Integrated Human Rights Impact

Assessment & Reflection Tool for *AI Systems*

A comprehensive tool combining formal assessment with ongoing reflection throughout the AI lifecycle

AI & EQUALITY

<AI & Equality> Human Rights Toolbox
Authored by Emma Kallina,
Sofia Kypraiou, & Caitlin Kraft-Buchman
Designed by Pilar Grant

About this tool

This integrated Human Rights Impact Assessment (HRIA) tool combines two essential approaches for responsible AI development:

Ongoing Reflection Tool:

Use the stage-specific reflection questions throughout development for team discussions, planning meetings, and design sprints. These prompts help internalize rights-based thinking and ensure human rights considerations "become relevant at each stage" rather than as an afterthought.

Formal Assessment Documentation:

Complete the structured assessment sections for official documentation, regulatory compliance (like EU AI Act requirements), and accountability. This creates a permanent record of your human rights due diligence process.

Three ways to use this tool:

1

Progressive Development Approach

- Work through each stage sequentially as you develop your AI system.
- Use reflection questions for team discussions and critical thinking sessions.
- Fill out formal assessment sections at the end of each stage.
- Build up a comprehensive HRIA document over time.

2

Milestone-Based Assessment

- Use reflection questions for ongoing team check-ins and planning.
- Complete formal assessment sections at key project gates (design review, pre-deployment, etc).
- Conduct comprehensive reviews at major milestones.

3

Comprehensive Evaluation

- Complete the entire tool at specific points (pre-development, pre-deployment, annually).
- Ideal for compliance requirements or external audits.
- Can be used retroactively to assess existing systems.

Key Principles

Keep it Simple

The tool uses plain language accessible to teams without legal expertise.

Make it Participatory

Always involve affected communities - they are the experts in their own context.

Document Everything

Record decisions, rationale, and changes to enable accountability and learning.

Stay Flexible

The AI lifecycle is iterative - revisit and update sections as you learn more.

Focus on Action

Every identified risk should have corresponding mitigation measures.

Who Should Use This Tool:

This tool can be adapted for any scale - from small pilot projects to national AI rollouts - and helps fulfill both ethical duties and emerging legal requirements for human rights impact assessments in AI systems.

- **Academic researchers** piloting AI systems in real-world contexts
- **Development organizations** working on AI for social good
- **Public sector agencies** developing or procuring AI systems
- **Private sector teams** committed to responsible AI practices
- **NGOs and civil society organizations** implementing technology solutions

Project Overview & Context

Fill out the spaces with the relevant information to your project.

AI System/Project Name:

Implementation Context Scale:

☐ Pilot ☐ Regional ☐ National ☐ International

Date Assessment Started:

Project timeline:

Start date: Full Deployment:

Last updated:

Assessment Team:

Sector:

☐ Healthcare ☐ Education ☐ Finance ☐ Justice

☐ Social Services ☐ Other:

Project Purpose and Social Goal:

What problem is this AI system addressing, and why does it matter for our community? Be clear about both the technical goal AND the human rights/ social goal.

Historical Context & Power Dynamics

Are there historical biases or power imbalances in this problem area? (e.g., discrimination in policing, credit scoring, healthcare) How will we learn from history to avoid repeating injustices?

1 Stage 1: **Objective + Team Composition**

Reflection Questions for Team Discussion

Use these questions for planning meetings, design sprints, and critical thinking sessions

Purpose and values:
- Have we clearly stated the social or human rights goal (not just the technical goal) of the project, derived in collaboration with affected groups?
- Is an AI system even the best way to address the issue? Have we explored alternatives, technical as well as non-technical?

Team Inclusivity:
- Who is involved in defining the project?
- Does our team include diverse perspectives (women, community members, domain experts, affected groups)?
- What key voices are missing, and how will we bring them in?

Community Agreement:
- Have affected communities been consulted and agreed that this AI system is needed?
- Do the most impacted or at-risk groups have the power to stop the project?
- Do they have actual decision-making power, or are we just extracting their input?

Formal Assessment: Stakeholders & Team Composition

Primary Responsible Organization:

Affected Communities and Stakeholders

List all groups who will be impacted by this system.

Direct Users:	**Vulnerable/Marginalized Groups:**
Communities Affected by Decisions:	**Key Partners/ Collaborators:**

Individuals accountable for potential harms:

Community Engagement Status

- ☐ Affected communities consulted and involved in design with decision-making power
- ☐ Limited consultation conducted - plan to expand: _____
- ☐ No consultation yet - planned for: _____
- ☐ No consultation planned

Red flag, reconsider approach.

Core Development Team

TIP: You can also create a spreadsheet

Name	Role	Relevant expertise	Demographic background

Team Diversity Assessment

- ☐ Gender diversity
- ☐ Domain/sector expertise
- ☐ Human rights expertise
- ☐ Cultural/ethnic diversity
- ☐ Lived experience with the problem
- ☐ Disciplinary diversity
- ☐ Social science expertise

Relationship of team members:
You should pursue all four!

- ☐ Flat hierarchy
- ☐ Shared language
- ☐ Close collaboration
- ☐ Mutual learning

Gaps Identified:	Plan to Address Gaps:

2 Stage 2: **Define System Requirements**

Reflection Questions for Team Discussion

Use these questions for planning meetings, design sprints, and critical thinking sessions

Community Needs:
- Who are the end-users or affected community, and have we asked them what they truly need?
- Are the system's requirements aligned with what these communities value, or are we imposing assumptions?

Participation in Design:
- Who is at the table when deciding features and requirements?
- How did you empower affected people to shape these requirements, especially representatives of at-risk groups?

Fairness & Trade-offs:
- When setting success metrics, did we consider equality measures?
- What other success criteria are we optimising for, e.g. privacy, accountability, interpretability, or transparency?
- Did you allow affected communities to substantially add to and influence this list of success criteria and how they can be fulfilled?
- What trade-offs are we making, and who could be negatively impacted?
- Are we documenting the rationale for these decisions?

Formal Assessment: System Requirements

Core System Requirements

Developed in dialogue with affected communities

1 Functional Requirements:	2 Performance Requirements:	3 Human Rights Requirements:

Ecosystem of Values - Managing Trade-offs

Accuracy vs. Other Values: Are there tensions between accuracy and other necessary metrics/ success criteria in this context?

How do you handle these trade-offs?

Explainability Requirements

Target audiences (including technical and non-technical groups):

What is the goal of explanations for each target audience?

Level needed:

- ☐ Global *(how system works generally)*
- ☐ Local *(individual decisions)*
- ☐ In-depth explanations for debuggin
- ☐ Accessible explanations for non-technical users
- ☐ Explanations accessible to affected non-users to increase transparency
- ☐ Other:

Fairness Considerations

Which fairness metrics are most relevant?

- ☐ Individual fairness
- ☐ Equal opportunity
- ☐ Other:
- ☐ Group fairness
- ☐ Demographic parity

For which (protected) attributes do you test these:

Rationale/justification of fairness-related decisions:

Privacy & Data Protection

Which measures did you take?

- ☐ Minimize data collection
- ☐ Anonymization required
- ☐ Informed consent mechanisms
- ☐ Right to deletion
- ☐ Data portability
- ☐ Other:

Transparency Commitments:

- ☐ Open methodology
- ☐ Training data documentation
- ☐ Technical performance metrics public
- ☐ Non-technical success criteria public
- ☐ Regular reporting
- ☐ Details on your business model
- ☐ Public HRIA results
- ☐ Other:

Accountability Structure

Who has oversight?:	Human-in-the-loop requirements:	Appeal/contest mechanisms:

What are responsibilities and timelines for reacting to feedback pointing to negative human rights impacts?

Community Feedback Integration

How have **community needs** shaped the above requirements?

3 Stage 3: Data Discovery & Preparation

Reflection Questions for Team Discussion

Use these questions for planning meetings, design sprints, and critical thinking sessions

Representation:
- Who is represented in our data – and who is not?
- Does it include different groups that might use or be subject to the AI system?
- If populations are missing, how will we address that?

Source & Consent:
- Where is the data coming from? Is it collected respectfully with informed consent?
- Are there privacy issues or data protection considerations?

Bias Analysis:
- Could the data contain systemic biases or historical prejudice?
- Have we done bias analysis and / or asked domain experts for insights on potential biases?
- What biases have we identified and how will we address them? Think of technical and non-technical methods!

Quality & Gaps:
- Are there limitations that might affect effectiveness for some groups?
- Might this contribute to affirming existing power structures?
- How will we deal with this, fill gaps, or adjust expectations?

Formal Assessment: Data Discovery & Preparation

Data Sources & Origin

TIP: You can also create a spreadsheet

Dataset	Source	Original Purpose	Consent Status	Sensitivity Level

Bias Assessment

Historical biases identified in data and/ or through domain expert involvement:

1. Bias/ Mitigation approach:
2. Bias/ Mitigation approach:
3. Bias/ Mitigation approach:

Pre-processing

Pre-processing steps taken:

- ☐ Data augmentation for underrepresented groups
- ☐ Re-balancing/ re-weighting
- ☐ Bias-aware sampling
- ☐ Synthetic data generation
- ☐ Other: _____

Representativeness Analysis

Demographics Included:

- ☐ Gender: _____
- ☐ Ethnic/ cultural groups: _____
- ☐ Age groups: _____
- ☐ Socioeconomic status: _____
- ☐ Geographic regions: _____
- ☐ Languages: _____
- ☐ Other relevant categories: _____

Demographics Underrepresented or Missing:	**Impact of gaps:**	**Mitigation strategy:**

Data Quality Assessment

Quality variations across groups:	**Limitations and their implications:**

Privacy & Rights Protection

- ☐ Data minimization applied
- ☐ Data retention policies defined
- ☐ Anonymization/pseudonymization used
- ☐ Right to deletion procedures
- ☐ Consent obtained and documented
- ☐ Cross-border transfer protections

4 Stage 4: Model Development & Selection

Reflection Questions for Team Discussion

Use these questions for planning meetings, design sprints, and critical thinking sessions

Model Choice:

- Why have we chosen this type of model/algorithm?
- Is it the right balance between complexity and explainability for the context and its risk-level?
- Can we justify our choice in terms of performance AND values alignment?

Fairness Interventions:

- Are we using techniques to mitigate bias, which ones, and have we documented these appropriately?
- What makes us believe that the model is fair enough?

Explainability & Accessibility:

- How do we ensure that affected groups are aware that AI is used on them?
- What's our plan to make decisions understandable to affected groups?
- How do we ensure that these explanations are appropriate for our audience and context?
- If it's a "black box," what compensatory measures do we have?

Formal Assessment: Model Development & Selection

Model Architecture & Rationale

Model Type Selected:

- ☐ Linear/Logistic Regression
- ☐ Decision Tree
- ☐ Random Forest
- ☐ Neural Network
- ☐ Deep Learning
- ☐ Ensemble
- ☐ Other: _____

Justification for Choice:	Technical reasons:	Non-technical reasons: *(explainability, fairness, etc.)*

Complexity Assessment

- ☐ Simple, highly interpretable
- ☐ Complex but explainable
- ☐ Moderately complex
- ☐ Black box - compensatory measures needed

Explainability Implementation

Based on what you planned at stage 2

Methods used to implement local explanations (if any):

- ☐ Feature importance
- ☐ LIME
- ☐ SHAP
- ☐ Example-based *(What would change outcome?)*
- ☐ Counterfactual explanations
- ☐ Natural language explanations
- ☐ Visual dashboards
- ☐ Decision trees
- ☐ Other:

Methods used to implement global explanations (if any):

- ☐ Overall feature weights
- ☐ Natural language explanations
- ☐ Visual dashboards
- ☐ Decision trees
- ☐ Other:

Fairness & Bias Mitigation

Techniques applied:

Pre-processing	In-processing	Post-processing	No specific techniques:
☐ Data augmentation	☐ Fairness constraints during training	☐ Output adjustment	Rationale: _____
☐ Re-weighting	☐ Adversarial debiasing	☐ Threshold optimization	
☐ Synthetic data			

Fairness Metrics Implemented:

- ☐ Demographic parity
- ☐ Individual fairness
- ☐ Equal opportunity
- ☐ Counterfactual fairness
- ☐ Equalized odds
- ☐ Other: _____

Intersectionality Considerations:

How are we addressing multiple, overlapping identities?

Audience-Specific Adaptations:

For which technical and non-technical audiences are your explanations intended?

How are your explanations tailored for these different groups?

Language/accessibility considerations:

Environmental & Resource Impact:

Computational resources required:

Environmental cost assessment:

Mitigation/offset measures:

5 Stage 5: Testing & Validation

Reflection Questions for Team Discussion

Use these questions for planning meetings, design sprints, and critical thinking sessions

Inclusive Testing:

- Who is testing the system? Are we including people beyond customers, e.g. intended users and those impacted?
- What feedback have we gotten and how are we incorporating it?

Performance Across Groups:

- Have we measured performance disaggregated by different subgroups?
- Are there disparities in error rates or outcomes?
- How are we addressing any disparities found?

Meeting Objectives:

- Does the system actually solve the problem we defined initially?
- Are there unintended outcomes? Are we ready to cycle back to earlier stages?
- How do we ensure that all technical and non-technical success criteria that we defined in stage 2 are met?

Transparency:

- Are we being transparent about limitations and uncertainties?
- Have we documented all known issues and incorporated them in a training manual for system users?

Formal Assessment: Testing & Validation

Testing Methodology

Test Datasets:

Training data:

Validation data:

Test data:

Real-world pilot data:

Participants Assessing whether the system meets our success criteria:

- ☐ Technical team only
- ☐ Domain experts
- ☐ Intended users
- ☐ Affected communities
- ☐ External auditors
- ☐ Other: _____

Performance Disparities Identified

1 Issue: _____

Groups affected: _____

Severity: ☐ Minor ☐ Moderate ☐ Significant ☐ Severe

Action taken: _____

2 Issue: _____

Groups affected: _____

Severity: ☐ Minor ☐ Moderate ☐ Significant ☐ Severe

Action taken: _____

Objective Achievement Assessment:

- ☐ Fully meets original objective, success criteria, and community needs
- ☐ Does not meet objective - requires significant changes
- ☐ Partially meets objective - gaps identified: _____
- ☐ Objective should be reconsidered based on learning strategy to address any identified gaps before deployment: _____

Unintended Consequences Identified:

Performance Analysis

Overall Performance Metrics:

Accuracy: _____ **Precision:** _____

Recall: _____ **F1-Score:** _____

Other relevant metrics: _____

Disaggregated Performance Analysis:

Performance broken down by relevant demographic/ social groups.

Group	Accuracy	Precision	Recall	Error Types	Sample Size

User Feedback & Community Testing

Feedback Collection Methods:

- ☐ Surveys
- ☐ Focus groups
- ☐ Individual interviews
- ☐ Community meetings
- ☐ Pilot programs
- ☐ A/B testing
- ☐ Other: _____

Key Feedback Themes:

1. _____
2. _____
3. _____

Changes Made Based on Feedback:

Limitations & Uncertainties

Known technical limitations:

Known bias or fairness limitations:

Uncertainty in predictions for specific contexts:

Other limitations:

Plan to communicate these to future system users and / or affected communities:

6 Stage 6: Deployment & Post-Deployment Monitoring

Reflection Questions for Team Discussion

Use these questions for planning meetings, design sprints, and critical thinking sessions

Final Pre-Launch Checks:

- Have we conducted a last review of potential harms before deployment including technical and non-technical assessments?
- Have all high-risk issues been addressed or clearly communicated?
- Who officially signs off, and is this decision informed by diverse reviews?

User Communication & Training:

- Are we informing users that an AI system is in use?
- Are we providing adequate training (e.g. manual) on use and limitations?
- Is there an easy way for system users to ask questions or report issues?

Ongoing Monitoring:

- What's our plan to monitor real-world performance?
- Who will track impacts over time and what will trigger a re-evaluation of the system's value?

Feedback & Recourse:

- How can individuals appeal or correct or interrogate AI decisions?
- How will we continue engaging with affected communitiesy to get their feedback?

Accountability:

- Do we have clear accountability if something goes wrong?
- What is the timeline for this?

Formal Assessment: Deployment Readiness

Pre-Deployment Final Review

Human Rights Impact Assessment Final Check:
Have you completed a full HRIA now that the system is finalized?:

☐ Yes, date completed: __/__/__

☐ No, planned for: __/__/__

Important: for systems in high-risk domains (following the EU AI Act), we recommend to conduct a more in-depth HRIA such as the HUDEIRA.

Outstanding High-Risk Issues:
Any unresolved high-risk human rights issues?

☐ No outstanding issues

☐ Issues identified but acceptable risk or appropriate mitigation strategies are implemented

☐ Issues require resolution before deployment. List any remaining issues and justification for proceeding:

_____

_____

_____

_____

_____

Deployment Approval

Final Sign-off Authority:

Primary decision-maker: _____

Ethics committee/board approval:	☐ Yes	☐ No	☐ N/A
Input from representatives of the most affected communities:	☐ Yes	☐ No	☐ N/A
Input from representatives/ experts in the issues of the most at-risk communities:	☐ Yes	☐ No	☐ N/A
Regulatory approval (if required):	☐ Yes	☐ No	☐ N/A

Deployment Conditions:

- ☐ All high-risk issues adequately addressed
- ☐ Mitigation measures in place and tested (technical and through input from affected communities)
- ☐ Monitoring systems operational
- ☐ Community engagement commitments fulfilled
- ☐ Staff training completed
- ☐ Legal/regulatory requirements met
- ☐ Avenues for future community feedback/alerts/complaints in place

User Communication & Training

Transparency Measures:

- ☐ Users informed AI system is in use
- ☐ Clear explanation of system purpose and limitations
- ☐ Terms of service and privacy policies accessible
- ☐ Information available in appropriate languages
- ☐ Contact information for questions/issues provided

Training Provided:

Target audience:

Training content:

Training method: ☐ In-person ☐ Online ☐ Documentation ☐ Other: _____

Training completion rate:

6 Stage 6: Deployment & Post-Deployment Monitoring (continued)

Formal Assessment: Monitoring & Accountability Framework

Performance Monitoring

Quantitative Indicators:

- ☐ Error rates by demographic group
- ☐ System usage and adoption rates
- ☐ Complaint volumes and types
- ☐ Response times and availability
- ☐ Other: _____

Qualitative Indicators:

- ☐ Satisfaction surveys with affected groups
- ☐ Community feedback themes
- ☐ Expert assessment findings
- ☐ Media coverage and public perception
- ☐ Other: _____

Monitoring Schedule:

- ☐ Continuous automated monitoring
- ☐ Weekly
- ☐ Monthly
- ☐ Quarterly
- ☐ Annually
- ☐ After incidents
- ☐ Other: _____

Review & Update Schedule

Next system performance review:

Next HRIA update:

Annual impact report due:

Accountability Structure

Primary Responsible Person:	**Oversight Body/ Committee:**
External Auditor/ Reviewer:	**Grievance & Remedy Mechanisms:**

Grievance & Remedy Mechanisms

How to Report Issues:

Contact method:

Languages available:

Response timeline commitment:

Review process

1. _____
2. _____
3. _____

Remedy Options

- ☐ System correction
- ☐ Decision reversal
- ☐ Human review
- ☐ Compensation
- ☐ Apology
- ☐ Process improvement
- ☐ Other: _____

Appeal Process:

Steps for appealing AI decisions

Feedback Integration:

How will ongoing feedback inform system updates?

Context Change Monitoring:

How will we detect changes in social/political context that might affect fairness?

System Retirement Criteria:

Under what conditions would we take the system offline or retire it?

Overall Assessment Summary

Human Rights Impact Analysis

Rights Potentially Affected

Check all that apply and rate impact level based on the highest possible impact for any of the groups affected by the system, including affected non-users.

Civil & Political Rights:	**Impact level:**
☐ Non-discrimination and Equality	☐ High ☐ Medium ☐ Low
☐ Privacy	☐ High ☐ Medium ☐ Low
☐ Information/Transparency	☐ High ☐ Medium ☐ Low
☐ Freedom of Expression	☐ High ☐ Medium ☐ Low
☐ Participation	☐ High ☐ Medium ☐ Low
Economic, Social & Cultural Rights	
☐ Education	☐ High ☐ Medium ☐ Low
☐ Health	☐ High ☐ Medium ☐ Low
☐ Decent Work	☐ High ☐ Medium ☐ Low
☐ Freedom of Expression	☐ High ☐ Medium ☐ Low
☐ Adequate Standard of Living	☐ High ☐ Medium ☐ Low

Overall Risk Assessment

Highest Priority Risks:

1 **Risk:** _____

Affected groups: _____

Likelihood: ☐ High ☐ Medium ☐ Low

Severity: ☐ High ☐ Medium ☐ Low

Mitigation status: ☐ Resolved ☐ Mitigated ☐ Ongoing ☐ Unaddressed

2 **Risk:** _____

Affected groups: _____

Likelihood: ☐ High ☐ Medium ☐ Low

Severity: ☐ High ☐ Medium ☐ Low

Mitigation status: ☐ Resolved ☐ Mitigated ☐ Ongoing ☐ Unaddressed

3 **Risk:** _____

Affected groups: _____

Likelihood: ☐ High ☐ Medium ☐ Low

Severity: ☐ High ☐ Medium ☐ Low

Mitigation status: ☐ Resolved ☐ Mitigated ☐ Ongoing ☐ Unaddressed

Overall Assessment Outcome

Human Rights Compliance Level

☐ **Promotes Human Rights:** System actively advances human dignity and rights.

☐ **Compliant:** Meets human rights standards with adequate protections.

☐ **Conditionally Acceptable:** Significant human rights concerns require major changes.

☐ **Unacceptable:** Deployment would likely cause substantial harm.

Recommendation:

☐ **Deploy as planned** - All requirements met

☐ **Deploy with conditions** -
List conditions: _____

☐ **Delay deployment** -
Address issues first: _____

☐ **Significant redesign required** - Major changes needed

☐ **Do not deploy** - Risks too high or benefits insufficient

Final Documentation and Sign-off

Assessment Completion

Lead Assessor: _____ **Date:** _____ **Signature:** _____

Team Review

Technical Lead: _____ **Date:** _____

Community Representative: _____ **Date:** _____

Ethics/Rights Expert: _____ **Date:** _____

Final Approval

Authorizing Official: _____ **Date:** _____

Title/Role: _____ **Date:** _____

Signature: _____

Conditions for Approval:

Next Review Date:

Appendices & Supporting Documents

Attached Documentation:

- ☐ Technical specifications
- ☐ Community consultation reports
- ☐ Expert reviews
- ☐ Legal analysis
- ☐ Risk mitigation plans
- ☐ Monitoring protocols
- ☐ Training materials
- ☐ Other: _____

Key References:

- Universal Declaration of Human Rights
- Relevant national/regional human rights legislation
- AI ethics guidelines and regulations (EU AI Act, etc.)
- Organizational policies and standards
- Academic research and best practices

Living Document Notes:

This assessment is a living document that should be updated as the system evolves, context changes, or new information becomes available. Regular reviews ensure continued alignment with human rights principles and community needs.

This integrated tool combines formal Human Rights Impact Assessment with ongoing reflection throughout the AI lifecycle. It is designed to be both a practical working document for development teams and a comprehensive assessment for accountability and compliance purposes.

Template Version: 2.0 - Integrated
Last Updated: [Date]
Next Review Due: [Date]

AI & EQUALITY

Join our ‹AI & Equality› Community

for **reading groups, panels, community publications and collaborative policy comments**

aiequalitytoolbox.com