ARTIFICIAL INTELLIGENCE FOR DEVELOPMENT AFRICA

What is gender-responsive AI, and why does it matter?

Learning Brief No. 2 • Alex Berryhill and Lucía Mesa Vélez (Ladysmith) • December 2022

Objectives of this resource: To provide an operational definition of gender-responsive projects in the context of AI-driven research and innovation, and synthesize available research that illustrates why gender equality and inclusion (GEI) considerations matter for the design and use of algorithmic decision-making. *This synthesis of the available literature is far from comprehensive, and is meant to serve as a starting point for discussions and actions around GEI excellence for AI4D labs and hubs.*

This is the second resource in the **AI4D GEI Support Team's learning brief series**:

- **Learning Brief No. 1** summarizes priority GEI considerations for the design of AI4D calls for proposals.
- **Learning Brief No. 3** in continuation of Learning Brief No. 2 (this resource), identifies key action areas for advancing gender-responsive AI research and innovation.

What is gender-responsive AI?

Gender-responsive projects advance gender equality and social justice via their organizational practices, methodologies, outputs, and outcomes. Such projects are core to the mission of IDRC, as illustrated by **IDRC's Equality statement**: "We support the generation of knowledge — including by individuals from diverse genders, communities, histories, and experiences — that ***tackles the systems that perpetuate inequalities on the basis of identity***" (emphasis added).

Gender-responsiveness appears to be a particular challenge for AI research and innovation projects. From Google's sexist image search results, to Twitter's racist image cropping algorithms, from AI-based video games that promote child abuse and sexual exploitation, to the misdiagnosis of life-threatening conditions[1]—a growing evidence base illustrates the many ways in which AI applications have contributed to and exacerbated gender, racial, economic and global inequalities.

[1] Mandal, A. (2021). "Algorithmic Origins of Bias." Women at the Table.

There are no fully conclusive 'checklists' on how to 'fix' the AI sector's 'gender problem'—although certain considerations can help 'move the needle' towards more gender-responsive AI research and innovation, some of which are outlined in Learning Brief No. 3. For more transformative and sustainable impact, teams leveraging AI must continuously reflect upon *their* roles and responsibilities in broader organizational, cultural, and political systems change.

Key features of gender-responsive projects	
Human rights-based	Human rights are at the core of gender-responsive projects. This means, among other things, protection concerns are prioritized over data desires.[2] For example, while data on gender-based violence can contribute to informing programs and policies, some data collection practices may present risks of creating additional harm. In such cases, a data subject's rights to protection must be prioritized above any other potentially competing priorities (such as well-intended interests to monitor trends in gender-based violence, or to collect data that could then contribute to relevant algorithmic models).[3]
Participatory	Gender-responsive projects effectively *respond* to local norms and achieve positive gender equality outcomes due precisely to their deep participatory nature with impacted communities. Such projects go beyond tokenistic or top-down approaches to participation (i.e., calling certain groups 'partners' but not allowing them to influence decision-making) and instead invest in equitable partnerships (i.e., partners can influence a project's direction and are treated as equitable collaborators, rather than research 'subjects').
Intersectional	Intersectionality is a term that refers to the cumulative way in which the effects of multiple forms of oppression, discrimination and exclusion (including but not limited to racism, sexism, and classism) combine, overlap, or intersect. Achieving positive

[2] For additional resources and discussion on AI and human rights, see Women at the Table's <AI & Equality> Human Rights Toolbox here.
[3] Zulver, J.M., et al. (2021). "COVID-19 and gender-based violence: reflections from a "data for development" project on the Colombia–Venezuela border." *International Feminist Journal of Politics* 23(2): 341-349.

	gender equality outcomes requires an intersectional gender analysis of the given topic and population of interest. In other words, gender-responsive AI-driven research and innovation should begin with an understanding of *how* and *which* factors (e.g., age, race, dis/ability, income, location) intersect to accentuate oppression among individuals and different sub-populations in the given context and across different levels, including interpersonal, community and institutional systems.
Interdisciplinary	Achieving many of these features requires bringing together diverse disciplines, or professional backgrounds and experience. AI teams are often composed of professionals from the natural sciences (e.g., mathematicians, engineers, biologists). Yet, perspectives from social science (e.g., sociologists, economists, psychologists) and humanities (e.g., history, visual arts, law) are critical for leading intersectional gender analyses, and connecting such analyses to more holistic and impactful AI design.[4] Interdisciplinary perspectives are also essential for identifying and mitigating potential protection risks with both qualitative and quantitative data collection methods[5], developing evidence-based theories of change, and engaging impacted communities, among other essential components of responsible and gender-responsive AI.
Reflexive	Gender-responsive projects recognize that achieving positive gender equality outcomes requires more than short-term technological fixes—instead, we need broader systems change.[6] As such, gender-responsive projects regularly reflect on their individual and collective roles in these systems, and their responsibilities towards broader organizational, cultural, and political systems change.
Action-driven	Beginning from the inception of their work, gender-responsive

[4] See Stanford's Gendered Innovations in Science and Engineering Project's discussion of intersectional approaches (here) and their resources to support Intersectional Design methods (here).
[5] Fuentes, L. and Cookson, T.P. (2020). "Counting gender (in)equality? a feminist geographical critique of the 'gender data revolution'." *Gender, Place & Culture* 27(6): 881-902.
[6] For example, see Gender at Work's Analytical Framework (here).

	projects are action-driven, and clear about desired gender equality outcomes. Gendered intersectional relations (i.e., an understanding of how gender and other systems of oppression shape relationships between and within sub-populations) are considered and built into their concept notes, proposals, evaluation grids, and other project materials to ensure project outputs support sustainable, positive gender equality outcomes.[7] Early on, gender-responsive projects engage key stakeholders, including power-holders within the particular system of interest.

Are gender-responsive projects "just about women"?

Gender-responsive projects are clear about desired gender equality outcomes from the beginning and through to the end of their project's life cycle, and treat GEI as an integral component of *how* they approach projects, rather than as an individual activity or 'add on'. This means that gender-responsive projects should go beyond an 'add in women and stir' approach[8] and instead analyze *why* women, among other oppressed groups (see principle of intersectionality above), are systematically excluded from the design and benefits of AI applications.

Consider the following theoretical project: A team develops new AI-based prediction tools that seek to strengthen community resilience to epidemic outbreaks. This project could be considered gender-responsive if (1) advancing GEI is a clear objective, and (2) it invests in organizational practices, methodologies, and activities that align with this objective, such as:

- The team's leadership is **representative of the communities** they wish to impact, features a diversity of disciplinary perspectives, and has created partnerships with local civil society organizations and community advocates;
- They understand that public health crises have **differential impacts** on groups based on multiple and intersecting identities, and make sure the data informing their predictive models is inclusive and captures local and

[7] For more on connecting (gender) data to action, see IDRC's Global AI4COVID program's Technical Brief #4: Connecting Gender Data to Action (Ladysmith 2021).
[8] See Cornwall, A., Harrison, E., and Whitehead, A. 2007. *Feminisms in Development: Contradictions, Contestations & Challenges.* London: Zed Books.

intersectional identities (for example, including data generated from **participatory action research**[9]);

- Because the team has strong and well-integrated **monitoring and reflection practices**, they are able to efficiently identify new risks, changes in local contexts (including dynamics that influence systems of oppression or create new forms of discrimination), or other key learnings, and based on this, continuously **iterate, adapt and strengthen** the project as needed;
- Lastly, the project team wishes to go beyond 'collecting data for data's sake',[10] so they invest in **data uptake** activities and work closely with key stakeholders (including power holders) from the inception of their project, to create spaces for deliberative dialogue, leverage research for advocacy, and consequently, inform more evidence-driven and inclusive public health services, programs and policies.

Why do GEI considerations matter for AI?

Along with ensuring that activities do no harm as they engage impacted communities, effectively integrating GEI considerations can enhance the quality, innovative nature and sustainable impact of projects. Here's how:

1. **Do No Harm**: Given the unique scale, speed and impact of AI applications across many facets of our daily lives (including impacts which often go unknown[11]), AI-based technologies are unquestionably powerful tools. AI has the potential to greatly benefit society—but it also has the power to (intentionally or unintentionally) enact significant harms. For example, medical professionals increasingly rely on AI to diagnose diseases. However, the data used to inform these algorithms often includes gender or racial biases, resulting in misdiagnoses that disproportionately impact women, transgender individuals, and individuals with darker skin—thus exacerbating health inequalities and rendering these groups at greater risk of

[9] Njuki, J., et al. (2022). "Meeting the challenge of gender inequality through gender transformative research: lessons from research in Africa, Asia, and Latin America." *Canadian Journal of Development Studies.*
[10] Cookson, T.P & Fuentes, L. (2021). "Without Actionable Data, Gender Equality Will Remain Out of Reach." *Stanford Social Innovation Review.*
[11] Once an AI-based technology is open to the public, individuals can also find harmful ways of using it, even if those were not intended by its original creators. For instance, Microsoft's conversational AI Tay, launched as a Twitter bot in 2016, was quickly weaponized by alt-right users to make misogynistic and racist tweets (Vincent 2016). Similarly, OpenAI's GPT-3 language model was taught to generate answers as if it were a member of far-right group QAnon (McGuffie & Newhouse 2020)

life-threatening conditions.[12] In other words: the principle of 'do no harm' has been severely overlooked in this particular use of AI.

Gender-responsive practices help identify and more effectively (and responsibly) mitigate such harms: Teams that meaningfully engage impacted communities throughout their project's design, implementation and monitoring phases are more likely to identify and mitigate potential risks or biases.[13] This is especially true when project team's are *not* fully representative of impacted communities, and thus less likely to intuitively understand the many ways in which their proposed research or AI applications may impact diverse groups and individuals. Partnerships with women's rights organizations and other gender equality stakeholders, in particular, can help with understanding the ways in which some AI-driven research and innovation projects may reinforce discriminatory gender norms and inequalities, *or* perhaps challenge these norms—and what the consequences of this may look like for women, girls, LGBTIQ+ individuals, and other marginalized groups.

2. **Quality:** Inclusive research is higher quality research—whether quality is defined from a scientific rigor, legitimacy, relevance, or actionability perspective, as illustrated by IDRC's **Research Quality Plus (RQ+)** approach. The relationship between quality and inclusivity is especially true for AI applications, given the fundamental importance of data for shaping the performance and impact of algorithmic decision-making models. The disproportionate exclusion (or stereotyped representation) of marginalized groups (such as women, girls, and LGBTIQ+ individuals) from datasets results in statistical and social biases that harm the overall quality of the AI applications trained on these datasets.[14] Visual datasets, for example, are often biased in favor of white male faces. Therefore, when facial recognition technologies are trained on this data, their algorithms are less likely to recognize women and non-white faces. This results in lower quality or less effective technologies, while also contributing to both gendered and racial micro- and macro-aggressions.[15] More participatory and rights-based approaches (i.e., engaging *all* key stakeholders in the targeted system or context) along with more inclusive

[12] See Larrazabal, A. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis; and Adewole, S. (2018). Machine Learning and Health Care Disparities in Dermatology.
[13] As illustrated in Technical Brief #3 for IDRC's AI4COVID Program, community-based stakeholders have unique insight into which data gaps should be prioritized to maximize impact. By working with community leaders in Cameroon, for instance, the Africa-Canada Artificial Intelligence and Data Innovation Consortium's (ACADIC) identified that data about COVID-19 hotspots did not include communities with a higher percentage of internally displaced persons, which they knew was not representative of the local reality. This lead to campaigns that increased access to testing centers and therefore had an impact on policy and resource mobilization for these communities.
[14] Mehrabi, et al. (2019). "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54(6): 1-35.
[15] Dickey, M.R. (2020). "Twitter and Zoom's algorithmic bias issues".

teams is essential for reducing such risks of harm, and enhancing opportunities for more positive, successful and impactful user uptake from the given AI application.

3. **Innovation:** Diverse, inclusive and more collaborative teams are also more innovative teams. These findings are supported by a growing body of research,[16] which find that more *meaningfully diverse* teams (i.e., team composition features a diversity of identities, backgrounds, and disciplines—and this diversity is also reflected in decision-making bodies and processes) are more likely to approach design challenges from new vantage points, spark new ideas, collaborate, and thus, innovate. Therefore, for organizations and researchers seeking to use AI applications to address society's most pressing challenges, it is highly practical and valuable to include a diverse range of perspectives—particularly those that are most often marginalized from the sector and/or those most likely to be impacted by proposed AI applications.

4. **Sustainable impact**: Lastly, gender-responsive research is also more likely to have sustainable, long-term impact. This is for several reasons: Partnerships and close engagement with impacted communities and the organizations that represent them, along with the inclusion of more interdisciplinary expertise in AI teams, is key for deeply understanding local contexts, and based on this, assessing and mitigating risks. Consequently, gender-responsive research is *less* likely to result in unintended negative impacts, which otherwise could (and at times, very much should) result in an early termination of a project.[17]

   Secondly, gender-responsive research is action-driven, meaning that positive social impact is a core driving motivation for gender-responsive research. With evidence-based theories of change and deep partnerships with impacted communities, civil society organizations, and community leaders, gender-responsive projects are more likely to understand the root drivers of social challenges and the key entry points for addressing these challenges. Consequently, the core practices that define gender-responsiveness are also essential practices for developing more relevant, sustainable, and impactful AI-driven research and innovation projects.

[16] Women at the Table. (2019). "We Shape Our Tools, Thereafter Our Tools Shape Us: Artificial Intelligence, Automated Decision-Making & Gender."
[17] Such as Microsoft's Twitter bot Tay (mentioned above), which closed 24 hours after being launched due to it's rapid transformation via discriminatory tweets (Vincent 2016).

Annex A: Recommended resources

- "Algorithmic Origins of Bias" (Mandal 2021)
- Women at the Table's <AI & Equality> Human Rights Toolbox (here)
- Stanford's Gendered Innovations in Science and Engineering Project (here)
- "We Shape Our Tools, Thereafter Our Tools Shape Us: Artificial Intelligence, Automated Decision-Making & Gender" (Women at the Table 2019)
- AI4COVID Technical Brief 1: Designing gender-responsive data projects (Ladysmith 2021)
- AI4COVID Technical Brief 2: A guide for more gender-responsive health research (Ladysmith 2021)
- AI4COVID Technical Brief 3: Stakeholder engagement for gender-responsive health research (Ladysmith 2021)
- AI4COVID Technical Brief 4: Connecting Gender Data to Action (Ladysmith 2022)
- "Seven intersectional feminist principles for equitable and actionable COVID-19 data" (D'Ignazio and Klein 2020)
- "Meeting the challenge of gender inequality through gender transformative research: lessons from research in Africa, Asia, and Latin America" (Njuki et al 2022)
- Transforming gender relations: Insights from IDRC research (IDRC 2019).

For more resources, we recommend visiting this Google Drive folder organized by IDRC, along with the AI4D program website. IDRC and the AI4D gender support team will continue adding publicly available resources to both of these pages throughout the AI4D program.

Annex B: Key terms

- **Gender**: Gender refers to the socially-constructed roles, responsibilities and relationships that society considers appropriate for women and men. It also has implications therefore for individuals and groups who identify as gender

non-conforming. Gender is upheld by political, economic, social, and cultural institutions. Gender is context and time-specific, and thus changeable as well.

- **Sex:** The sum of biological and physiological characteristics that typically define men and women, such as reproductive organs, hormonal make-up, chromosomal patterns, hair-growth patterns, distribution of muscle and fat, body shape and skeletal structure.
- **Intersectionality:** The cumulative way in which the effects of multiple forms of oppression, discrimination and exclusion (including but not limited to racism, sexism, and classism) combine, overlap, or intersect.
- **Inclusion:** The aim of inclusion is to embrace all people irrespective of race, gender, disability, medical or other need. It is about giving equal access and opportunities and getting rid of discrimination and intolerance (removal of barriers). It affects all aspects of public life.