

Reparative Algorithmic Impact Assessments

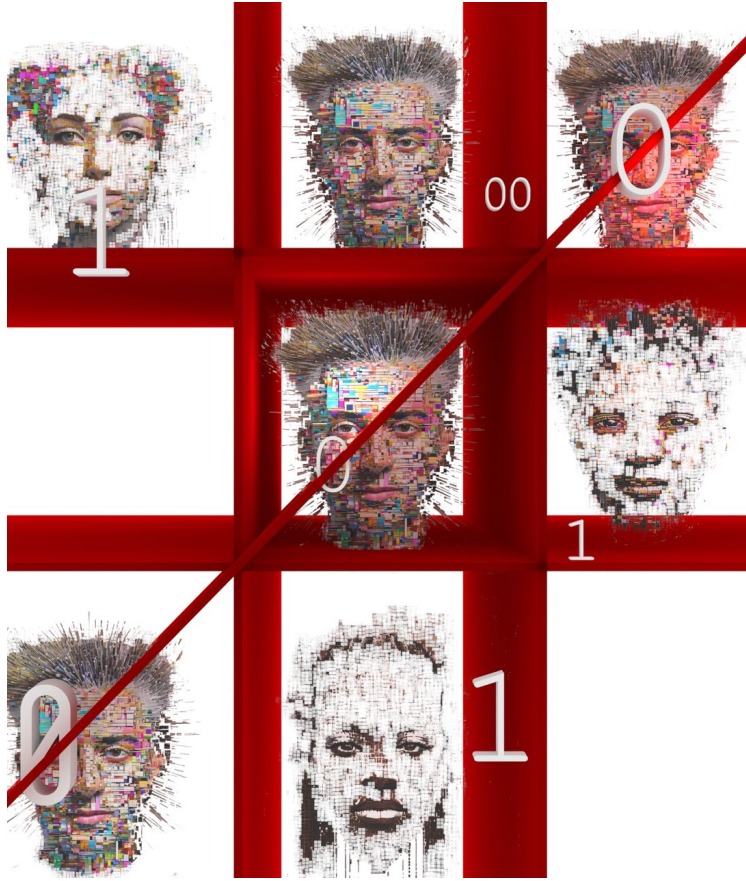
*A Human-Centered,
Justice-Oriented
Accountability Framework*

Elise Racine, MPA, MSc
Ethics in AI Institute & Ethox Centre
PhD Candidate, University of Oxford



Clarote & AI4Media / Better Images of AI / Power/Profit / CC-BY 4.0

✱ Apart + Johns Hopkins AI Policy Hackathon



Amritha R Warrior & AI4Media / Better Images of AI / tic tac toe /
CC-BY 4.0

The Challenge of Global AI Inequity

Global majority critically underrepresented in AI lifecycle

Western-centric AI Development creating systemic harm

Urgent need for new accountability frameworks

This is leading to documented harms...

Commercial facial-analysis technologies consistently misidentify individuals who are not white cisgender men (Birhane, 2022; Buolamwini & Gebru, 2018), enabling discrimination, privacy violations, wrongful policing, and the reinforcement of harmful stereotypes.

Credit: Dr Joy Buolamwini (CC BY 4.0)

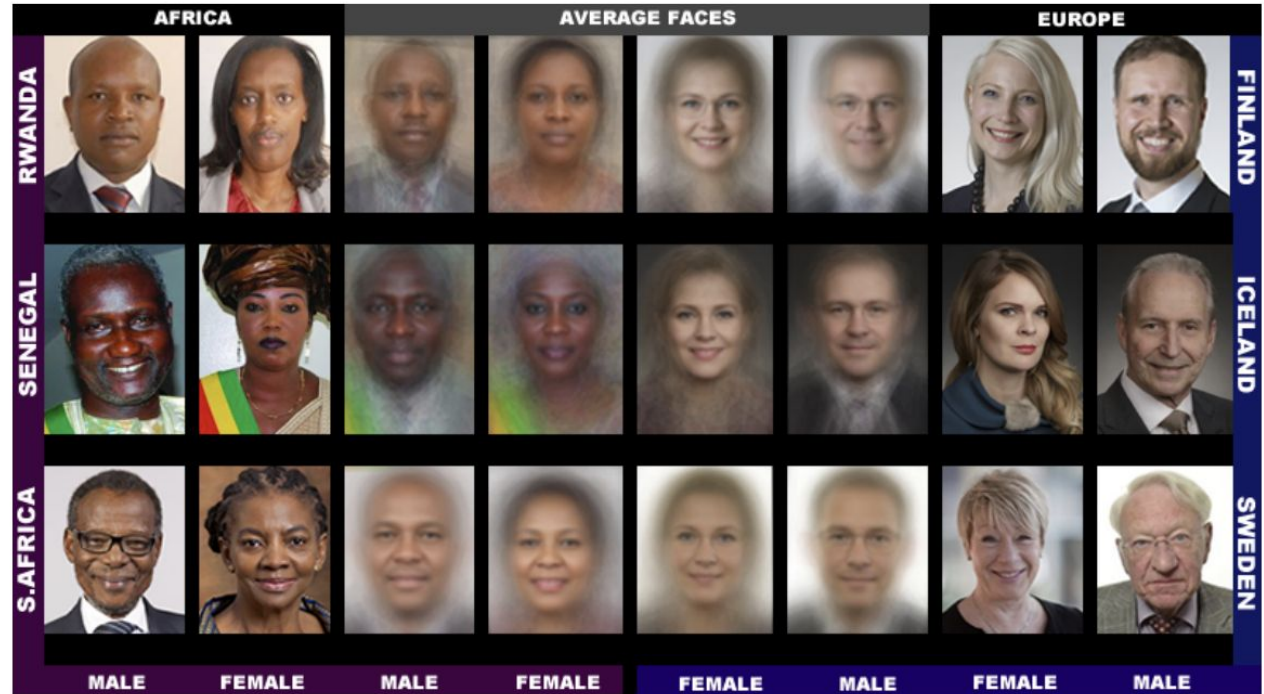


Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

And raises critical questions re: safety alignment....

Current AI safety efforts often fail to incorporate diverse conceptions of human values and well-being, creating systems that may be 'aligned' with Western values but potentially harmful in other cultural contexts.

For example, AI mental health chatbots trained on Western psychological frameworks may recommend individualistic coping strategies that conflict with collectivist cultural values, or fail to recognize the importance of spiritual practices and community healing methods common in many Global Majority societies. In general, Healthcare AI systems trained primarily on Western populations deliver suboptimal care in Global South contexts, directly affecting patient outcomes and reinforcing global health disparities. Another example illustrated how Content moderation systems struggle to interpret culturally specific expressions, leading to both unwarranted censorship/suppression and failure to identify truly harmful content in diverse cultural contexts (Sambasivan et al., 2021).

Other examples...

Critical AI functions like data labeling are routinely outsourced to the Global Majority, often subjecting workers to traumatic conditions and low pay while limiting career advancement opportunities (Igarapé Institute, 2024).

The concentration of AI power—with 14 of the 15 largest AI companies being US-based (Stash, 2024)—severely limits technological sovereignty for many in the Global Majority and reinforces global power imbalances.



Clarote & AI4Media / Better Images of AI / Labour/Resources / CC-BY 4.0

Key Limitations of Traditional AIAs...

1

Superficial Compliance: (a) Treats impact assessment as technical checkbox, (b) Lacks transformative potential, and (c) Focuses on documentation over change

2

Power Blindness: (a) Fails to address underlying power asymmetries, (b) Ignores historical context, and (c) Maintains existing power structures

3

Limited Participation: (a) Lacks mechanisms for meaningful community engagement, (b) Often relies on superficial consultations, and (c) Missing authentic community voice in assessment.

4

Reactive vs. Proactive: (a) Focuses only on avoiding harm, (b) Missing opportunities for community empowerment, and (c) Fails to build community capability and power

Traditional AIAs become performative exercises that fail to drive systemic, structural change.



Introducing Reparative Algorithmic Impact Assessments (R-AIAs)...

A Transformative Solution

R-AIAs combine robust accountability mechanisms with reparative praxis to create a methodology that is more:

- * Culturally sensitive
- * Justice-oriented
- * Human-centered

Key Components

1. Deep consideration of historical context and lived experiences,
2. Thorough analysis of power dynamics and asymmetries,
3. Commitment to meaningful community engagement and harm co-construction,
4. Incorporation of decolonial, Intersectional principles that recognizes human diversity and the complex interplay of various aspects of identity, and
5. Ongoing monitoring, evaluation, and adaptation based on real-world impacts,
6. Focus on sustainable development, long-term impacts, and opportunities for redress.

R-AIAs also incorporate the following decolonial, Intersectional principles:

- 1.** Epistemological diversity: Actively incorporating diverse knowledge systems into AI development.
- 2.** Data sovereignty: Respecting the rights of communities to control their data.
- 3.** Technological self-determination: Empowering communities to develop and deploy AI that aligns with their values and needs.
- 4.** Cultural preservation: Ensuring AI respects and promotes cultural diversity.
- 5.** Reciprocity: Establishing mutually beneficial relationships between AI developers and communities.

Through these principles, R-AIAs help create AI systems that enhance rather than diminish human agency, reflecting and respecting the full diversity of human experience.



While R-AIA implementation requires rethinking traditional approaches, we propose a framework that is **technically feasible and policy-actionable**. Building upon existing impact assessment mechanisms and governance structures, we outline **six interconnected steps** that both **incorporate the above principles** and **can be readily operationalized within current regulatory frameworks**. Each key step also includes sample strategie(s) and/or practice(s) that align or misalign with the R-AIA approach, illustrating the framework in action.

To exhibit practical implementation, we use the following case study...

A US-based technology company is piloting an AI-powered chatbot providing 24/7 mental health support to underserved communities across rural India.

This case study demonstrates how R-AIAs can effectively prioritize human dignity and agency while deploying AI systems in Global Majority contexts.



Key Considerations:

Cross-cultural mental health support

Rural accessibility challenges

Traditional healing practices

Community health structures

Language and cultural nuances

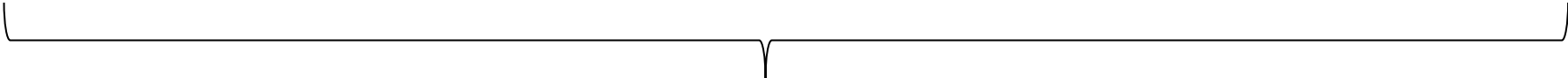
Data privacy and sovereignty

Local healthcare integration

Sustainable support systems

Step 1: Socio-Historical Research

This foundational step involves thorough investigation into the context in which the AI system will operate.



Implementation Example: *Librarians and information specialists with expertise in data curation conduct comprehensive research into... Traditional mental health practices in rural India, Previous technological interventions in community healthcare, Historical patterns of medical colonialism, Local healing systems and knowledge frameworks, Community health worker networks, Existing mental health support structures, Cultural attitudes toward mental health, and Language-specific expressions of mental well-being*

Step 2: Power Analysis

Examining multiple dimensions of power reveals how the AI system might reinforce or transform existing structures.

Key Power Dimensions Analyzed

(for our case study)...

Institutional Power: Decision-making authority over mental health services

Knowledge Power: Whose mental health expertise is valued

Data Power: Control over sensitive health information

Economic Power: Resource distribution in healthcare

Technological Power: Access to and control of AI systems

Narrative Power: Who shapes mental health discourse

Local Power Structures: Community healthcare hierarchies

Step 3: Participant Engagement and Co-Construction

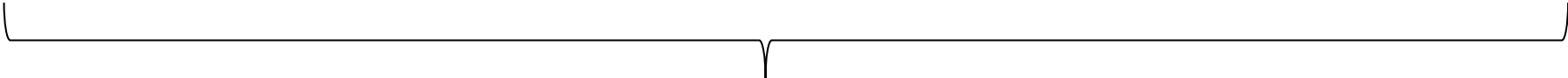
Meaningful engagement ensures impacts and potential harms are co-constructed with affected communities.

Implementation Approach: *Utilize socio-historical research to inform participant recruitment. Implement flexible participation options, Provide compensation for community input, Offer mental health support for participants, Ensure accessibility needs are met, Create clear feedback channels, and Enable direct influence on chatbot functionality.*

Avoid: *Superficial consultations with high-prestige experts while neglecting input from affected communities.*

Step 4: Sovereign and Reparative Data Practices

Ensuring communities retain control over their mental health data while building inclusive data methods.



Implementation Strategies: *Establish community-controlled data trusts, Develop local data governance frameworks, Create transparent data usage policies, Implement strong privacy protections, Enable community audit rights, Support local data infrastructure, and Build data literacy programs.*

Key Focus: *Respect Indigenous data sovereignty principles while ensuring communities maintain control over sensitive mental health information.*

Step 5: Ongoing Monitoring and Adaptation

Continuous evaluation and adjustment based on real-world impacts and community feedback.

Implementation Measures:

- Develop culturally-appropriate success metrics
- Monitor community-defined outcomes
- Track system performance across groups
- Measure cultural competency
- Evaluate accessibility goals
- Assess community satisfaction
- Enable rapid adjustments
- Document learnings and adaptations

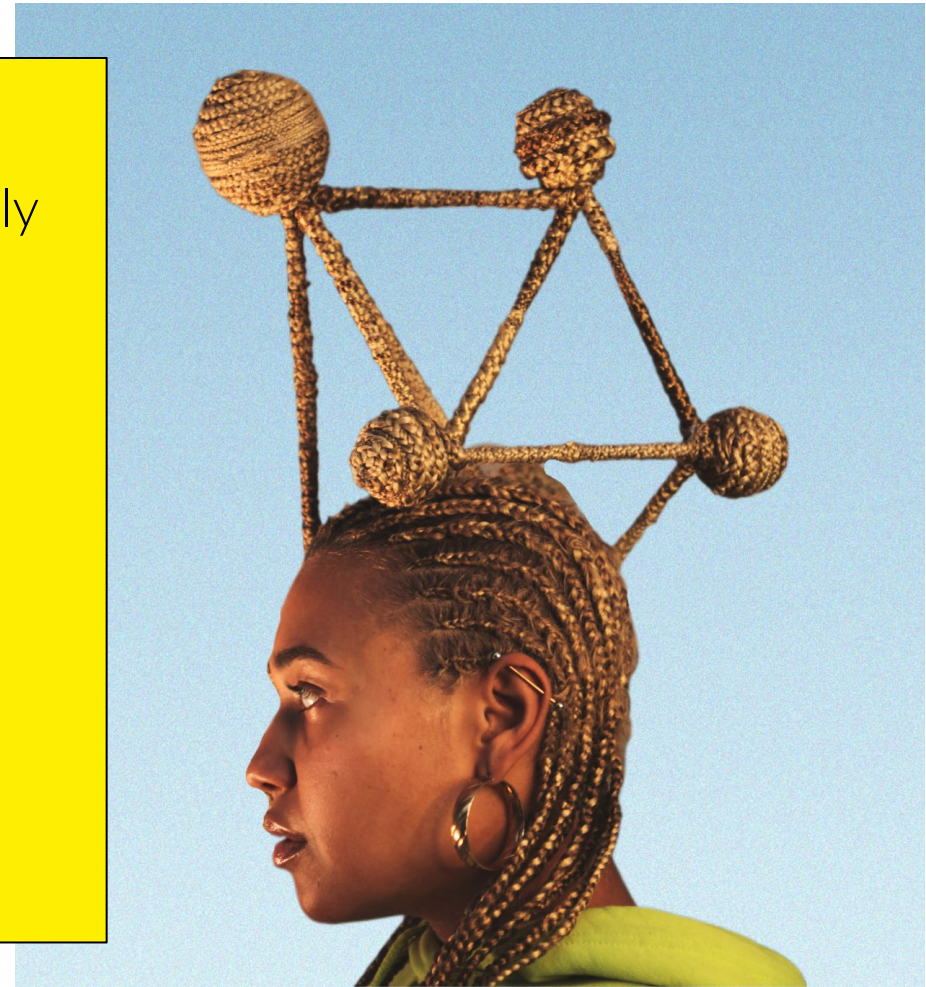
Aligned Practice: *Develop new ways of measuring chatbot performance that incorporate diverse cultural values and priorities, challenging the dominance of Western-centric benchmarks in AI evaluation.*

Step 6: Redress

Moving beyond identification to actively address and repair potential harms.

Sample Implementation Actions:

- Partner with local AI research institutes
- Offer training and education programs
- Invest in local infrastructure
- Provide cloud computing access
- Create scholarship opportunities
- Support local mental health initiatives
- Enable community-led development
- Build sustainable support systems



The **dominance of Western-centric** AI development has created systems that often reinforce rather than reduce global inequities. Through R-AIAs, we present a **feasible and impactful solution** that moves beyond traditional accountability frameworks. In combining culturally sensitive participatory methods with a reparative praxis and decolonial, Intersectional principles, our framework **addresses critical gaps in current AI governance**. Specifically, it (a) ensures meaningful participation, (b) redistributes power, and (c) protects diverse knowledge systems. Through our **phased implementation approach**, organizations can systematically transform their AI development practices to create more equitable and effective systems.

Most importantly, R-AIAs advance both the technical and social aspects of AI safety by ensuring alignment mechanisms reflect global diversity in human values and conceptions of well-being.

By prioritizing **human flourishing** for all those impacted by these systems, we can work toward an AI future that truly serves and empowers the Global Majority. The time for this transformation is now, and R-AIAs provide the structured pathway to make it possible.



Image by Alan Warburton / © BBC / Better Images of AI / Quantified Human / CC-BY 4.0



Thank you!

Elise Racine

elise.racine@ndph.ox.ac.uk

www.linkedin.com/in/eliseeracine