Integrating Human Rights Considerations Along the AI Lifecycle

A Framework to AI Development

Authored by Emma Kallina, Sofia Kypraiou, & Caitlin Kraft-Buchman
From the <AI & Equality> Human Rights Toolbox, December 2024

Abstract

Current research highlights the potential for AI systems to adversely affect individual and collective Human Rights if developed without careful consideration. By incorporating critical analysis and reflection points regarding Human Rights impacts during AI development, such harms can be mitigated or prevented completely. This white paper outlines our <AI & Equality> framework that enables such an approach, going even further and promoting AI development that is driven by the wish to promote human dignity. The framework consists of the essential questions and reflection points that are relevant at each of the six stages of the AI lifecycle, ensuring that Human Rights impacts are considered as they become relevant (vs after the system is already completed). Integrating the Human Rights Impact Assessment of the Alan Turing Institute with our practical Human Rights-based approach to the AI LifeCycle and AI Development, this methodology facilitates compliance to upcoming policy requirements such as the Human Rights Impact Assessment of the EU AI Act.

However, our goal is to move beyond mere compliance and **towards a paradigm of AI development that proactively promotes the achievement of Human Rights** – vs mitigating risks as an add-on or after harms have already occurred. By **involving affected communities** from the outset and with substantial decision agency, we promote and enable the development of systems that center Human Rights, equality, and inclusion at the core of code, capable of creating new opportunities and innovative correction of inequities. We hope to bring social programs in line with 21st century research and values, united in finding ways to make AI more effective – not merely more 'accurate' and 'efficient'.

What is the Purpose of a Human Rights-based approach?

AI is affecting all parts of society and even when well-intentioned has repeatedly harmed or exploited communities, and especially vulnerable groups[1]. We believe that many of these harms can be prevented through **critical reflection points** from the conceptual phase, throughout, and post AI development. These reflection points promote a **paradigm shift** in AI creation away from primarily stand alone technology-driven objectives towards a socio-technical system creation in **collaboration with the communities** that the system will interact with and affect.

This approach is likely to result in systems that are more robust, resulting in more effective uptake, use and evolution of the technology with the potential to **empower** communities and citizens in achieving and enjoying their Human Rights. It will also result in systems and solutions that bear less risk of negatively impacting the Human Rights of communities the technology is designed to serve.

Why a Human-Rights based approach vs "Ethical" or Responsible AI?

Ethics, which are crucially important, are also **situational**[2]. Ethical and Responsible AI principles, authored by a wide range of bodies (e.g. academia, civil society organizations, research institutes, governments, and the private sector) are the most common response to concerns around the ethics of AI[3], however, they are under major critique from academia[4][5] and AI practice[6][7]. Their **abstract** nature allows for diverging interpretations and implementation, impeding or even undermining accountability.

We avoid this ambiguity by focusing on Human Rights, an agreed body of international (and national) law that reflects a **universal understanding** of aspects required to ensure human dignity with a focus on equality and non-discrimination, participation and inclusion, accountability and the rule of law which are indivisible and interdependent principles of human rights[8]. Thus, Human Rights provide a **common and concrete starting point** to align different actors, disciplines, and cultures.

Further, new policies such as the EU AI Act require **Human Rights Impact Assessments** (HRIA) by the deployers or procurers of high-risk technologies such as AI used in human resources, education, financial decisions, or healthcare[9]. Since currently, no official HRIA is available as part of the EU AI Act or elsewhere, various bodies and research institutes are developing their versions of HRIAs. After reviewing several, we decided to **integrate the very thorough HRIA of the Alan Turing Institute[10] in our framework**, i.e. prompt the questions and reflections covered by the HRIA at the lifecycle stages

[1] AI Incident Database, accessed on 07.12.24 at https://incidentdatabase.ai/
[2] Sadek, M., Kallina, E., Bohné, T. et al. Challenges of responsible AI in practice: scoping review and recommended actions. AI & Society (2024). https://doi.org/10.1007/s00146-024-01880-9
[3] Jobin, A., Ienca, M. and Vayena, E. (2019) The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2
[4] McNamara, A., SmithJ., and Murphy-Hill, E (2018). Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?, https://doi.org/10.1145/3236024.3264833
[5] Munn, L. The uselessness of AI ethics. AI Ethics 3, 869–877 (2023). https://doi.org/10.1007/s43681-022-00209-w
[6] Ibáñez, J., Olmeda, Mónica (2022). Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. AI and Society 37 (4):1663-1687
[7] Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices." Proceedings of the ACM on Human-Computer Interaction 5 (April 13, 2021): 1–23. https://doi.org/10.1145/3449081
[8] United Nations. 1948. Universal Declaration of Human Rights.
[9] The EU Artificial Intelligence Act, Article 27: Fundamental Rights Impact Assessment for High-Risk AI Systems (2024), accessed on 19.12.2024 at https://artificialintelligenceact.eu/article/27/
[10] The Alan Turing Institute, Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal (2022). P.251-276, https://doi.org/10.5281/zenodo.5981675

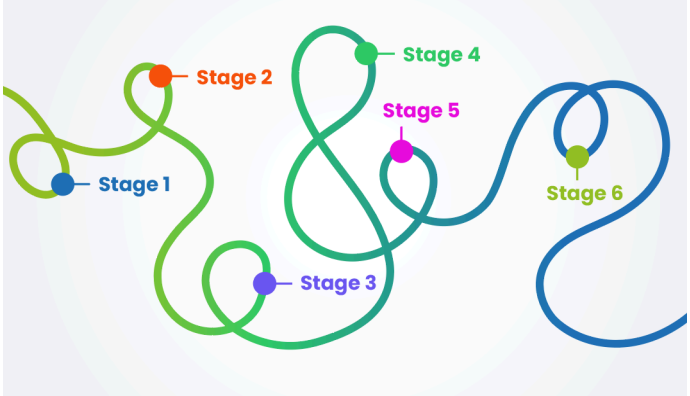
at which they become relevant. Thus, we enable an approach to AI development that considers relevant aspects **throughout the development process** – instead of as an add-on after the system has been developed, i.e. at the point of procurement. In this manner, deployers or procurers can review all actions taken, vastly facilitating accountability, transparency, as well as the process of conducting HRIAs before deployment. Consequently, orienting our framework along Human Rights has the further benefit that it **facilitates the compliance with upcoming AI regulation.**

The AI Lifecycle

To ensure that our recommendations are **actionable for AI practitioners**, we anchored our **<AI & Equality> reflective questions** along the AI lifecycle, combining them with the HRIA of the Alan Turing Institute.[11] The lifecycle is not strictly linear but **interwoven** and **cyclical**, resembling a thread looping back repeatedly. This emphasizes the importance of **reflecting, revisiting, and refining** as we learn more about the socio-technical context, the data, the model, and integration of Human Rights-based considerations **throughout the AI lifecycle** – instead of as an add-on after the system has been developed or even contemplated or slated for use.

We distinguish following **six stages** of the lifecycle:

1. Objective + Team Composition
2. Defining System Requirements
3. Data Discovery
4. Selecting and Developing a Model
5. Testing and Interpreting Outcome
6. Deployment & Post-Deployment Monitoring



[11] The Alan Turing Institute, Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal (2022). P.251-276, https://doi.org/10.5281/zenodo.5981675

Essential Questions per AI Lifecycle Stages

In the following sections, we will provide a short overview over the six stages, crucial concepts, and the essential questions that AI creators should reflect on at each specific stage (purple table). We urge you to **additionally complete our free online course**, with special emphasis on module 2 and 3, to get a more comprehensive understanding of why we recommend these reflection points in addition to purely technical measures. Both modules elaborate on the actions and thought patterns that contribute to some currently harmful practice.

How to address the reflective questions?

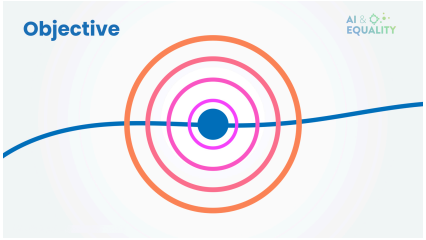
It is essential that you do not answer the questions only by yourself or with your team. Instead, for many questions it is essential to **discuss the questions and potential answers with representatives from the specifically affected communities and especially with historically marginalized groups**. Further, your answers may change as you learn new things, so **do not hesitate to revisit and amend** your answers.

The Alan Turing Institute's HRIA

The Alan Turing Institute published a working version of their Human rights, democracy, and the rule of law assurance framework for AI systems. We **locate the areas covered in their HRIA template (see p. 251 to 276[12]) along the AI lifecycle** to enable AI development that considers these **prior** to deployment, **and also** at the stage of the AI lifecycle at **which they become relevant**. In this manner, we help to build systems with Human Rights at their core, **not only implying HRIA compliance but making the process of conducting pre-deployment HRIAs easier, more efficient and effective**.

Stage 1: Defining Objective & Team Composition

A. Defining Objective



It is **essential to start with the objective and purpose of a system:** It should always be clear why a specific system is required, which issue it solves, and for whom. Too often, this vision only reflects the needs of the people developing the system in isolation holding great power in this context – as opposed to the needs of the communities the system is designed to serve and affect.

Therefore, it is essential to engage affected communities early on through participatory development practices (see box). To begin, the affected community should be consulted and agree that **an AI system is the best way to help solve their problem** as there may be simpler, more efficient and cost effective ways to tackle the core problem.

[12] The Alan Turing Institute, Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal (2022). P.251-276, https://doi.org/10.5281/zenodo.5981675

Participatory Development in this context describes the process of creating technology in collaboration with affected communities[13]. This includes an exploration of their needs, values, and concerns in the application context and addressing these in the system's design.

Affected communities can be system customers (e.g. hospital, bank, government), system users (e.g. radiologists, employee of a bank, civil servant), the people the system is used on (e.g. patient, someone applying for a loan, citizen), as well as the most vulnerable communities.

Here, it is essential that **all affected communities** (vs only revenue-critical groups) are involved and have **actual decision power and agency in the process**. This prevents an extractive form of participatory development where community needs are collected but their implementation is disregarded by commercial interests or internal agendas.

B. Team Composition

Numerous people are involved in the creation and operation of an AI system - more than just people writing code! The **objective** of a system **should fundamentally inform** the composition of its team of creators, in other words, what types of expertise and lived experience are required to fully make the intended objective a reality. This would include not only the required knowledge and technical skills, but the **diverse backgrounds, perspectives, and experiences** with the environment for which your system is developed. We want to highlight two roles that are often forgotten: affected communities & social scientists.

Impacted / Affected Communities

Affected communities are the **experts in the context where the system will be deployed** (i.e. in their *lived experience*) and will carry the consequences of the system's deployment. Special attention should be given to already marginalized communities since AI systems may have particularly adverse effects on these communities' ability to participate fully and meaningfully in the new systems that are created[14]. Input from affected communities helps to create better suited systems[15], ensures more uptake, and helps in foreseeing risks and harms.

Social Scientists

Your team should include members that are experts in the social or human rights-aspects of your application context. This is **required to understand the social contexts as well as power imbalances and inequalities** that might disadvantage historically marginalized communities, especially women and girls. Having an expert in social dynamics in your team will help the entire team, flag potential issues, and emphasize a core commitment to a collaborative team effort as the entire group to promote and protect human rights.

---

[13] Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023, October). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*

[14] Buolamwini & Gebru (2018) http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf; Angwin et al. (2016) 'Machine Bias'. ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[15] Lee, Min Kyung, et al. "WeBuildAI: Participatory framework for algorithmic governance." *Proceedings of the ACM on human-computer interaction* 3.CSCW (2019): 1-35.; Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023. Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. ACM Trans. Comput.-Hum. Interact. 30, 2, Article 33 (April 2023), 39 pages. https://doi.org/10.1145/3582430

Essential Questions: Defining Objective & Team Composition

Purpose & Context of the System	• What **problem** is the system trying to solve?
	  ○ Does the domain have a **history of discrimination?**
	  ○ Is there a risk that your system might **enhance or enforce** historically unequal outcomes?
	  ○ How can you **counteract** such historical discrimination?
	• Will the system have an **essential or high-risk function** or be implemented in a high impact or safety critical sector (see e.g. EU AI Act)?
	  ○ How do you ensure safe operation, both in design as well as in case of system outage?
	• Have **communities** affected by the system been engaged in dialogue about the system?
	  ○ Is an AI system even the best way to address the issue?
	  ○ Does it address the community's most pressing needs?
	  ○ Are some of the communities vulnerable, e.g. due to protected characteristics?
	• Is the system supposed to be implemented at **scale**? Is this wise?
	• Is using the system or the system being used on someone **voluntary** (direct & indirect use)?

Effects of the System	• Who **benefits** from the system and who can be **disadvantaged**?
	  ○ Does this reflect or level **current power structures**?
	  ○ How can we involve communities and especially historically marginalized groups?
	• Does the system actively **contribute** to **Human Rights**?
	  ○ Have you conducted a first screening of Human Rights Impacts to identify risks before resources have been invested (p.21 to 47 in[16])? Potential risks include manipulation, discrimination, or guarding current power structures.
	  ■ What if the system is used in unintended ways?
	  ○ Does the system help to promote Human Rights principles and priorities?
	  ○ Who should be included in / consulted during this assessment?
	  ○ How do you ensure that identified risks are eliminated or mitigated?
	• Who is **accountable** for inaccuracies and resulting harm?
	  ○ How do you document system design decisions, accountabilities, and general responsibilities so they can be traced back?
	  ○ Have you considered all above questions (especially Human Rights impacts) for your **system's entire value chain**, e.g. for suppliers, subcontractors, auditors, ...?
	  ■ How do you ensure the ongoing and thorough scrutiny of the value chain?

Empowering Affected	• How can the **impacted communities be represented in the team** so that the team can benefit from their insights and real world experience?

[16] The Alan Turing Institute. *Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal* (2022), https://doi.org/10.5281/zenodo.5981675

Communities	• Besides via team membership, how does the team involve affected communities?
	  ○ Do these communities receive the necessary **agency** to impact decisions?
	  ○ Does the development team have the **mindset** and **skills** to achieve this?

Team Composition	• What **expertise** do you need in your team?
	• Do you have diversity in **culture**, **demographics, lived experience, disciplines** & **skills** (socio-technical, legal, anthropological, UX, technical,...)?
	• How do you ensure flat **hierarchies** & **communication** between the disciplines?
	• Does the team have:
	  ○ Awareness of the **risks that AI system**s pose to Human Rights and underlying reasons?
	  ○ Insights into / experiences with the problem they are trying to solve?
	  ○ Insights into / experiences with potential solutions for this problem?

Stage 2: Defining System Requirements

At the second stage, the system's objective is formalized into a list of requirements, again, **developed in dialogue** between various roles and communities. This includes managing trade-offs between different needs and desired requirements as systems exist in an **ecosystem of values**.

Ecosystem of Values.

Different aspects of a system make it responsible. Examples are that its decisions are fair (fairness), that its decisions are easy to understand (explainability), that its development process and underlying motivations are clear (transparency), or it operates with little error (accuracy). You can find a list of these aspects with more detailed definitions and examples in Module 2 of our online course. it is **impossible to optimize all of these aspects simultaneously** in equal measure, therefore **trade-offs** are required[17] (Although these trade offs do not necessarily reduce accuracy in any fundamental way[18]). For example, highly explainable models often have less accuracy than more opaque forms of AI models[19].

In some contexts, explainability might be as important (or even more important) than the minimization of errors (accuracy): only if the human overseeing the system can understand and question the output, she can detect and correct the errors - thus ultimately leading to less errors than high accuracy alone. Thus, it is essential to not focus solely on one metric (such as often done with accuracy), but instead to **make a conscious decision** about metric hierarchy and importance in the specific context.

[17] Whittlestone, Jess, et al. "Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research." London: Nuffield Foundation (2019).https://www.nuffieldfoundation.org/wp-content/uploads/2019/02/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf
[18] Rodolfa, K.T., Lamba, H. & Ghani, R. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nat Mach Intell* 3, 896–904 (2021). https://doi.org/10.1038/s42256-021-00396-x
[19] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://www.mdpi.com/1099-4300/23/1/18

Importantly, **accuracy should never be considered without fairness** as it can hide unequally distributed accuracy, e.g. that the system is highly accurate for the majority of cases while being very inaccurate for a minority group[20]. This can lead to negative Human Rights impacts, in healthcare, facial recognition, finance, subsidy, and other important sectors.

The process of defining the system's requirements should be iterative and fluid; it is very likely that the **list of requirements may change as more details about the social context and the needs of impacted communities become apparent.** Thus, it is important to **provide a platform** where operators and affected communities can notify the team of new pieces of information that might influence the requirements.

Essential Questions: System Requirements

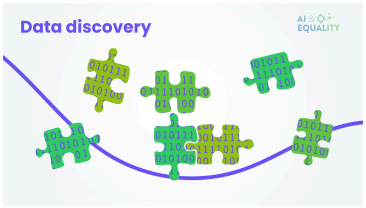
Involving Affected Communities	• **Who** should be involved in the definition of the system requirements? Think beyond operators, users or revenue-critical parties! • Are there **tensions** between the system's goals and the needs of affected communities? How can these be addressed, always prioritising Human Rights? • Have you **revisited** your initial Human Rights Impact Assessment, now where more capabilities are planned? • Have you arranged **expert input**, e.g. from affected communities with lived experience, a government department (or allied government department), academia, or public body?

Explainability Considerations	• **What is the goal** of explanations?
	  ○ Who is the **audience** and why?
	  ○ Will explanations be available for all affected communities, aiding public scrutiny?
	  ○ Are provided explanations **easy to process** for all intended audiences?
	• Have you considered **which aspects** of explainability are the most relevant?
	  ○ E.g. how decisions are made in general, how an individual decision was made, ...
	• How can you use explanations to **increase the agency** of affected communities, e.g. via detailing what would have to change for a different outcome (counterfactual explanation)?
	  ○ How do you ensure that your explanations help affected communities to understand the limits and impacts of the system?

[20] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. PMLR, 2018.

Ecosystem of Values	• Are there **tensions between accuracy and other, more necessary metrics** in this context?
	• **Fairness:** Which fairness metrics do you expect to be useful in this context? Explore several!
	• **Privacy:** Is the privacy of all affected communities and data subjects respected?
	  ○ How can you minimize the data collection in private spheres, e.g. homes?
	  ○ Is the remaining intrusion worth it?
	• **Transparency:** How will you enable impacted communities to access information about your methodology, e.g. training data, analytical process, how the model was trained, metadata of various metrics?
	  ○ How can you ensure that affected communities are aware that they are using an AI system /or it is used on them?
	• **Accountability:** What is the accountability structure?
	  ○ Which human oversight should be aimed for?
	  ○ What expertise and training will the human in the loop require?
	  ○ How can you enable affected communities to contest an outcome?
	• **Usability:** How can we ensure that the interface is intuitive and accessible for all?

Stage 3: Data Discovery



A valuable system objective and its requirements can be undermined if the dataset used to train the AI system is not **representative** of your use case and context. A **good socio-cultural fit** of the dataset includes various aspects such as the demographics of the individuals in the dataset, their culture, or environmental factors[21]. Consulting **domain experts** will be imperative to ensure relevant aspects are appropriately captured.

If no dataset with a good fit is available, the team may have to generate a **new dataset**, either by collecting new data, and/or by improving or augmenting existing datasets through **pre-processing** (i.e. mathematical) steps.

Pre-Processing refers to **the manipulation and transformation of raw data before feeding it into a model.** It involves various techniques to enhance the quality, relevance, and fairness of the data, e.g. by balancing the frequency of a specific class (e.g. gender or race) in the dataset so that the model is equally trained on them.

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. ACM 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723

Essential Questions: Data Discovery

Data Origin	• **Who** collected the data and for which **purpose**?
	• Did the data subjects **consent** to use of their data?
	  ○ Was their **privacy** respected?
	• How **sensitive** is the information, e.g. does the data reveal sensitive attributes such as racial or ethnic origins, sexual orientations, health status, or religious beliefs?
	  ○ Is there a way to **anonymize** the personal data so that privacy is respected AND insights on age, gender, geography can be captured ?

Data Bias	• Who is **included** in the data? Who is **excluded**? Why might that be?
	  ○ Which **geographic regions** and **cultures** are included and which not?
	  ○ Which **consequences** does this have for your system's operation?
	• Which historical / present **bias** might be in the data, risking to compromise Human Rights?
	• Which data **pre-processing** steps are required to create a model that is **fair** in this context?
	• In your specific use case, is it most beneficial to ignore (show potential unfairness in data), 'erase' (remove potential unfairness in data), or even counteract (counteract this bias in a way that the disadvantaged group is now advantaged) in this bias?

Documentation	• Have you **documented** which datasets you are using and why you choose them so that potential deployers can assess whether your training data fits their context? • Have you documented all **pre-processing** steps you took (essential information for future uses of your system or code)? • Have you saved your **"raw" data** – in addition to the preprocessed data – to support future uses?

Stage 4: Selecting and Developing a Model

It is time to consider **what type of AI model is the best** to satisfy the system requirements. Note: it is not always the most complicated deep-learning algorithm!

Instead, it is about choosing the most suitable model for the required scope while managing **trade-offs**. For example, less complex models are often more explainable but might achieve a slightly lower accuracy. Since explainability is a prerequisite for good error and bias detection, such models seem especially important in **high-stakes scenarios**. For example, the European Central Bank requires a high level of explainability for credit scoring decisions[22], and therefore excludes neural

[22] Dessain, J., Bentaleb, N., and Vinas, F (2023). *Cost of Explainability in AI: An Example with Credit Scoring Models.* https://doi.org/10.1007/978-3-031-44064-9_26

networks and other types of less explainable algorithms that **impede the discovery of discriminatory outcomes** and scrutiny.

Model development itself is an **iterative process** in which different aspects of the model are adjusted to **meet different system requirements** (e.g. via in- or post-processing methods or by adjusting the weights or parameters of a model). It is important here to **reflect about earlier stages** to ensure that your objective, requirements, data, and model are all aligned.

In-Processing methods are designed to mitigate bias / increase fairness while the model is being trained while **Post-Processing** methods include modifying the model's output after training has been completed.

Essential Questions: Selecting and Developing a Model

Model Type and Explainability Requirements	• Does your model…
	  ○ … Achieve appropriate **explainability**, considering the stakes of the situation?
	  ○ …. Minimise **complexity**?
	  ○ … **Alert** the user if it is **uncertain** with a decision and / or when it is confronted with an instance that is not reflected sufficiently in its training data (e.g. model only trained on light skin with little pigment is presented with an instance of dark skin with more pigment, thus alerting the user that it does not know how to classify this instance)?

Fairness Aspects (see module 3 of our free online course)	• What is the most suitable **fairness metric** and why? Have you experimented with a variety of different metrics and outcomes? • Which **aspects** of fairness are in focus, e.g. based on gender, ethnicity, education…? ○ Have you considered relevant **intersectionalities**? • Have you ensured that the model does not rely on variables or **proxies** that might be unfairly discriminatory? For example, a person's postcode might allow you to infer ethnicity. • Why have certain **in- (model) and post (evaluation)-processing steps** been chosen?
Other	• Is the model **transparent** to affected communities, i.e. who funded it, its objective, who was involved, training data, performance, … • What is the **environmental impact** of the model? Is it worth the cost? ○ Have there been efforts to **minimize** or offset environmental impact?

Stage 5: Test and Interpret Outcome

After the model has been developed, we have to test whether it **fulfills the system requirements** defined by the team in stage 2. For some metrics, this can be done via **technical tests**, others require the **feedback of affected communities**[23], e.g. whether the intended level of explainability was achieved.

For the technical tests, it is important that the **testing dataset is as representative** of the context as the training dataset. Including **extreme examples / cases** can help to uncover potential issues that may not be apparent during routine testing, thereby revealing any limitations or weaknesses in the model's performance[24].

Insights gained should inform a **'manual' handed to the future system users / operators.** Through stating the contexts for which the system has been trained (expected to operate well) and which are not (inaccuracies likely), the operators can **calibrate their trust and adherence accordingly**. Further, the manual should include recommendations on the required level of **human oversight**, thus allowing appropriate training of the operators.

Essential Questions: Test and Interpret Outcome

Testing Context and Outcomes	• Does the system **meet the objective and the system requirements**?
	  ○ What **measures** of model performance are included and why were they selected over others (including quantitative AND qualitative aspects)?
	  ○ Does this selection **still apply** after we learned more about the application context? Should we **add** something?
	  ○ **Whose opinion** was included in these tests?
	• Can the trained model be released to the public or external experts to allow them to **test and scrutinize** it to highlight issues?
	• Has the model been tested as **close to its actual application context as possible** (including its actual users) to identify potential harms?
	  ○ Have resulting learnings and feedback points been **included**?

Operation Manual	• Is an **easily understandable** manual available to the operators? • What can we recommend as best practices around operation, e.g. how much **human oversight** is required and with which expertise ? • For **which contexts** has the system been trained? ○ Where might it become unfair or inaccurate? • How will you **train** operators on how to use and interpret the system, including how to calibrate their trust in and ability to question the system's operation? • How will you **log future changes** to the system?

[23] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

[24] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B.,Spitzer, E., Deborah Raji, I. and Gebru, T. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, January 2019. ACM. http://dx.doi.org/10.1145/3287560.3287596

Stage 6: Deployment & Post-Deployment, Auditing and Monitoring

Deployment: The deployment step is the **last sanity check**, i.e. whether all harms, discriminatory impacts and consequences have been considered, communicated, and are accounted for. Revisit your initial **Human Rights Impact Assessment** and conduct it more thoroughly now that you know the full system to ensure that the system has been assessed for negative Human Rights impacts in its final form.

The decision as to whether the system is ready to be deployed is powerful. We recommend truly empowering affected communities - after all, **they have to bear the consequences of a faulty operation**! Additionally, it is crucial to set up pathways that enable **operators and strongly affected communities to alert issues** they experience around the system.

Post-Deployment: The system should be **audited and tested regularly in post-deployment audits**, including opportunities for affected communities to provide feedback. This is **especially relevant shortly after deployment** as the newly deployed system might expose previously unknown challenges or problems.

Even if the system operates as expected, the model's application context is **likely to change over time**. This can not only alter the input data or which outputs are considered fair, but even impact the objective, e.g. make the objective obsolete so that the system should be retired. Therefore, **it is essential to continuously audit the system**, including both quantitative audits as well as qualitative audits in collaboration with affected communities (see e.g.[25] for a framework to operationalise such audits). A thorough overview over different types of audits - also including audits by external parties - can be viewed here[26]

Essential Questions: Deployment & Post-Deployment, Auditing and Monitoring

Deployment	• Who decides that the model is **ready to be deployed**?
	  ○ Have regulators, domain experts, affected communities **agreed** to deployment?
	  ○ Do the most affected communities have the **agency to delay / stop** deployment?
	• Have you revisited your initial **Human Rights-Impact Assessment** and conducted a **more thorough** one, now where the full model capabilities are known? (following [27])
	• Before deployment: Are there processes in place to **detect potential system failures or unexpected harms**?

[25] Yurrita, M., Murray-Rust, D., Balayn, A., & Bozzon, A. (2022, June). Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 535-563). https://dl.acm.org/doi/abs/10.1145/3531146.3533118
[26] Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024, April). AI auditing: The broken bus on the road to AI accountability. In 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) (pp. 612-643). IEEE. https://ieeexplore.ieee.org/abstract/document/10516659
[27] The Alan Turing Institute, Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal (2022). P.251-276, https://doi.org/10.5281/zenodo.5981675

	○ Are the deciders **accountable** for harm that might be caused? ○ What mechanisms are in place for after an issue has been identified? ○ Who is **responsible** for addressing upcoming harms? What is the **timeline**?
Monitoring	• Are there processes or features in place that **allow operators and impacted communities to alert** suspected system inaccuracies or failures? ○ How can you ensure that affected communities can **opt out of system use**? • How are you **monitoring context changes**? ○ What is your process to learn about **new risks or harms**? ○ What is your mechanism to learn about **new user needs** in the field? ○ How can we include them in the **requirements** and account for them? ○ In which cases is it better to take the system **offline** until risks have been accounted for? ○ How will you test that the model **continues to fulfill its objective**? ■ How would you know that it is time to retire the system?

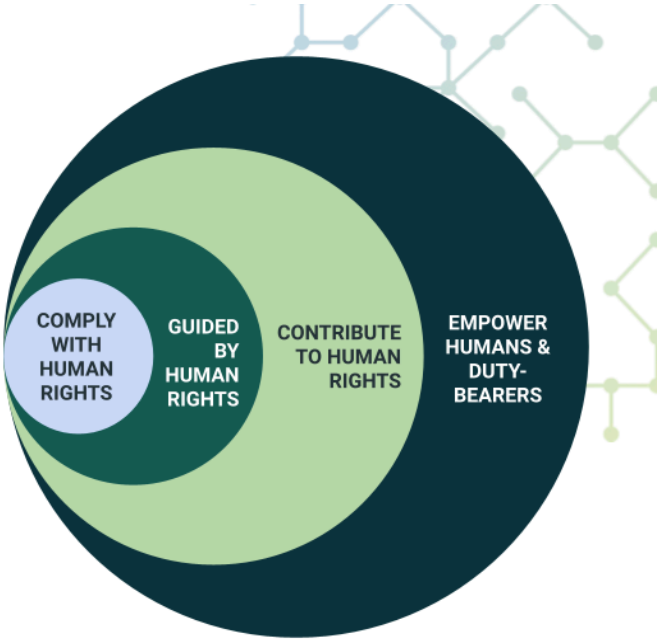
Summary

We highlighted essential questions along the six stages of the AI lifecycle to enable AI creators to **reflect about the objectives, Human Rights impacts, and wider societal effects of the systems they create in collaboration** with the communities affected by their system.

We want to emphasise that these questions - at the bare minimum – facilitate the creation of technology that **complies with the Human Rights** principles of Equality & Non-Discrimination, Participation & Inclusion, Accountability & the Rule-of-Law. However, these questions may help to go beyond mere compliance and allow the creation of technologies that are:

- **guided by** Human Rights principles,
- **contribute to** their access and fulfillment, and
- aspire to **empower humans & duty-bearers** to achieve and enjoy their Human Rights.

Going forward, this may allow us to not only 'leave no one behind', but to bring everyone with us, **enhancing human dignity as we create new technologies.**



Join our

<AI & Equality> Community

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

AI & EQUALITY

A ***fair future*** starts **here**

This brief was developed in partnership with Women At The Table for the Swiss Federal Department of Foreign Affairs Peace and Human Rights Division, and the AI4D AI for Development Program funded by The International Development Research Centre (IDRC) .

WOMEN AT THE TABLE

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Foreign Affairs FDFA

ARTIFICIAL INTELLIGENCE FOR DEVELOPMENT AFRICA

IDRC · CRDI