Responsible AI: Framework for Evaluation

Oai Quoc Tran[1] and Farhana Faruqe[1]

[1]School of Data Science, University of Virginia, 1919 Ivy Rd, Charlottesville, 22903, VA, Country.

Contributing authors: dzn7nf@virginia.edu; xrh6cm@virginia.edu;

Abstract

Here is the introduction and a high-level literature Review, and what I am thinking.

At this stage, I think a shift from Responsible AI for Education to a Responsible AI framework for Evaluation. This is so we can be more flexible in the research question. This is due to the reading and the recent EO from the president to open up federal data.

Research Question:

RQ1. How transparent are U.S. federal agencies about the design, data, and risks of the AI systems they deploy, and what agency characteristics predict higher transparency?

I do use grammarly to assist with writing.

Keywords: AI, Education, Risk

1 Introduction

Recent federal mandates require agencies to disclose their AI use cases as part of a broader push toward responsible and transparent AI adoption. However, early evidence suggests that these inventories are uneven in quality, inconsistent across agencies, and limited in their capacity to support evaluation, oversight, or risk management.

AI has moved from experimental pilot to mainstream infrastructure in digital government. Agencies now apply AI for tasks such as workforce forecasting, education-to-employment matching, benefits administration, fraud detection, citizen engagement, document analysis, and operational optimization. Governments increasingly rely on machine learning (ML), natural language processing (NLP), and now foundation models to support high-stakes public decisions.

However, this expansion has occurred without the corresponding governance capacity necessary to ensure responsible, equitable, and transparent implementation. Federal reviews, including GAO audits, OMB reports, and independent examinations, repeatedly show that government agencies vary dramatically in AI readiness, transparency, and maturity. Some agencies provide detailed disclosures on model design, risk mitigation, fairness assessments, and governance oversight. Others provide almost nothing.

Public-sector organizations are under increasing pressure to disclose how AI systems are designed, deployed, and governed. However, the inventories remain largely descriptive and vary substantially across agencies in depth, specificity, and quality. Agencies differ not only in what they disclose, but also in the clarity with which they communicate data sources, model design, risks, and governance practices. These significant gaps in AI documentation, governance practices, literacy readiness, and evaluation maturity generate hidden risks.

This project provides the first system-level analysis of transparency across the federal AI inventory. By combining the consolidated AI use case listings with agency-level characteristics such as resource capacity, workforce composition, governance maturity, and external oversight. I want a quantitative baseline of the current state of disclosure. Where transparency is strong, where it is inconsistent, and where it is structurally limited by agency capacity. I assume Agencies with higher technical staffing, clearer AI governance structures, or external oversight tend to provide richer and more structured disclosures, while resource-constrained agencies report significantly less detail.

These findings will reveal a deeper challenge: the federal government lacks a shared evaluation and literacy framework for responsible AI adoption. Agencies disclose what they can articulate, not what is required for consistent governance. Critical dimensions of responsible AI risk mitigation, downstream impact, data provenance, fairness considerations, human oversight, and accountability mechanisms are often missing or only loosely referenced. The absence of a standardized reporting structure means that transparency is unevenly distributed across government, creating a governance gap with direct implications for equity, public accountability, and system safety.

The Responsible AI Framework for Evaluation (RAIFE) highlighted systemic problems in how agencies articulate model purpose, disclose risks, and align AI systems with public-value principles such as equity, inclusion, and learner well-being. The same gaps persist and magnify when extended to broader digital government. We present it as a conceptual next step rather than a replacement for existing inventories.

In this way, RAIFE establishes both the baseline and the solution. The empirical results show what exists; the conceptual framework shows what is missing and how to address it. RAIFE is therefore positioned as a governance tool that operationalizes responsible AI principles through a practical, evaluative rubric. It provides agencies with a structured way to map system characteristics to public-value principles, document risks, articulate accountability structures, and support literacy among system developers, implementers, and policymakers.

1. Policymakers lack AI literacy. Most senior leaders agency heads, program directors, legislative analysts cannot reliably interpret model behavior, risk exposure, or system limitations. They frequently rely on vendor claims or internal technical staff without having the conceptual tools to ask meaningful questions. This leads to underspecified procurement, overtrust in algorithms, and failure to identify public-value misalignment.

2. Systems cannot be effectively evaluated. Transparency disclosures vary widely across agencies. Some systems include risk assessments, governance plans, and evaluation metrics; others list only a vague purpose and a brief benefits statement. Without standardized evaluation criteria, agencies cannot compare system maturity, identify gaps, or determine readiness.

3. Risks emerge across pathways, not single models. The greatest harms in digital government rarely come from an isolated model behaving badly. They emerge from AI pathways—multi-step processes where:

- biased or incomplete data
- imperfect classifiers
- policy rules
- resource allocation
- human discretion
- institutional constraints

combine to produce unintended public outcomes. Examples include:

- education-to-employment matching systems funneling low-income students into low-wage occupations
- Risk models reinforce administrative burdens on marginalized groups.
- Workforce forecasts misallocate training resources.
- routing algorithms prioritizing rural or underserved communities

These failures are structural, not solely technical.

This paper introduces Governing AI Pathways, a public-value governance framework grounded in three components:

- AI Literacy for Policymakers
- A structured guide enabling leaders to interrogate AI proposals intelligently.

- System Evaluation Framework
- A multi-dimensional scoring model assessing transparency, fairness, accountability, context alignment, and public-value alignment.
- AI Pathway Risk Mapping ???
-

A method to identify where harm emerges within multi-step AI pathways using data-driven indicators.

We apply this framework to a multi-dataset empirical analysis using federal transparency data, workforce capacity data, budget data, and education-to-employment pathway data.

This paper makes three contributions:

1. A public-value–aligned governance model extending RAIFE to more transparent.???

2. A data-driven evaluation of transparency, workforce capacity, and governance maturity across federal agencies.???

3. A pathway-level risk analysis demonstrating how AI systems can distort opportunity structures in education-to-work transitions.???

Overview Map

Phase 1
Current AI Status: Here I will aim to map all the spending, AI policy etc.

Phase 2
Conceptual Matrix, scorecard, and method to score them.

Phase 3
Propose the solution

2 Background and Research Outline

2.1 Literature Review

Artificial intelligence (AI) literacy has emerged as a foundational requirement for responsible AI integration in education. Early frameworks such as [**?**] highlight competency-based approaches that define what learners and educators must understand to interact safely and effectively with AI systems. Classical perspectives on intelligent tutoring systems [**?**] also provide useful historical grounding for thinking about how AI systems structure learning interactions.

Recent work extends these foundations. Long and Magerko [**?**] identify core competencies for AI literacy, arguing that both conceptual and socio-technical understanding are essential. Research in K–12 contexts [**? ?**] shows that AI literacy is unevenly developed across schools and educational systems, with significant challenges around educator preparedness. Likewise, Zhang and Aslan [**?**] survey AI applications in education and point to the growing need for transparent models and human-centered design.

Beyond literacy, policy frameworks are gaining momentum. UNESCO's 2021 Recommendation on the Ethics of AI [**?**] provides global guidance for human-centered, rights-based AI governance. These principles inform modern educational AI systems and support the argument for structured evaluation frameworks such as RAIFE. The broader educational ecosystem is also evolving rapidly, with more recent texts [**?**] highlighting how leaders plan to integrate AI across classrooms and administrative systems.

Responsible AI Governance in the Public Sector

Responsible AI frameworks from the OECD (2019), UNESCO (2021), NIST (2023), and IEEE (2019) emphasize transparency, fairness, accountability, and human oversight, yet they remain largely at the principle level and offer limited pathways for operationalizing governance in public institutions. Floridi and Cowls' (2019) five ethical principles, beneficence, non-maleficence, autonomy, justice, and explicability, provide a normative foundation, but public agencies continue to struggle with translating these ideals into procurement, risk assessment, documentation, and day-to-day oversight.

Administrative law scholars argue that algorithmic systems challenge traditional accountability structures by obscuring responsibility and reducing contestability (Veale & Brass, 2019). Selbst et al. (2019) show that abstraction errors, the removal of systems from their real institutional contexts, produce governance failures. Raji et al. (2022) further demonstrate that internal audits are insufficient, calling for ecosystem-level third-party oversight to evaluate public-sector AI. These insights collectively argue for context-sensitive, institution-specific frameworks that move beyond broad ethical principles.

AI Literacy & Human-Centered Governance

A parallel body of work highlights a critical literacy gap among policymakers and public administrators. Long and Magerko (2021) define AI literacy as the knowledge, skills, and attitudes needed to meaningfully govern AI-mediated systems. Ng et al. (2021) argue that the public sector requires a distinct form of AI literacy, focused on oversight, risk interpretation, and accountability rather than technical development. Shneiderman (2020) advances the concept of "human-centered AI," emphasizing that non-technical actors must be able to understand and interrogate system behavior to ensure democratic control. This literature suggests that governance failures often stem not from malicious design but from capacity deficits within public institutions.

Transparency, System Evaluation, and Governance Maturity

Transparency tools such as accountable algorithms (Kroll et al., 2017), model documentation (Mitchell et al., 2019), and dataset disclosures (Gebru et al., 2021) offer mechanisms for evaluating system behavior. Edwards and Veale (2017) warn, however, that many automated systems cannot be adequately governed through existing legal and administrative processes. The U.S. Government Accountability Office (GAO) repeatedly identifies gaps in agencies' AI inventories, documentation practices, risk assessments, and oversight plans, revealing a mismatch between policy expectations and institutional readiness. Wirtz and Müller (2019) extend this argument by framing AI adoption in government as a socio-technical process requiring organizational capability, data maturity, and public-value alignment. Their work shows that governance quality depends not only on system performance but on the maturity of institutional processes supporting transparency, evaluation, and accountability.

These lay the foundation for this paper; however, they reveal a deeper challenge: the federal government lacks a shared evaluation and literacy framework for responsible AI adoption. Agencies disclose what they can articulate, not what is required for consistent governance. Critical dimensions of responsible AI risk mitigation, downstream impact, data provenance, fairness considerations, human oversight, and accountability mechanisms are often missing or only loosely referenced. The absence of a standardized reporting structure means that transparency is unevenly distributed across government, creating a governance gap with direct implications for equity, public accountability, and system safety.

2.2 Data Sources

Here are a few data sources for stage 1, this is important because this will establish the foundation for my works.

Dataset Where to Access Format Effort Purpose

- Federal AI Use Case Inventory → CIO.gov, this aims to get the funding related to AI, as AI.gov data is no longer aviable.
- USAspending(budget) api.usaspending.gov → Agency capacity predictor, this future examing government spending.
- FedScope (staffing) fedscope.opm.gov → Agency workforce predictor
- CDO Council list cdo.gov → Data governance maturity

- NIST RMF pilots nist.gov → Risk governance maturity, I need to establish the risk level.
- GAO audits gao.gov → External oversight variable

2.3 Data Steps for Stage 1:

At this stage, I only can share my ideas and how I would go about dealing with the data. Each dataset above should be able to connect. There is a link to connect them so I plan to do a close-read of those datasets.

- Analysis each data sets
- merge and find the missing link
- trace dataset and find the links

3 RAIFE Framework

4 Matrix

Here are some examples I was thinking.

Component	**Policymaker Question**	**Indicator(s)**	**Data Needed**	**Method/Analysis**	**Output**
Purpose	Placeholder	Placeholder	Placeholder	Placeholder	Placeholder
Representation	Placeholder	Placeholder	Placeholder	Placeholder	Placeholder
Model Design	Placeholder	Placeholder	Placeholder	Placeholder	Placeholder
Fairness & Disparities	Placeholder	Placeholder	Placeholder	Placeholder	Placeholder
Governance & Accountability	Placeholder	Placeholder	Placeholder	Placeholder	Placeholder
AI-Risk & Action	Placeholder	Placeholder	Placeholder	Placeholder	Placeholder

Table 1 This is a temp guide to answer. a) what to do? b)with what data? c) What methods? d) the expect outcomes. * Public values define as advance equity, trust, equalities, opportunities for all.

Level	Description	Typical Evidence
0 – Absent	No meaningful documentation; policymakers and public cannot see how the system works.	AI inventory entry is blank or one sentence; no public docs; no description of data or model.
1 – Ad hoc	Some documentation exists but incomplete, inconsistent, or highly technical; policymakers still largely "in the dark."	Minimal inventory fields; an internal slide deck; unclear data provenance.
2 – Defined	Clear documentation of purpose, data sources, model type, and limitations exists and is accessible internally.	Filled inventory; written model card; data sources listed; known limitations documented.
3 – Integrated	Transparency is standardized, updated, and accessible internally and externally; explanations adapted for non-technical audiences.	Public-facing documentation; model cards; FAQs; plain-language summaries; regular updates when system changes.

Table 2 Transparency Matrix

Level	Description	Typical Evidence
0 – Absent	No fairness analysis; no disaggregation by group; "we treat everyone the same" assertions.	No subgroup metrics, no disparity analysis etc.
1 – Ad hoc	Some subgroup metrics are examined informally, but not systemically; no clear fairness thresholds or remediation plan	One-off fairness checks; isolated charts; no standard fairness metrics or process.
2 – Defined	Regular fairness evaluations using agreed metrics; known thresholds for concern; some mitigation steps implemented. listed; known limitations documented.	Documented fairness metrics (e.g., outcome by income/quintile, race, gender); internal fairness review.
3 – Integrated	Fairness is integrated into system design, evaluation, and governance; results influence system changes and policy.	Fairness built into development lifecycle; continuous monitoring; fairness metrics reported to leadership; decisions adjusted based on findings.

Table 3 Fairness - I need to define this !!!!

References

[1] Kasneci, E., et al.: ChatGPT and other LLMs in education **4**, 100111 https://doi.org/10.1016/j.lindif.2023.102274

[2] Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education **16**(1), 39 https://doi.org/10.1186/s41239-019-0171-0

[3] Liu, R., Rivière, J., Roll, I.: Ethics of AI in education: Towards a community-wide framework **32**, 138–170 https://doi.org/10.1007/s40593-021-00239-1

[4] Siemens, G.: Learning analytics: The emergence of a discipline **57**(10), 1380–1400 https://doi.org/10.1177/0002764213498851

[5] Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., Heo, J.: EdNet: A large-scale hierarchical dataset in education https://doi.org/10.48550/arXiv.1912.03072

[6] Long, D., Magerko, B.: What is AI literacy? competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20), pp. 1–16. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

[7] Zhang, K., Aslan, A.: AI technologies for education: Recent research & future directions **2**, 100025 https://doi.org/10.1016/j.caeai.2021.100025

[8] Maher, M.L., Young, M.F.: Artificial intelligence and literacy development in k–12 schools: Opportunities and challenges **36**, 100450 https://doi.org/10.4018/979-8-3693-0205-7.ch004

[9] Stilgoe, J., Owen, R., Macnaghten, P.: Developing a framework for responsible innovation **42**(9), 1568–1580 https://doi.org/10.1016/j.respol.2013.05.008

[10] Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate **3**(2), 1–21 https://doi.org/10.1177/2053951716679679

[11] Hsu, C.-C., Sandford, B.A.: The delphi technique: Making sense of consensus **12**(10), 1–8

[12] UNESCO: Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[13] OECD: AI in Education: Policy Perspectives for Equitable Adoption. https://www.oecd.org/education/ai-in-education.htm

[14] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design: A Vision for Prioritizing Human Well-being. https://ethicsinaction.ieee.org/

[15] Zhai, X., Chu, X., Chai, C.S., Jong, M.S.Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., Li, Y.: A review of artificial intelligence (AI) in education from 2010 to 2020 **2021**(1), 8812542 https://doi.org/10.1155/2021/8812542 . Accessed 2025-12-01

[16] Faruqe, F., Watkins, R., Medsker, L.: Competency Model Approach to AI Literacy: Research-based Path from Initial Framework to Model. arXiv. Version Number: 1. https://doi.org/10.48550/ARXIV.2108.05809 . https://arxiv.org/abs/2108.05809 Accessed 2025-12-01

[17] Schiff, D.: Education for AI, not AI for education: The role of education and ethics in national AI policy strategies **32**(3), 527–563 https://doi.org/10.1007/s40593-021-00270-2 . Accessed 2025-12-01