety of AI applications;
- Limits on data, including impediments associated with issues of traceability, accessibility, and interoperability; and
- Lack of mechanisms to develop and retain the needed AI talent.

All of these organizational readiness factors are potentially applicable to HHS's efforts to promote the use of AI in health and human services domains. A structured assessment of organizational readiness across HHS would be valuable in identifying areas for improvement and OPDIV-by-OPDIV challenges to be addressed.

2. Foster an HHS-wide AI Infrastructure

As identified in the OIG AI infrastructure case study, the agency is pursuing efforts for developing infrastructure for AI. OIG is working to support innovation by providing access to container images⁴ and libraries that can be used in the development of AI applications, enabling the consistent deployment of production models. These approaches could be expanded HHS-wide. This could include establishing a repository (similar to DoD's Iron Bank) for sharing code, AI models, or security-hardened containers for AI tools (Air Force n.d.). While the OCAIO has been involved in creating a regularly updated inventory of AI initiatives, additional efforts to build HHS-wide infrastructure for fostering AI efforts would be valuable.

² RAND assessed six factors of DoD's AI efforts: the Organization (vision, strategies, organization structures, stakeholder mandates, authorities and roles), Advancement (research and development portfolio and activities, prototyping, testing, evaluation, verification, and validation), Adoption (procurement, fielding, sustainment, and life-cycle management, development of doctrine), Innovation (internal culture, mechanisms to leverage external innovation), Data (storage, computing, governance of data collection and use), and Talent (talent needed to develop, acquire, sustain and operate, recruitment, retention, cultivation, and growth).

³ As of 2021, the responsibilities of the JAIC have been transferred to the DoD Office of the Chief Digital and Artificial Intelligence Officer (CDAO). The CDAO now serves as the lead organization within DoD for fostering responsible AI (DoD 2022).

⁴ Pre-configured software installations that can be easily managed, deployed, and used at scale to improve performance of production AI systems.

3. Promote and Coordinate Best Practices and Good Governance Approaches across OPDIVs and across the Federal Government

FDA's ongoing regulatory efforts regarding AI/ML-enabled medical devices represent an approach that may be considered for replication HHS-wide. FDA is working to define, for example, *Algorithm Change Protocols* as a mechanism for tracking how an AI-enabled medical device would update while remaining safe and effective or *Good Machine Learning Practice* to establish best practices for using ML. Although FDA's approach is rooted in its specific regulatory authorities, it may be replicable HHS-wide or by other HHS OPDIVs in setting guidelines for or providing technical assistance to State, local, Tribal, and territorial partners.

There may also be opportunities to share information beyond HHS with other Federal entities. For example, with respect to security, the HHS *TAI Playbook* lists sample stakeholders in maintaining the safety and security of AI as being the HHS OCAIO, OPDIV/StaffDIV Office of the Chief Information Officer, Information System Security Owner or Chief Information Security Officer, Database Owner, and the System Owner (HHS 2021b, pages 39, 49, and 60). In addition to these entities, HHS may want to consider engaging with other Federal agencies, such as the DoD CDAO and the Federal Chief Data Officer Council, to coordinate plans and share lessons learned regarding AI security.

Another opportunity for interagency collaboration may occur with respect to TEVV infrastructure. The Johns Hopkins Applied Physics Laboratory, Sandia National Laboratories, and the CDAO are all conducting efforts to build TEVV infrastructure that could potentially be accessed by HHS. HHS has specific needs to ensure the protection of personally identifiable information and compliance laws such as the Health Insurance Portability and Accountability Act that may require HHS to create its own TEVV infrastructure. Existing efforts, however, could serve as starting points for HHS and coordinating with those other Federal TEVV efforts may generate lessons learned for setting up testing infrastructure and point to available existing software tools that HHS could use or adapt.

4. Build Contracting Platforms Specific to Responsible AI

Building responsible use principles into AI is facilitated by contracting vehicles specific for the purpose. HHS should explore adapting (or adopting) the Tradewind acquisition platform being developed by DoD. Tradewind is a specialized AI procurement platform, managed by the Indiana Innovation Institute, that is intended to serve as a bridge between DoD AI procurement needs and developers in industry and academia. The approach makes use of Other Transaction Authority to facilitate rapid, flexible contracting (Tradewind n.d.). DoD is working to pilot processes for ethical and responsibility review for incorporation into the Tradewind platform (DoD 2021). While Tradewind and any processes for incorporating responsible AI practices into DoD's acquisition of AI technologies are still emerging, this example represents a single-agency approach to using procurement policy to ensure that ethical principles are incorporated into Federal agency practice that HHS may benefit from monitoring to identify whether similar approaches could be incorporated into its own implementation of trustworthy AI.

5. Support Social Science Research regarding AI Use in Health and Human Services Contexts

Both the broader recommendations and the incremental options in this chapter reflect concerns that the challenges and risks associated with AI are hindering the optimal development and deployment of AI in health and human services contexts. The tools themselves are changing, as is the societal context in which these tools are embedded. A program of research on using AI in health and human services contexts can help HHS to understand the changing context and to assist the OCAIO and individual OPDIVs in developing and adapting approaches and guidance. Given that most of the opportunities and options identified in this synthesis document focus on the interactions between stakeholders and AI tools rather than on the technologies themselves, any central research program should involve and be rooted in the social sciences, though research should encompass interdisciplinary work that incorporates computing and data sciences as appropriate.

Appendix A

HHS Governance of AI to Date in Light of STPI’s Objective 1 Analyses

STPI’s Objective 1 analysis of Federal AI guidelines (Pratico et al. 2021) identified a set of gaps in the 10 documents reviewed and of practices identified in individual documents that may be relevant for HHS, given its mission to promote health and human services. While STPI’s Objective 1 document was not intended to influence directly the development of guidance by HHS, many of these recommendations align with areas of emphasis in the HHS *TAI Playbook*. In the section below, we discuss the *TAI Playbook* and other elements of HHS governance of AI in light of findings from the analysis of Federal guidelines.

Best-Practice Recommendations from Objective 1 That Are Already Incorporated into Some HHS AI Guidance

Trustworthiness

While STPI’s Objective 1 document noted the importance of trust and trustworthiness of AI solutions, these concepts are central to HHS’s implementation guidance. The Playbook is titled, “Trustworthy AI Playbook” and the introductory message from the HHS Chief AI Officer focuses on the value of maintaining public trust as AI solutions are developed: “As we use AI to advance the health and wellbeing of the American people, we must maintain public trust by ensuring that our solutions are ethical, effective, and secure” (HHS 2021b, page 4). The subsequent page identifies the importance of trustworthy AI in protecting against strategy and reputation risks, cybersecurity and privacy risks, legal and regulatory risks, and operational risks (HHS 2021b, page 5). The balance of the document attempts to further operationalize “trustworthy” AI and how to implement it in health and human services contexts.

Addressing Bias

The *TAI Playbook* primarily addresses concerns regarding bias as part of the operationalization of the Fair/Impartial principle (“AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants”). The *TAI Playbook* identifies approaches to mitigate bias across all stages of the AI development lifecycle:

- at the initiation stage (e.g., checking whether data on marginalized or underrepresented groups may be less accurate given the business problem the system will address, identifying how laws and regulations related to bias and discrimination may apply to the business problem);
- at the design stage (e.g., conducting data bias reviews, involving stakeholders to mitigate unintended bias);
- at the development stage (e.g., checking that training data do not introduce unintended bias); and
- at the operational stage (e.g., checking that the system is not producing unintended bias in outcomes).

Stakeholder Involvement

The *TAI Playbook* makes extensive reference to the importance of involving stakeholders across the initiation, design, development, and operational stages. Stakeholder involvement is referenced explicitly in four of the six principles:

- Fair/Impartial:
 - Stakeholder involvement is part of the definition of the principle (“AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants”).
 - Engagement of stakeholders to define requirements, validate assumptions, and identify potential sources of bias in the model is one of the four key considerations
- Transparent/Explainable: Stakeholder involvement included in three of the four key considerations:
 - Technical and Functional Design (“Have you consulted with stakeholders to ensure that the technical and functional design documentation is understandable?”)
 - Stakeholder Needs (“What people or groups [e.g., regulatory bodies] have an interest in the outputs of your AI solution? Have you engaged them to understand what they need to know about the model to trust the outputs [e.g., decision-making criteria]?”)
 - Security Risks (“How are you communicating information about the model to stakeholders?”)

- **Responsible/Accountable:** Although stakeholder involvement is not included in the principle definition or the key considerations, the rationale for the principle is to ensure that stakeholders are able to gain redress if required (“If an AI medical device fails to identify macular degeneration in a patient who later develops vision problems, there should be clear roles and responsibilities in place to respond to the issue?”)
- **Robust/Reliable:** Stakeholder involvement included in one of the four key considerations: Consistency (“How do you manage new versions of the AI solution? If a new version produces different results, how do you resolve performance issues and communicate this with stakeholders?”)

Focus on Accuracy and Reliability, Understandability, and Transparency

A finding from the Objective 1 analysis is that many of the Federal guidance documents studied provided less detail related to three of the EO 13960 principles. These three principles, however, are strongly incorporated into the *TAI Playbook*, embodied in the Transparent/Explainable and Robust/Reliable principles.

Trade-offs between Principles

The Objective 1 analysis identified the importance for guidelines to consider trade-offs between principles, especially with respect to increasing transparency while preserving privacy, increasing accuracy while preserving fairness, and increasing accuracy while preserving understandability. The *TAI Playbook* makes explicit reference to trade-offs between accuracy and fairness and between accuracy and understandability.⁵ While Privacy is one of the six *TAI Playbook* principles and the document includes an appendix (Appendix II) that describes privacy-protecting technologies in addition to discussing concepts such as consent, protecting personally identifiable information and personal health information, and data sharing, the *TAI Playbook* does not explicitly discuss privacy-transparency trade-offs and how best to address them.

AI in Legislation and Regulatory Context

STPI’s Objective 1 document noted the guidance of the Federal Trade Commission was explicitly grounded in relevant legal and regulatory authorities. The *TAI Playbook* makes explicit reference to grounding AI development in relevant legal and regulatory requirements as part of both the Privacy and the Fair/Impartial principles. In addition, the document includes an appendix (Appendix III) that identifies statutory authorities relevant to the implementation of AI solutions.

A-3

⁵ “There can be tradeoffs between model performance and both fairness and explainability. A model may have a high percentage of accurate predictions, but the model may be replicating historical biases present in the data. Similarly, a deep learning or other similarly complex model may have strong performance metrics, but it may be more difficult to understand and explain the model’s outputs” (HHS 2021b, page 96).

Specific Implementation Guidance

Many of the specific implementation topics that STPI identified in the U.S. Agency for International Development guidance document are addressed in the *TAI Playbook* as well. Avoiding reproducing or enhancing societal inequities, identifying responsible decision makers, developing formal mechanisms for stakeholder communication, rights of redress, and model drift are all incorporated explicitly into the *TAI Playbook*, including as part of implementation checklists. A notable exception is guidance related to accuracy-understandability trade-offs. While implementation checklists refer to AI fairness metrics and bias mitigation algorithms, there are no comparable references to tools for assessing trade-offs between accuracy and model complexity.

AI Governance Structures

HHS launched the OCAIO in March 2021. That office’s functions as of April 2022 are to: “drive implementation of the HHS AI strategy, stand up the HHS AI governance structure, coordinate the HHS response to AI-related federal mandates, and foster collaboration among HHS agencies and offices” (HHS n.d.). While the DoD JAIC (and now the CDAO) is 3 years older, from the time of its initiation in 2018 it was given four missions: (1) delivering AI-enabled capabilities for DoD; (2) establishing a common foundation for scaling the impact of AI including through enterprise-level data stores, frameworks, tools, and standards; (3) facilitating AI planning, policy, and governance; and (4) attracting and cultivating subject-matter expertise within the JAIC and across DoD (DoD 2019). These goals are substantially broader than those of the OCAIO—while they share a common facilitation and collaboration goal, the JAIC and now the CDAO in addition are expected to spur the development of capabilities in support of DoD missions, develop an enterprise AI architecture, and build the DoD’s AI workforce and capabilities. While assessments of DoD’s AI efforts have identified limitations in the implementation of the JAIC and to DoD’s AI implementation to date (Tarraf et al. 2019), the DoD approach is more expansive in its goals—specifically with respect to its responsibility for fostering the incorporation of AI-enabled approaches across DoD’s mission—than HHS’s creation of the OCAIO.

References

- Department of Defense (DoD). 2019. *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*. February 12, 2019. Available from: <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>
- . 2021. “Joint Artificial Intelligence Center to Pilot a Responsible AI Procurement Process.” JAIC Public Affairs, July 2021. Available from: https://www.ai.mil/news_07_27_21-jaic_to_pilot_a_responsible_ai_procurement_process.html
- . 2022. “U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway.” DoD Responsible AI Working Council, June 2022. Available from: <https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF>
- Engstrom, David F., Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government By Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Administrative Conference of the United States. <https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf>
- European Commission High-Level Expert Group on AI. 2019. “Ethics Guidelines for Trustworthy Artificial Intelligence.” Available from: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Executive Office of the President (EOP). 2020. *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*. Executive Order 13960. December 3, 2020. Available from: <https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government>
- . 2021. “Guidance for Regulation of Artificial Intelligence Applications.” White House Office of Management and Budget Memorandum M-21-06, November 17, 2020. Available from: <https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf>
- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. “Datasheets for Datasets.” Available from: <https://arxiv.org/abs/1803.09010v8>
- Karabin, James M., Brian L. Zuckerman, Morgan A. Livingston, Rachel A. Parker, Sharon R. Williams, William E. J. Doane, and Andrew B. Ware. 2022. *Emerging Issues and Needs Associated with Artificial Intelligence (AI) in the Health and Human Services Sectors*. IDA Science and Technology Policy Institute: Washington, DC. (Note: Internal, pre-decisional analysis for HHS)

- Mittra, Arjun and Jeffrey A Moscow. 2019. “Future Approaches to Precision Oncology-Based Clinical Trials.” *Cancer Journal*. 25(4):300–304. doi: 10.1097/PPO.0000000000000383.
- National Institutes of Health. n.d. “Precision Medicine Initiative: Privacy and Trust Principles.” Available from: <https://allofus.nih.gov/protecting-data-and-privacy/precision-medicine-initiative-privacy-and-trust-principles>
- Pratico, Logan M., Andrew B. Ware, Grace E. Hildebrand, Brian L. Zuckerman, Sharon R. Williams, William E.J. Doane, James M. Karabin, and Ian D. Simon. 2021. *Analysis of Artificial Intelligence Guidelines Across the Federal Government*. IDA Science and Technology Policy Institute: Washington, DC. Document D-21619, June 2021. (Note: Internal, pre-decisional analysis for HHS)
- Pratico, Logan M. Andrew B. Ware, Brian L. Zuckerman, Sharon R. Williams, William E.J. Doane, James M. Karabin, Morgan A. Livingston, and Ian D. Simon. 2022. *Comparison of AI Use Cases between HHS and other Federal Agencies*. IDA Science and Technology Policy Institute: Washington, DC. (Note: Internal, pre-decisional analysis for HHS)
- Tarraf, Danielle C., William Shelton, Edward Parker, Brien Alkire, Diana Gehlhaus, Justin Grana, Alexis Levedahl, Jasmin Léveillé, Jared Mondschein, and James Ryseff. 2019. *The Department of Defense Posture for Artificial Intelligence: Assessment and Recommendations*. RAND Corporation, RR-4229-OSD. <https://doi.org/10.7249/RR4229>
- Tradewind. n.d. “About Tradewind.” <https://tradewindai.com/about-tradewind/>
- Urbina, Fabio, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. “Dual use of artificial-intelligence-powered drug discovery.” *Nature Machine Intelligence* 4: 189–191. <https://doi.org/10.1038/s42256-022-00465-9>
- U.S. Air Force. n.d. “Platform One Products and Services.” <https://software.af.mil/dsop/services/>
- U.S. Census Bureau. n.d. “Data Protection and Privacy Policy.” <https://www.census.gov/about/policies/privacy.html>
- U.S. Department of Health and Human Services. n.d. “About the HHS Office of the Chief Artificial Intelligence Officer (OCAIO).” Available from: <https://www.hhs.gov/about/agencies/asa/ocio/ai/ocaio/index.html>
- . 2021a. *Artificial Intelligence (AI) Strategy*. January 2021. Available from: <https://www.hhs.gov/sites/default/files/hhs-ai-strategy.pdf>
- . 2021b. “Trustworthy AI (TAI) Playbook.” September 2021. Available from: <https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf>

Wong, Richmond Y. and Deirdre K. Mulligan. 2018. "Using A Multi-Dimensional Analytic for Privacy Theory, Design, and Analysis." Position Paper for CSCW Workshop on Privacy in Context: Critically Engaging with Theory to Guide Privacy Research and Design. Available from:
https://networkedprivacycsw2018.files.wordpress.com/2018/10/wong_2018_using-a-multi-dimensional-analytic-for-privacy-theorypdf.pdf

Abbreviations

ACF	HHS Administration for Children and Families
ACUS	Administrative Conference of the United States
AI	artificial intelligence
ASPE	HHS Office of the Assistant Secretary for Planning and Evaluation
CDAO	DoD Office of the Chief Digital and Artificial Intelligence Officer
DoD	Department of Defense
EO	Executive Order
EOP	Executive Office of the President
FDA	Food and Drug Administration
IDA	Institute for Defense Analyses
JAIC	DoD Joint Artificial Intelligence Center
HHS	Department of Health & Human Services
ML	machine learning
NIH	National Institutes of Health
OCAIO	HHS Office of the Chief AI Officer
OIG	HHS Office of the Inspector General
OPDIV	HHS Operating Division
OSTP	Office of Science and Technology Policy
PPML	Privacy Preserving Machine Learning
StaffDIV	HHS Staff Division
STPI	Science and Technology Policy Institute
TAI	trustworthy artificial intelligence
TEVV	Testing Evaluation Verification Validation