

To: AI Policy and Regulatory Bodies

From: Elise Racine, University of Oxford

Date: November 20, 2024

RE: Implementing a Human-centered AI Assessment Framework (HAAF) for Equitable AI Development

With: Howard University & Apart Research

Hackathon: Howard University AI Safety Summit & Policy Hackathon, 2024

Implementing a Human-Centered AI Assessment Framework (HAAF) for Equitable Artificial Intelligence Development

Elise Racine, MPA, MSc
University of Oxford

With
Howard University & Apart Research

Executive Summary

Current AI development, concentrated in the Global North, creates measurable harms for billions worldwide. Healthcare AI systems provide suboptimal care in Global South contexts, facial recognition technologies misidentify non-white individuals (Birhane, 2022; Buolamwini & Gebru, 2018), and content moderation systems fail to understand cultural nuances (Sambasivan et al., 2021). With 14 of 15 largest AI companies based in the US (Stash, 2024), affected communities lack meaningful opportunities to shape how these technologies are developed and deployed in their contexts.

This memo proposes mandatory implementation of the Human-centered AI Assessment Framework (HAAF), requiring pre-deployment impact assessments, resourced community participation, and clear accountability mechanisms. Implementation requires \$10M over 24 months, beginning with pilot programs at five organizations. Success metrics include increased AI adoption in underserved contexts, improved system performance across diverse populations, and meaningful transfer of decision-making power to affected communities. The framework's emphasis on building local capacity and ensuring fair compensation for community contributions provides a practical pathway to more equitable AI development. Early adoption will help organizations build trust while developing more effective systems, delivering benefits for both industry and communities.

Keywords: Artificial Intelligence Governance; Community Participation; Data Sovereignty; Impact Assessment; Global Majority; Algorithmic Harm; Human-centered Design; Technology Policy; Participatory Governance; AI Ethics

1. Introduction

1.1 Problem Statement

AI systems increasingly shape critical decisions in healthcare, financial services, and public safety worldwide. However, current AI governance faces two fundamental challenges: (1) extreme concentration of development power - 14 of 15 largest AI companies are US-based (Stash, 2024), and (2) systematic exclusion of affected communities from development and oversight processes. This combination leads to systems that fail in diverse contexts and actively harm vulnerable populations. The scale of impact is significant and well-documented. Healthcare AI systems provide inadequate care in Global South contexts, while facial recognition systems consistently misidentify non-white individuals (Birhane, 2022). Content moderation systems fail to understand cultural contexts, leading to both harmful content and unwarranted censorship (Sambasivan et al., 2021). Meanwhile, communities face ongoing data extraction without compensation or consent.

1.2 Current State Analysis

Existing oversight mechanisms rely primarily on voluntary guidelines and technical audits. These approaches lack enforcement power and treat community input as optional rather than essential. Most critically, they provide no pathway for ensuring communities benefit from AI systems deployed in their contexts. Effective governance must ensure meaningful community participation, provide clear accountability mechanisms, and protect data sovereignty while building local capacity. The root causes of current failures lie in the concentration of power in the Global North, combined with a lack of resourced community participation and insufficient accountability mechanisms. Technical rather than human impacts dominate assessment criteria. Without intervention, rapid AI deployment will further entrench global inequities. This means continued algorithmic discrimination, economic exploitation through

data extraction, erosion of cultural knowledge systems, and limited technological sovereignty for affected communities.

1.3 Stakeholder Analysis

The implementation of AI systems affects multiple stakeholder groups with distinct needs and interests. Global Majority communities face direct impacts on their healthcare, financial access, and daily lives, yet lack meaningful voice in development and deployment decisions. AI developers express concerns about increased costs and extended timelines, while potentially benefiting from improved system effectiveness and broader adoption in currently underserved markets. Regulatory bodies require clear standards and enforcement mechanisms while building local oversight capacity. Civil society organizations play a crucial role in advocating for community interests and supporting implementation. Current power dynamics heavily favor industry interests, with resources and technical knowledge concentrated in the Global North, while communities lack both decision-making authority over deployed systems and mechanisms for ensuring fair benefit-sharing.

This proposal aligns directly with regulatory bodies' mandate to ensure AI benefits all populations while preventing harm. By providing a structured framework for community participation and accountability, HAAF supports both innovation and protection of public interests. The framework particularly advances goals of equitable development, consumer protection, and sustainable technological growth.

2. Methods

2.1 Data Sources & Methodology

Our development of HAAF draws from three primary sources. First, systematic review of existing algorithmic impact assessments reveals significant gaps in implementation effectiveness (Stahl et al., 2023). Second, analysis of documented algorithmic harms provides evidence of current governance failures and their impacts on Global Majority communities (Birhane, 2022; Buolamwini & Gebru, 2018). Third, examination of successful participatory governance models demonstrates pathways for meaningful community engagement (Metcalf et al., 2021). Key limitations in our analysis include the emerging nature of AI governance frameworks and limited data on long-term impacts of community participation models. We address these limitations through pilot testing and iterative framework development with affected communities.

2.2 Key Findings

Analysis of current approaches reveals critical governance gaps. Traditional Algorithmic Impact Assessments (AIAs) have shown promise as accountability mechanisms (Ada Lovelace Institute, 2021; Watkins et al., 2021), but consistently fail to address power imbalances or provide mechanisms for community benefit. Current frameworks treat assessment as a technical checkbox rather than a transformative process, leading to superficial community engagement and limited impact. Successful elements from existing models include structured evaluation protocols, clear enforcement mechanisms, and dedicated resources for community participation. However, these elements typically exist in isolation rather than as part of comprehensive governance frameworks.

2.3 Alternative Approaches

Several alternative governance models were considered. Industry self-regulation through voluntary guidelines offers flexibility but lacks enforcement mechanisms and consistently fails to prevent documented harms. Technical auditing frameworks provide clear metrics but ignore crucial social and cultural impacts. Traditional impact assessments offer structured evaluation but typically lack meaningful community participation.

HAAF improves upon these alternatives by combining three essential elements: participatory design, data sovereignty, and capacity building. The participatory design component ensures communities have real decision-making power through Advisory Boards and compensated participation. Building on established principles of Indigenous data sovereignty (Carroll et al., 2024; Kukutai & Taylor, 2016), HAAF requires community control over data collection and use, with fair compensation for data contributions and protection of cultural knowledge. Long-term success requires sustained investment in community capability through technical training programs, assessment certification pathways, and local governance infrastructure. This integrated approach addresses limitations of existing frameworks while creating clear pathways for community benefit and control. By focusing

on both immediate assessment needs and long-term capacity building, HAAF offers a more sustainable and equitable approach to AI governance.

3. Results (Recommended Solution)

3.1 Policy Recommendation

The Human-centered AI Assessment Framework requires three core elements for effective implementation. Organizations must complete comprehensive pre-deployment assessments that examine power dynamics, cultural impacts, and opportunities for community benefit. Meaningful community participation must be resourced, with a minimum percentage of AI development budgets allocated to engagement and benefit-sharing. Clear data governance frameworks must protect community rights and data sovereignty principles (Carroll et al., 2024), including transparent collection practices and fair compensation for data contributions. Innovation aspects incorporate shifting from voluntary to mandatory assessment, establishing clear resource commitments for community participation, and creating concrete mechanisms for community benefit-sharing. This approach transforms traditional impact assessment from a compliance exercise into a tool for building trust and ensuring equitable AI development.

3.2 Implementation Roadmap

Implementation proceeds through three carefully structured phases:

Phase 1 (Months 1-6): Foundation Building Initial groundwork focuses on infrastructure development and early capacity building. An oversight committee with majority community representation guides initial framework development and pilot selection. This phase emphasizes assessment tool development and establishing sustainable funding mechanisms for community participation.

Phase 2 (Months 7-18): Pilot Implementation Each pilot site establishes a Community Advisory Board with real decision-making power and clear compensation structures. Regular evaluation ensures continuous improvement of both process and outcomes.

Phase 3 (Months 19-24): Scaling and Institutionalization Regional assessment hubs provide local expertise and support while ensuring framework adaptation to specific contexts. Successful practices inform development of compliance guidelines and certification programs.

Required Resources:

- Assessment Platform Development: \$2M
- Community Engagement Programs: \$3M
- Regional Assessment Hubs: \$3M
- Training and Certification: \$2M

Key Dependencies:

- Community participation in oversight committee
- Industry cooperation for pilots
- Technical infrastructure development
- Local capacity building

Risk Mitigation: Industry resistance addressed through clear ROI demonstration and tax incentives for early adopters. Resource constraints managed through multi-stakeholder funding model combining industry contributions, government matching funds, and development grants.

3.3 Success Metrics

Progress will be measured through clear performance indicators across four dimensions:

Trust and Adoption: Increased AI system deployment in previously underserved contexts, higher user satisfaction rates, and growth in local AI initiatives demonstrate framework effectiveness.

Empowerment Indicators: Number of community members in decision-making roles, amount of resources transferred to community control, and level of influence on system design track real power shifts.

System Performance: Reduced instances of algorithmic harm and improved functionality across diverse populations show technical impact.

Economic Benefits: Job creation in affected communities, local AI development initiatives, and fair compensation for data and expertise measure concrete community gains.

Regular monitoring through both quantitative metrics and qualitative community feedback ensures accountability and enables framework adaptation as needed.

4. Discussion (Impact Assessment)

4.1 Expected Outcomes

Implementation of HAAF will generate impacts across multiple timeframes. In the immediate term (Years 1-2), structured pathways for community participation will prevent harmful AI deployments while building trust between developers and communities. Pilot programs indicate that organizations adopting this framework see rapid improvements in system performance across diverse populations, alongside increased community trust in their AI initiatives. As the framework becomes standard practice in the medium term (Years 3-5), we expect to see fundamental shifts in global AI development and deployment. Regional assessment hubs will emerge as centers of expertise, while standardized assessment processes will streamline the development of trustworthy AI systems. Community Advisory Boards will transition from primarily oversight roles to become active partners in AI development. Long-term impacts (5+ years and beyond) include transformation of global AI development practices, emergence of new innovation centers in Global Majority regions, and sustainable mechanisms for community-led AI governance. This shift promises to fundamentally alter power dynamics in global AI development while ensuring technologies better serve diverse human needs.

4.2 Stakeholder Impacts

Implementation affects stakeholder groups differently. Communities gain direct input into AI systems affecting their lives, fair compensation for their contributions, and increased technical capacity. AI developers initially face increased costs but benefit from improved system effectiveness, broader market adoption, and reduced risk of harmful impacts. These challenges can be addressed through tax incentives, clear ROI demonstration, and technical support during transition periods.

Regulatory bodies receive clear standards and metrics for enforcement, while building local oversight capacity. Civil society organizations gain structured ways to advocate for community interests. Potential unintended consequences include initial slowdown in AI deployment timelines and possible resistance from industry stakeholders. Regional technical assistance centers and flexible implementation guidelines help organizations navigate these challenges while maintaining framework integrity.

4.3 Feasibility Analysis

Political Feasibility: Framework builds on existing impact assessment mechanisms while addressing clear public harms, increasing likelihood of political support. Multi-stakeholder funding approach distributes costs and benefits across participants.

Economic Feasibility: Initial \$10M implementation cost is modest compared to AI industry resources and potential benefits. Return on investment through increased adoption and reduced harm provides clear business case.

Technical Feasibility: Open-source assessment tools and standardized protocols make implementation straightforward. Existing impact assessment infrastructure can be adapted for framework requirements.

Administrative Feasibility: Phased implementation allows organizations to build capacity gradually. Clear protocols and regional hubs provide ongoing support for assessment processes.

5. Conclusion

The rapid global deployment of AI systems demands immediate action to ensure these technologies benefit all populations. HAAF provides a practical, implementable solution to current governance gaps, offering structured protocols for meaningful community participation in AI development. The initial investment of \$10M and 24-month implementation timeline represents a modest commitment compared to both the potential benefits and costs of inaction. Early adoption of HAAF can help organizations build trust while developing more effective and widely adopted AI systems. By emphasizing genuine community participation, fair compensation, and local capacity building, this framework transforms how AI systems are developed and deployed. Now is the time for implementation—before further entrenchment of harmful practices and inequities. This investment in human-centered AI governance will pay dividends not only in prevented harms but in increased trust, adoption, and effectiveness of AI systems globally.

6. References

- Ada Lovelace Institute. (2021). Algorithmic impact assessment: A case study in healthcare. Retrieved from <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/02/Algorithmic-impact-assessment-a-case-study-in-healthcare.pdf>
- Birhane, A. (2022). The unseen Black faces of AI algorithms. *Nature*, 610, 451-452. <https://doi.org/10.1038/d41586-022-03050-7>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77-91). PMLR.
- Carroll, S. R., Duarte, M., & Liboiron, M. (2024). Indigenous data sovereignty. In *Keywords of the datafied state* (pp. 207-223). Data & Society.
- Kukutai, T., & Taylor, J. (2016). *Indigenous data sovereignty: Toward an agenda*. ANU Press.
- Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 735-746).
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
- Stahl, B.C., Antoniou, J., Bhalla, N., et al. (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, 56, 12799-12831.
- Stash. (2024, August 8). 15 largest AI companies in 2024. Stash Learn.
- Watkins, E. A., Moss, E., Metcalf, J., Singh, R., & Elish, M. C. (2021). Governing with algorithmic impact assessments: Six observations. *AIES '21*.