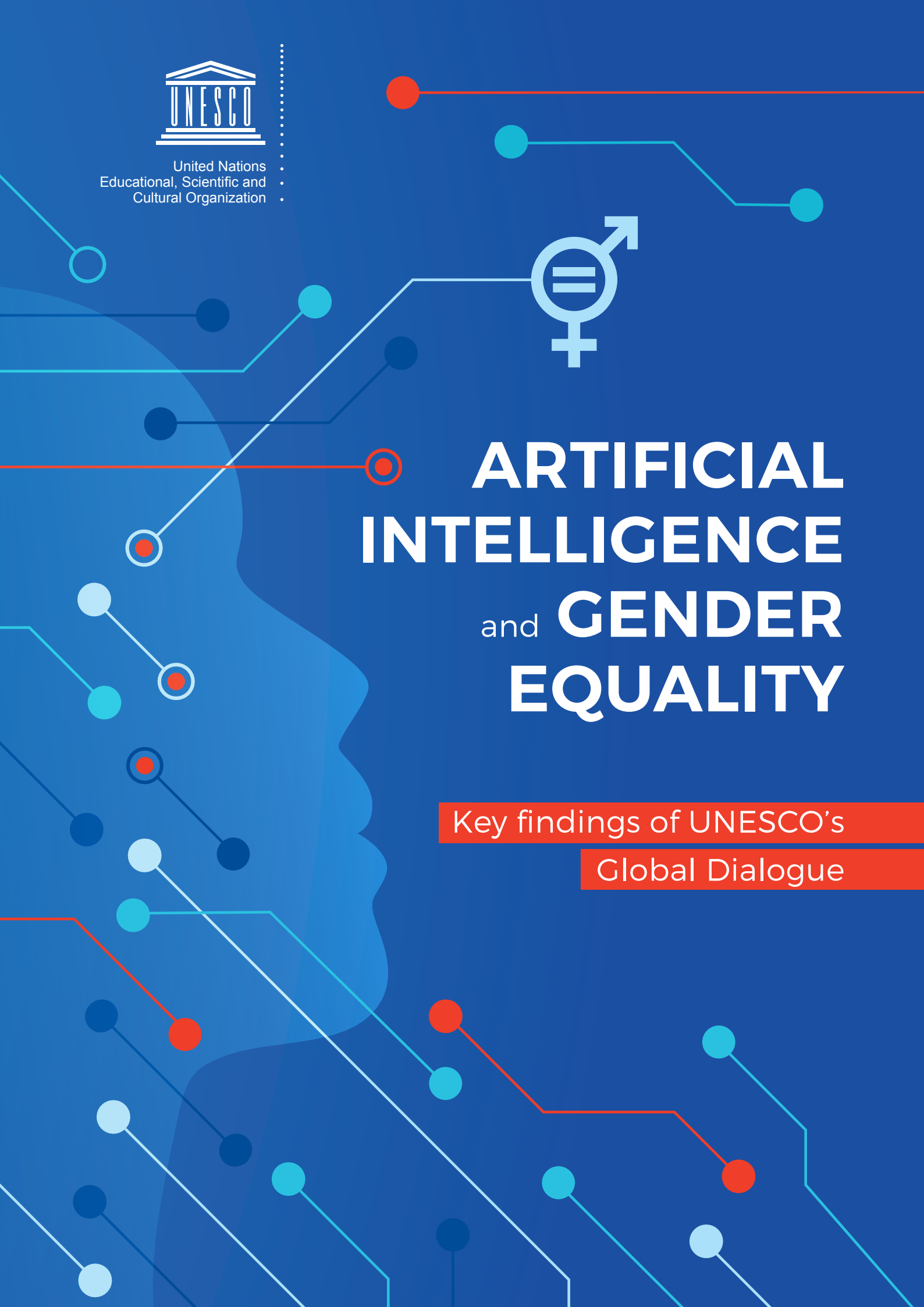
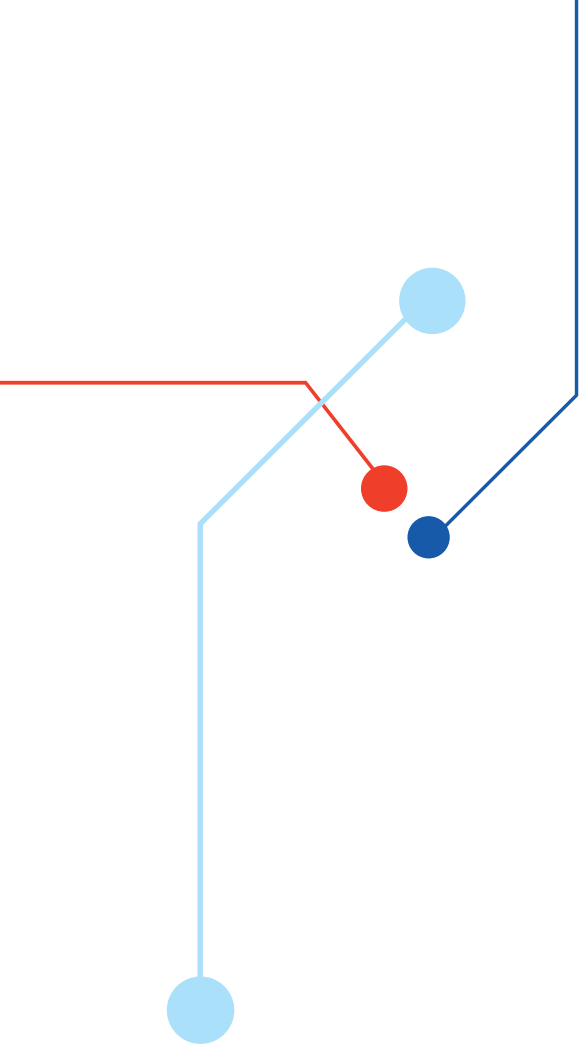
United Nations
Educational, Scientific and
Cultural Organization



# ARTIFICIAL INTELLIGENCE and GENDER EQUALITY

## Key findings of UNESCO's Global Dialogue



# ARTIFICIAL INTELLIGENCE and GENDER EQUALITY

## Key findings of UNESCO's Global Dialogue

Prepared in August 2020 by the United Nations Educational, Scientific and Cultural Organization,
7, place de Fontenoy, 75352 Paris 07 SP, France



This Report was prepared by the Division for Gender Equality, UNESCO

Graphic design: Anna Mortreux

Printed by UNESCO
The printer is certified Imprim'Vert®, the French printing industry's environmental initiative

GEN/2020/AI/2 REV

# CONTENTS

# Foreword

**Advancing gender equality through education, the sciences, culture, information and communication lies at the heart of UNESCO's mandate, with Gender Equality constituting one of the two Global Priorities of the Organization since 2008. This means that UNESCO applies a gender equality lens to all its initiatives, including its normative work.**

The present report builds on UNESCO's previous work on gender equality and AI and aims to continue the conversation on this topic with a select group of experts from key stakeholder groups. In March 2019, UNESCO published a groundbreaking report, *I'd Blush if I Could: closing gender divides in digital skills through education*, based on research funded by the German Federal Ministry for Economic Cooperation and Development. This report featured recommendations on actions to overcome global gender gaps in digital skills, with a special examination of the impact of gender biases coded into some of the most prevalent AI applications.

The recommendations concerning AI's gender biases are urgent in light of the explosive growth of digital voice assistants such as Amazon's Alexa and Apple's Siri. Almost all voice assistants are given female names and voices, and express a 'personality' that is engineered to be uniformly subservient. As the UNESCO report explains, these biases are rooted in stark gender imbalances in digital skills education and are exacerbated by the gender imbalances of the technical teams developing frontier technologies by companies with significant gender disparities in their C-suites and corporate boards.

The release of *I'd Blush if I Could* has helped spark a global conversation on the gendering of AI technology and the importance of education to develop the digital skills of women and girls. Over 600 media reports have been published on it by outlets across the world, including *The New York Times*, *The Guardian*, *Le Monde*, *El País*, *Der Spiegel*, and many others. The conversation was subsequently taken up by influential global fora, such as the Web Summit in Lisbon, Portugal, considered as the largest tech event in the world. At the Summit's 2019 edition, which gathered over 70.000 participants, I had the honor of discussing the gender biases and sexism displayed by digital voice assistants during a fireside chat with the journalist Esther Paniagua, launching the DeepTech stage of the Summit.

Building on this momentum, UNESCO planned a follow-up conference to coincide with International Women's Day in March 2020. The conference entitled 'Gender Equality and the Ethics of Artificial Intelligence: What solutions from the private sector", and funded by the German Government and Google-Exponent, aimed to continue the conversation with experts from the technology industry as well as from academia and research institutes. Unfortunately, we had to cancel this global event due to the COVID-19 pandemic.

Since rescheduling the conference was not possible, we decided to reorient our work and launch a Global Dialogue on Gender Equality and Artificial Intelligence (the Dialogue) with leaders in AI, digital technology and gender equality from academia, civil society and the private sector. We structured the Dialogue around eight questions that participants could answer either in writing or through a virtual interview session that was recorded.

The present report shares the main findings from experts' contributions to UNESCO's Dialogue on Gender Equality and AI, as well as additional research and analysis conducted by an external consultant, Jennifer Breslin1. The report provides recommendations on how to address gender equality considerations in AI principles. It also offers guidance to the public and private sectors, as well as to civil society and other stakeholders, regarding how to operationalize gender equality and AI principles. For

1 Executive Director of Futuristas, a policy organization that advocates for a broader science, technology, engineering, arts, and mathematics (STEAM) ecosystem that is inclusive, just and responds to the needs of society.

example, it highlights the need to increase general awareness within society regarding the negative and positive implications of AI for girls, women and gender non-binary people. Regarding the education sector, it also insists on the need to develop curricula and pedagogy that better integrate cross-disciplinary social sciences, ethics and technology literacy at the secondary and tertiary educational levels.

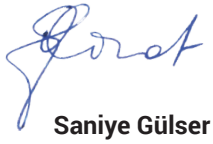
The timing of the Global Dialogue on Gender Equality and AI is also propitious to take the conversation forward in order to ensure that AI and AI codes of ethics are part of the solution, rather than part of the problem, in global efforts to achieve gender equality. In November 2019, at the 40th session of the General Conference, UNESCO Member States unanimously decided to mandate the Organization with developing a global normative instrument on the ethics of artificial intelligence to be submitted to the 41st session of the General Conference for its approval in November 2021. While UNESCO is preparing its draft Recommendation on the Ethics of Artificial Intelligence, the findings and recommendations of the Dialogue will provide the stakeholders with the opportunity to reflect on how best to integrate gender equality considerations into such global normative frameworks.

This report is the result of teamwork. First, I am grateful to the experts and leaders in the field of AI for taking the time to either talk to me via video or respond to our questions in writing. Without their input, we would not have been able to produce this document. These experts are (in alphabetical order):

- Rediet Abebe, Junior Fellow, Society of Fellows, Harvard University
- Getnet Aseffa, CEO, iCog Labs
- Maria Axente, Responsible AI and AI for Good Lead, PwC United Kingdom
- Samara Banno, Tech/AI Expert, Women Leading in AI
- Daniela Braga, Founder and CEO, DefinedCrowd
- Margaret Burnett, Distinguished Professor, School of Electrical Engineering and Computer Science, Oregon State University
- Christine Chow, Director, Global Tech Lead, Head of Asia and Emerging Markets, Federated Hermes
- Lori Foster, Research Partner and Advisor, pymetrics
- Allison Gardner, co-Founder of Women Leading in AI
- Sara Kassir, Senior Policy and Research Analyst, pymetrics
- Genevieve Macfarlane Smith, Associate Director, Center for Equity, Gender and Leadership, UC Berkeley Haas School of Business
- Saralyn Mark, Founder and President, iGIANT
- Mike McCawley, Chief Architect, IBM Watson in Support
- Frida Polli, CEO, pymetrics
- Londa Schiebinger, John L. Hinds Professor of History of Science, Stanford University
- Elizabeth Stocks, Tech/AI Expert, Women Leading in AI
- Kelly Trindel, Head of Policy & I/O Science, pymetrics

I am indebted to Jennifer Breslin for travelling to Manhattan in March, during the scary early days of the pandemic to meet with me and for accepting without hesitation to collaborate on this project. She put in an extraordinary effort under difficult and tight timelines to conduct stellar research and prepare a draft. Blandine Bénézit has been indispensable in this process since November 2019. Despite the disappointment of the cancellation of the global conference she helped organize for 7 March 2020 at UNESCO, she rallied behind the idea of re-orienting the work and provided support with this Dialogue by preparing the questions, transcribing recordings and editing, with Anne Candau, the final report.
I would be remiss if I did not mention Bruno Zanobia and Mary Joy Brocard from my team for their logistical, communications and design support for the Conference and for the Dialogue. I also wish to express my sincere gratitude to the staff, consultants and interns of UNESCO's Division for Gender Equality; International Women's Day Programme Working Group; and the AI Intersectoral Task Team for their input and feedback at different stages of the process.

This report is the final product of a 3-year partnership and generous financial support of the German Government. Our collaboration that started with a wonderful, trusting working relationship with Norman Schraepel in 2017 continued with Dr Michael Holländer, both from the German Agency for International Cooperation (GIZ). I am particularly grateful to Michael for his unwavering support, trust, solidarity and flexibility without which I could not have been able to navigate this initiative through unexpected developments, including a pandemic.



**Saniye Gülser Corat**
Director for Gender Equality, UNESCO
Paris, 26 August 2020

# Introduction

**Simply put, artificial intelligence (AI) involves using computers to classify, analyze, and draw predictions from data sets, using a set of rules called algorithms. AI algorithms are trained using large datasets so that they can identify patterns, make predictions, recommend actions, and figure out what to do in unfamiliar situations, learning from new data and thus improving over time. The ability of an AI system to improve automatically through experience is known as Machine Learning (ML).**

While AI thus mimics the human brain, currently it is only good, or better than the human brain, at a relatively narrow range of tasks. However, we interact with AI on a daily basis in our professional and personal lives, in areas such as job recruitment and being approved for a bank loan, in medical diagnoses, and much more.

The AI-generated patterns, predictions and recommended actions are reflections of the accuracy, universality and reliability of the data sets used, and the inherent assumptions and biases of the developers of the algorithms employed. AI is set to play an even more important role in every aspect of our daily lives in the future. It is important to look more closely at how AI is, and will affect gender equality, in particular women, who represent over half of the world's population.

Research, including UNESCO's 2019 report *I'd Blush if I Could: closing gender divides in digital skills through education*, unambiguously shows that gender biases are found in Artificial Intelligence (AI) data sets in general and training data sets in particular. Algorithms and devices have the potential of spreading and reinforcing harmful gender stereotypes. These gender biases risk further stigmatizing and marginalizing women on a global scale. Considering the increasing ubiquity of AI in our societies, such biases put women at risk of being left behind in all realms of economic, political and social life. They may even offset some of the considerable offline progress that countries have made towards gender equality in the recent past.

AI also risks having a negative impact on women's economic empowerment and labour market opportunities by leading to job automation. Recent research by the IMF[1] and the Institute for Women's Policy Research[2] found that women are at a significantly higher risk of displacement due to job automation than men. Indeed, the majority of workers holding jobs that face a high-risk of automation, such as clerical, administrative, bookkeeping and cashier positions, are women. It is therefore crucial that women are not left behind in terms of retraining and reskilling strategies to mitigate the impact of automation on job losses.

On the other hand, while AI poses significant threats to gender equality, it is important to recognize that AI also has the potential to make positive changes in our societies by challenging oppressive gender norms. For example, while an AI-powered recruitment software was found to discriminate against women, AI-powered gender-decoders help employers use gender-sensitive language to write job postings that are more inclusive in order to increase the diversity of their workforce. AI therefore has the potential of being part of the solution for advancing gender equality in our societies.

Building on the momentum of the report *I'd Blush if I Could* and subsequent conversations held in hundreds of media outlets and influential conferences, UNESCO invited leaders in AI, digital technology and gender equality from the private

sector, academia and civil society organizations to join the conversation on how to overcome gender biases in AI and beyond, understanding that real progress on gender equality lies in ensuring that women are equitably represented where corporate, industry and policy decisions are made. UNESCO, as a laboratory of ideas and standard setter, has a significant role to play in helping to foster and shape the international debate on gender equality and AI.

The purpose of the UNESCO's Dialogue on Gender Equality and AI was to identify issues, challenges, and good practices to help:

- Overcome the built-in gender biases found in AI devices, data sets and algorithms;
- Improve the global representation of women in technical roles and in boardrooms in the technology sector; and
- Create robust and gender-inclusive AI principles, guidelines and codes of ethics within the industry.

This Summary Report sets forth proposed elements of a Framework on Gender Equality and AI for further consideration, discussion and elaboration amongst various stakeholders. It reflects experts' inputs to the UNESCO Dialogue on Gender Equality and AI, as well as additional research and analysis. This is not a comprehensive exploration of the complexities of the AI ecosystem in all its manifestations and all its intersections with gender equality. Rather, this is a starting point for conversation and action and has a particular focus on the private sector.

It argues for the need to

1. Establish a whole society view and mapping of the broader goals we seek to achieve in terms of gender equality;

2. Generate an understanding of AI Ethics Principles and how to position gender equality within them;

3. Reflect on possible approaches for operationalizing AI and Gender Equality Principles; and

4. Identify and develop a funded multi-stakeholder action plan and coalition as a critical next step.

# FRAMING THE LANDSCAPE OF GENDER EQUALITY AND AI

**'Algorithmic failures are ultimately human failures that reflect the priorities, values, and limitations of those who hold the power to shape technology. We must work to redistribute power in the design, development, deployment, and governance of AI if we hope to realize the potential of this powerful advancement and attend to its perils.'**
**Joy Buolamwini[3]**

## AI AND SOCIAL GOOD

While AI can be used to make our daily lives easier, for example by driving our car or helping us find the perfect match on a dating app, it can also be used to help solve some of the world's biggest and most pressing challenges. To this end, technology companies, such as Google and Microsoft, have put in place 'AI for Good' or 'AI for Social Good' programmes that seek to use AI to solve humanitarian and environmental challenges. For example, AI could contribute to predicting natural disasters before they happen, protecting endangered species or tracking diseases as they spread in order to eliminate them sooner.[4]

Although the term 'AI for good' is increasingly used by technology companies and civil society organizations, there is much less discussion about what actually constitutes 'social good'. While using AI for social good is indeed commendable, it must start with a conversation about what is 'good' and with an understanding of the linkages between AI and our societal goals.

What are our values and what are the transformative goals to which AI is being applied? Are we continuing what has been described as an AI race and 'frontierism'[5], or moving to a more thoughtful model that better serves humanity? Aspirations and considerations have been raised around public interest and social good as a driver of AI (versus being largely at present concentrated in, and driven by, the private sector). UNESCO, for example, is advocating for a humanistic approach to AI that would ensure that AI contributes to the realization of the Sustainable Development Goals (SDGs) and of human rights frameworks. Others are asking for the decolonization of AI[6], arguing that AI has become a tool of digital colonialism, whereby 'poorer countries are overwhelmed by readily available services and technology, and cannot develop their own industries and products that compete with Western corporations'.[7] Others still are calling for a use of AI that protects data rights and sovereignty, expands collective as well as individual choices, dismantles patriarchy, the neo-liberal order and late stage/extractive capitalism, and promotes human flourishing over relentless economic growth.[8] Gender equality is necessary for the realization of any and all of the goals above, as well as being an objective in and of itself. It thus needs to be mainstreamed and considered at the highest level of outcomes and societal imperatives.

However, establishing this is not necessarily a straightforward process.

What happens when there are competing or different views of human values and social good? Or when ethical imperatives clash with each other or with national or local laws? And while it might be theoretically possible to find consensus on

grave harms to be avoided through AI, are we prepared to designate areas of social use where harm carries greater repercussions and apply higher standards of caution and accountability, or not pursue it at all?[9] One must also be attentive to the potential negative consequences of AI in the name of social good. For example, an AI designed to increase efficiency or effectiveness could cause people to lose their jobs and thus their livelihoods. Moreover, if an AI system is poorly trained or designed, or used incorrectly, flaws may arise. In medicine, for example, 'bad AI' can lead to misdiagnoses. In that case, false-positive results could cause distress to the patients, or lead to wrong and unnecessary treatments or surgerie. Even worse, false-negative results could mean patients go undiagnosed until a disease has reached its terminal stage.[10] AI also risks reproducing sexist, racist or ableist social structures if those developing and deploying it do not critically interrogate current practices. The complexity and nuances want for easy solutions.

Secondly, we must understand the AI and technology ecosystem. What are the components and prevailing practices around incentives and investments, governance and power, and broader issues around technology and society? Who has access? Whose priorities are reflected? Who benefits and who is harmed? Who takes decisions? This requires an examination of systemic structural issues like governance, policy, regulation, funding, societal norms, avenues for participation, and the like.[11]

## THE IMPERATIVES OF GENDER EQUALITY

Likewise, there is a need to increase our understanding of the landscape and imperatives of gender equality and women's empowerment in order to appreciate and address how it interfaces with AI.

At the level of global women's human rights, the Beijing Platform for Action, the UN Convention on the Elimination of All Forms of Discrimination against Women, and the work of the UN Commission on the Status of Women, are essential reference points for establishing a more comprehensive understanding of the persistent and entrenched structural and micro-level challenges of gender equality around the world.[12] The UN Sustainable Developing Goals provide another guide to realizing gender equality as a goal in and of itself (SDG 5) and as a necessary lever for achieving all the other interlinked SDGs[13]

These bodies of work reflect commitments to and measurement of progress around issues such as:

- Women's political participation and representation in decision-making;
- The elimination of discriminatory practices in institutions, budgets, law and access to justice;
- Changing negative social norms and gender stereotypes and valuing different 'ways of knowing' and social practices;
- Eliminating gender-based violence;
- Women's unpaid care and domestic work;
- Women's economic empowerment and financial inclusion;
- Gender responsive social protection and access to basic services, infrastructure and digital inclusion;
- Women's roles in peace and security and in humanitarian contexts;
- Strengthening women's roles in environmental sustainability and resilience;
- Access to health care and sexual and reproductive rights;
- Access to quality education and life-long learning; and
- Women's participation in culture, media and STEM, including traditional knowledge.

Context and issues of intersectionality[14] must also drive efforts for transformative gender equality. Women are a multifaceted and heterogeneous group and have different experiences based on realities or characteristics which include: women

living in rural and remote areas; indigenous women; racial, ethnic or religious minority women; women living with disabilities; women living with HIV/AIDS; women with diverse sexual orientations and gender identities; younger or older women; migrant, refugee or internally displaced women or women in humanitarian settings. The importance of taking intersectionality into account in AI principles was raised by a number of participants in the UNESCO Dialogue (Abebe, Schiebinger, Smith). For example, Abebe noted that 'much of the discourse around gender equality considerations can be narrow and assume uniform experience of people belonging to one gender and, as a result, AI principles for women, for instance, can be designed with specific kinds of women in mind.' In addition, participants pointed to the importance of understanding gender as non-binary to ensure that gender equality principles for AI are as inclusive as possible.

It cannot be expected that the AI industry (nor even gender practitioners themselves) will become expert in each and every area of intervention necessary to achieve transformative gender equality. However, there needs to be a robust level of awareness of the vastness, complexity, and persistency that is gender inequality so as not to ever essentialize it or approach it in an overly simplistic way. It is critical that there are commitments and mechanisms to bring in experts and affected groups. Otherwise, it may only be the most egregious, obvious, and familiar forms of gender inequality – as understood by a more select group or based on select assumptions – that are addressed.

In order for these gender equality goals, values and considerations to be manifested in the technology space, some argue that there needs to be a 'critique and framework that offers not only the potential to analyse the damaging effects of AI, but also a proactive understanding on how to imagine, design and develop an emancipatory AI that undermines consumerist, misogynist, racist, gender binarial and heteropatriarchal societal norms'.[15]

Developing an understanding of the gap between needs and action, and between promise and harm, in AI and gender equality to date would also prove informative and help to cut through hype, platitudes, and reductive approaches. This should include an examination of the displacing effect that AI has on women in the workforce who may be disproportionately represented in sectors that are undergoing automation. It should also include a more nuanced assessment of data representativeness and the systems-level efforts required to fix this. If one looks at women's representation in the online contents that comprise the data training our AI, it is clear that we have a long way to go.[16]

Finally, a discussion on gender and AI must include a more serious interrogation of why gaps between men and women's contributions to the development and use of AI, and digital technology in general, are not adequately shrinking and why gaps such as the digital gender divide are actually growing.[17]

# GENDER EQUALITY AND AI PRINCIPLES

## GENDER EQUALITY IN EXISTING PRINCIPLES

Direct references to gender equality and women's empowerment in existing AI and ethics principles are scarce. A scan of several major AI and ethics principles, as well as meta-level analyses combining data from these principles in addition to the examination of principles from other relevant sources, point to a few different ways in which gender equality is generally being treated.

### SUMMARIES OF THE META-ANALYSES

Several meta-level analyses have been undertaken – by Harvard, Nature Machine Intelligence, the AI Ethics Lab, and UNESCO[18] – of the AI ethics principles that have been developed over the past few years by the private sector, governments, inter-governmental organizations, civil society and academia.

The following are their findings with respect to the categories of principles and those most commonly found across frameworks. While there are commonalties, there is also divergence in meaning and interpretation. For example, while fairness is often included as one of the key principles, its definition usually varies across frameworks.

**UNESCO** undertook a review of AI principles and frameworks in 2020 and suggested the following five foundational values from which flow closely linked principles and then policy actions to implement them.

The foundational values are Human dignity; Human Rights and Fundamental Freedoms; Leaving no one Behind; Living in harmony; Trustworthiness; and the Protection of the Environment.[19]

UNESCO also references Recommendation 3C regarding AI, of the UN Secretary-General's High Level Panel on Digital Cooperation which 'has preliminarily identified consensus on the following fifteen principles: accountability, accessibility, diversity, explainability, fairness and non-discrimination, human-centricity, human-control, inclusivity, privacy, reliability, responsibility, safety, security, transparency, and trustworthiness. COMEST[20] has also identified the following generic principles, many of which overlap with the ones above: human rights, inclusiveness, flourishing, autonomy, explainability, transparency, awareness and literacy, responsibility, accountability, democracy, good governance, and sustainability'.[21]

### A FEMINIST VIEWPOINT

Other AI and ethics principles have been drafted by organizations that speak more specifically to gender equality and feminist theory. These perspectives and frameworks have not been fully considered in the above analysis (although it is possible that these views were provided in an effort to inform the development of the frameworks considered here).

Some call for feminist approaches that challenge the neo-liberal order, data exploitation, colonialism and AI's lack of transparency.[22] These approaches also warn against the risk of thinking that because AI is based on abstract mathematical algorithms, it can therefore reveal pure, objective and universal truths. In order to prevent people from relying and trusting AI too much, scholars recommend 'maintaining the focus on the humans behind the algorithms'.[23] According to Sareeta Amrute, Associate Professor of Anthropology at the University of Washington, 'making humans accountable for the algorithms they design

## AI Ethics Lab

**Autonomy**

- Power to decide (whether to decide)
- Human control
- Human oversight
- Transparency (to understand)
- Openness (to understand)
- Explainability
- Explicability
- Liberty
- Freedom
- Fundamental rights
- Personal privacy
- Privacy protection
- Fundamental rights
- Human values

**No Harm**

- Control risks
- Safety
- Security
- Capability caution
- Data protection
- Privacy (to avoid harm)
- Explicability
- Transparency (to avoid harm)
- Reproducibility
- Accuracy
- Reliability
- Responsible deployment
- Prevent arms race

**Benefit**

- Promoting well-being
- Benefit society
- Generating net benefits
- Sustaining the planet
- Impact
- Efficacy
- Explicability
- Scientific excellence
- User-centered design (for user benefit)
- People-first approach

**Justice**

- Fairness
- Fundamental rights
- Equality
- Non-discrimination
- Avoiding bias
- Inclusivity
- Diversity
- Data neutrality
- Representative data
- Shared benefit / prosperity
- Social & economic impacts
- Avoid disparity
- Mitigating social dislocation
- Preserving solidarity
- Accessibility
- Explicability
- Transparency (for accountability)
- Openness (for accountability)
- Accountability
- Auditability
- Liability
- Inclusive
- Judicial transparency
- Open governance

## Harvard

The principles within each theme are:

**Privacy**:
Privacy
Control over Use of Data
Consent
Privacy by Design
Recommendation for Data Protection Laws
Ability to Restrict Processing
Right to Rectification
Right to Erasure

**Accountability**:
Accountability
Recommendation for New Regulations
Impact Assessment
Evaluation and Auditing Requirement
Verifiability and Replicability
Liability and Legal Responsibility
Ability to Appeal
Environmental Responsibility
Creation of a Monitoring Body
Remedy for Automated Decision

**Safety and Security**:
Security
Safety and Reliability
Predictability
Security by Design

**Transparency and Explainability**:
Explainability
Transparency
Open Source Data and Algorithms
Notification when Interacting with an AI
Notification when AI Makes a Decision about an Indi
Regular Reporting Requirement
Right to Information
Open Procurement (for Government)

**Fairness and Non-discrimination**:
Non-discrimination and the Prevention of Bias
Fairness
Inclusiveness in Design
Inclusiveness in Impact
Representative and High Quality Data
Equality

**Human Control of Technology**:
Human Control of Technology
Human Review of Automated Decision
Ability to Opt out of Automated Decision

**Professional Responsibility**:
Multistakeholder Collaboration
Responsible Design
Consideration of Long Term Effects
Accuracy
Scientific Integrity

**Promotion of Human Values**:
Leveraged to Benefit Society
Human Values and Human Flourishing

## Nature Machine Intelligence

**Table 3 | Ethical principles identified in existing AI guidelines**

| Ethical principle | Number of documents | Included codes |
|---|---|---|
| Transparency | 73/84 | Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing |
| Justice and fairness | 68/84 | Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversi plurality, accessibility, reversibility remedy, redress, challenge, acces and distribution |
| Non-maleficence | 60/84 | Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion |
| Responsibility | 60/84 | Responsibility, accountability, liability, acting with integrity |
| Privacy | 47/84 | Privacy, personal or private information |
| Beneficence | 41/84 | Benefits, beneficence, well-being, peace, social good, common good |
| Freedom and autonomy | 34/84 | Freedom, autonomy, consent, choice, self-determination, liberty, empowerment |
| Trust | 28/84 | Trust |
| Sustainability | 14/84 | Sustainability, environment (nature), energy, resources (energ |
| Dignity | 13/84 | Dignity |
| Solidarity | 6/84 | Solidarity, social security, cohesio |

shows that the biases the algorithms produce are far from inevitable'.[24] These approaches also challenge AI designers' tendency to put the onus on the individual. As stated by Amrute: 'most often, designers of technical systems begin with a standard user [in mind] and, in doing so, set into motion patterns of discrimination that are hidden by the assumption of system neutrality'.[25]

Instead, feminist scholars call for a reworked political and economic approach, as well as a relational approach to ethics. This means that rather than ethical imperatives being developed in the abstract, in a top-down fashion, they ask for more applied and embodied ethics that take into account people's lived experiences, and 'asks whose knowledge counts and in what ways'.[26] Feminist theory thus 'moves the discussion of ethics from establishing decontextualised rules to developing practices to train sociotechnical systems—algorithms and their human makers—to begin with the material and embodied situations in which these systems are entangled, which include from the start, histories of race, gender and dehumanisation'.[27] As such, relational ethics allows for adjustments responding to 'how subjects and technologies are aligned and realigned, attached and reattached to one another [as] a method for practicing ethics that critically assesses a situation, imagines different ways of living, and builds the structures that make those lives possible.'[28]

## FEMINIST INTERNET PRINCIPLES

Though the Feminist Internet Principles[29] are not geared toward AI, many of them – in broad content or in spirit – are relevant when one replaces 'Internet' with 'AI' and undertakes extrapolation. These should be adapted and built on. They further inform, and in some cases already reflect, the principles in frameworks under consideration in this paper.

In a more condensed form, they include:

- **Access:** universal, open, affordable, unconditional, meaningful and equal access to the internet; to information in diversity of languages, abilities, interests and contexts; and to usage of technology including the right to code, design, adapt and critically and sustainably use and reclaim, as well as a platform to challenge cultures of sexism and discrimination in all spaces;
- **Movements and Political Participation:** Technology is a place where **social norms are negotiated**, performed, imposed – and currently shaped by patriarchy – and is a **space for feminist resistance** and movement building; Governance of technology requires challenging patriarchal processes, **democratizing policy making**, **diffusing ownership** of and power in global and local networks, and **requires gender equality advocates/specialists at the decision-making table**;
- **Economy**: Challenge the capitalist logic that drives technology towards further privatization, profit and corporate control, and **create alternative forms of economic power that are grounded in principles of cooperation, solidarity, commons, environmental sustainably, and openness**; as well as commitment to free and open source software, tools and platforms;
- **Expression:** Claim the power of technology to **amplify women's narratives and lived realities, and defend the right to sexual expression. Object to efforts of state and non-state actors to control, surveil, regulate and restrict this sexual expression**, for example through the labeling of content that restricts expression of sexuality (e.g. as pornography or harmful content);
- **Agency:** build an **ethics and politics of consent into the culture, design, policies and terms of technology with women's agency lying in their ability to make informed decisions on what aspects of their public or private lives to share.** Support the right to **privacy and full control over personal data** and information online at all levels, with a **rejection** of practices by states and private companies to use data for profit and to **manipulate behavior, as well as surveillance practices** by any actor to control and restrict women's bodies, speech and activism;
- **Memory:** Right to exercise and retain **control over personal history and memory**, including being able to access all personal data and information online, be able to exercise control over this data, including who has access to it and under what conditions, and the ability to delete it forever;

- **Anonymity**: Defend the **right to be anonymous** and reject all claims to restrict anonymity;
- **Children and Youth**: Call for the **inclusion of the voices and experiences of young people in the decisions made about safety and security online and promote their safety, privacy, and access to information.** Recognize children's right to healthy emotional and sexual development, which includes the **right to privacy and access to positive information about sex, gender and sexuality at critical times in their lives**;
- **Online Violence**: Call all stakeholders to **address harassment and technology-related violence.**

## GENDER DIMENSIONS

When asked whether there exists any AI normative instruments or principles that successfully address gender equality, UNESCO Dialogue participants argued that these were either inexistent (Braga) or that current practices were insufficient (Abebe). Further reflections include:

- Smith: 'I have not seen any [AI normative instruments or principles that successfully address gender equality] and think it's a big gap.'
- Abebe: 'There are insufficient efforts and guidelines to ensure equitable benefit from and mitigation of harm of the development, deployment, access, and use of AI technologies.'
- Kassir: 'Legal standards need to be formalized, so that terms like "ethical" and "de-biased" are formally defined.'

Participants acknowledged that the existing AI principles of fairness, transparency, accountability and explainability tended to include gender equality, although sometimes only implicitly. As encouraging examples, participants pointed to existing frameworks such as the EU's Ethics Guidelines for Trustworthy Artificial Intelligence, the UN Human Rights Framework, the Asilomar AI Principles, Google's AI Principles, those of the Partnership on AI, and of the Australian Government.

In examining the discussions on the meta-analyses, as well as reviewing the principles of the Institute of Electrical and Electronics Engineers (IEEE), the European Union (EU), UNESCO, the Organization for Economic Co-operation and Development (OECD), Microsoft and IBM as a sampling, the following indicates where and how gender equality, sex discrimination, or women's empowerment tends to surface.

### EXPLICIT

Direct reference to gender can be found in provisions on non-discrimination, diverse teams, data privacy, vulnerable groups, accessibility, achieving the SDGs, human values/beneficence, among others. However, these sometimes appear in background sections or in referencing gender-related issues alongside a long list of other issues.

Most frequently, these gender-related issues are categorized under 'fairness' which varies across frameworks but includes bias and discrimination.[30] The Harvard meta-study noted: 'the Fairness and Non-discrimination theme is the most highly represented theme in our dataset, with every document referencing at least one of its six principles: "non-discrimination and the prevention of bias," "representative and high-quality data," "fairness," "equality," "inclusiveness in impact," and "inclusiveness in design."'[31]

With regards to fairness, Tannenbaum et al. point out that 'to date, there is no unified definition of algorithmic fairness.' [32]This echoes Collett and Dillon's finding that there are 'more than twenty different definitions of fairness circulating in academic work, some focusing on group fairness and others focusing on individual fairness.'[33] Given the absence of a unified definition, Tannenbaum et al. argue that 'the best approach is to understand the nuances of each application domain, make transparent how algorithmic decision-making is deployed and appreciate how bias can arise.'[34]

A list of examples of explicit references to gender equality in selected normative texts is presented in Annex 1.

## IMPLICIT OR ALIGNED VALUES AND OBJECTIVES WITH GENDER EQUALITY

The second category is that of implicit reference, or generally aligned values and objectives between stated principles (or their implementation mechanisms) and reducing gender-based discrimination or realizing gender equality and women's empowerment. These include:

- **Gender as Part of 'Groups' or under Bias/ Discrimination/Equality-related principles:** Women and gender status are assumed to be included in the often mentioned categories of marginalized groups, vulnerable groups, those suffering from discrimination and inequalities, excluded groups, and in framings such as 'regardless of group (…) or certain attribute' type language, reference to equality, fairness (as noted above), and inclusiveness. The Harvard study notes that: 'there are essentially three different ways that equality is represented in the documents in [their] dataset: in terms of human rights, access to technology, and guarantees of equal opportunity through technology. In the human rights framing, the Toronto Declaration notes that AI will pose "new challenges to equality" and that "[s]tates have a duty to take proactive measures to eliminate discrimination." In the access to technology framing, documents emphasize that all people deserve access to the benefits of AI technology, and that systems should be designed to facilitate that broad access. (…) The Montreal Declaration asserts that AI systems "must help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge" and "must produce social and economic benefits for all by reducing social inequalities and vulnerabilities." This framing makes clear the relationship between the "equality" principle and the principles of "non-discrimination and the prevention of bias" and "inclusiveness in impact.'"[35]

- **Values and Impact - Human Values, Well-Being, Beneficence, Impact:** A range of principles that speak further to inclusiveness, impact, social good, and human well-being, all of which implicitly – in theory – include women and the realization of (aspects of) gender equality.
  - **Inclusiveness in Impact:** According to the Harvard report, '"Inclusiveness in impact" as a principle calls for a just distribution of AI's benefits, particularly to populations that have historically been excluded.'[36] The study revealed remarkable consensus in the language that documents employ to reflect this principle, including concepts like 'shared benefits' and 'empowerment'.

| DOCUMENT | LANGUAGE |
|---|---|
| ASILOMAR AI PRINCIPLES | Shared Benefit: AI technologies should benefit and empower as many people as possible. |
| MICROSOFT'S AI PRINCIPLES | Inclusiveness - AI systems should empower everyone and engage people. If we are to ensure that AI technologies benefit and empower everyone, they must incorporate and address a broad range of human needs and experiences. Inclusive design practices will help system developers understand and address potential barriers in a product or environment that could unintentionally exclude people. This means that AI systems should be designed to understand the context, needs and expectations of the people who use them. |
| PARTNERSHIP ON AI TENETS | We will seek to ensure that AI technologies benefit and empower as many people as possible. |
| SMART DUBAI AI PRINCIPLES | We will share the benefits of AI throughout society: AI should improve society, and society should be consulted in a representative fashion to inform the development of AI. |
| T20 REPORT ON THE FUTURE OF WORK AND EDUCATION | Benefits should be shared: AI should benefit as many people as possible. Access to AI technologies should be open to all countries. The wealth created by AI should benefit workers and society as a whole as well as the innovators. |
| UNI GLOBAL UNION'S AI PRINCIPLES | Share the Benefits of AI Systems: AI technologies should benefit and empower as many people as possible. The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity. |

*Source:* Fjeld, J et al. 2020.

- **Benefits:** 'The European High Level Expert Group guidelines add some detail around what "benefits" might be shared: "AI systems can contribute to well-being by seeking achievement of a fair, inclusive and peaceful society, by helping to increase citizen's mental autonomy, with equal distribution of economic, social and political opportunity." There is a clear connection to the principles we have catalogued under the Promotion of Human Values theme, especially the principle of "leveraged to benefit society."'[37]
- **Inclusiveness in Design:** Design teams and societal decision-making should contribute to the discussion on what we use AI for, and in what contexts, 'specifically, that there should be "a genuinely diverse and inclusive social forum for discussion, to enable us to democratically determine which forms of AI are appropriate for our society." The Toronto Declaration emphasizes the importance of including end users in decisions about the design and implementation of AI in order to "ensure that systems are created and used in ways that respect rights – particularly the rights of marginalized groups who are vulnerable to discrimination." This interpretation is similar to the Multistakeholder Collaboration principle in [the] Professional Responsibility category, but it differs in that it emphasizes bringing into conversation all of society – specifically those most impacted by AI – and not just a range of professionals in, for example, industry, government, civil society organizations, and academia.'[38]

- **Human Values:** According to the authors of the Harvard study, 'the Promotion of Human Values category consists of three principles: "human values and human flourishing," "access to technology," and "leveraged to benefit society."'[39], often with links to human rights.
  - 'The principle of "human values and human flourishing" is defined as the development and use of AI with reference to prevailing social norms, core cultural beliefs, and humanity's best interests.'[40] It should be noted however, that national level 'prevailing social norms' may in fact be discriminatory, contradict international frameworks on gender equality and women's rights. How is this reconciled?
  - 'Access to technology': includes access to technology itself, education, training, workforce and economic dimensions, and ability to use (Harvard).
  - 'Leveraged to benefit society' and human rights: to illustrate the prevalence of this principle the Harvard study states that 'Twenty-three of the documents in our dataset (64%) made a reference of this kind. We also noted when documents stated explicitly that they had employed a human rights framework, and of the thirty-six documents (14%) did so.'[41]

- '**Justice** is mainly expressed in terms of fairness, and of prevention, monitoring or mitigation of unwanted bias and discrimination, the latter being significantly less referenced than the first two by the private sector. Whereas some sources focus on justice as respect for diversity, inclusion and equality, others call for a possibility to appeal or challenge decisions or the right to redress and remedy. Sources also emphasize the importance of fair access to AI, data, and the benefits of AI. Issuers from the public sector place particular emphasis on AI's impact on the labour market and the need to address democratic or societal issues. Sources focusing on the risk of biases within datasets underline the importance of acquiring and processing accurate, complete and diverse data, especially training data.'[42]

- **Beneficence:** 'While promoting good ("beneficence" in ethical terms) is often mentioned, it is rarely defined, though notable exceptions mention the augmentation of human senses, the promotion of human well-being and flourishing, peace and happiness, the creation of socio-economic opportunities, and economic prosperity.'[43] Strategies include: 'minimizing power concentration or, conversely, using power "for the benefit of human rights", working more closely with "affected" people, minimizing conflicts of interests, proving beneficence through customer demand and feedback, and developing new metrics and measurements for human well-being.'[44]

- **Solidarity:** 'Solidarity is mostly referenced in relation to the implications of AI for the labour market. Sources call for a strong social safety

net. They underline the need for redistributing the benefits of AI in order not to threaten social cohesion and respecting potentially vulnerable persons and groups. Lastly, there is a warning of data collection and practices focused on individuals that may undermine solidarity in favour of "radical individualism".'[45]

## THE LESS OBVIOUS BUT STILL IMPORTANT

While not every principle may have an explicit or immediately evident link to gender equality, this does not mean it does not exist. In other principles that address Accountability, Responsibility, Transparency, Human Control of Technology, Privacy, Security, Remedy, and the like, if one correctly takes the view that AI is in its entirety a socio-technical system, i.e. a system that includes software, hardware and data, but also the people who build and use the technology and the laws and regulations that determine how and what technology can be developed and for what purpose, then issues of gender equality are inevitably present throughout.

For example, in her contribution to UNESCO's Dialogue, Margaret Burnett, Distinguished Professor at Oregon State University, emphasized that to reveal its gender biases, an AI system has to be:

- Transparent, meaning that it is able to communicate on how it is reasoning;
- Accountable, which is only possible if it is transparent; and
- Explainable to all, not just the educated and affluent few.

According to Tannenbaum et al., AI transparency is not just about being able to understand how an AI is reasoning, it is also about being 'completely transparent where and for what purpose AI systems are used', and about '[characterizing] the behavior of the system with respect to sex and gender.'[46]

The accessibility of AI also came up as a core principle in participants' contributions to UNESCO's Dialogue. Braga for example stated that 'everyone should have the right to access AI, like access to healthcare and education.' She said: 'Insofar as it is a service you have to pay for, it's difficult. It has to be more democratized. And no one should be discriminated against (age, race, colour, knowledge…). There is a level of digital skills to access this technology and some economic wealth, but these barriers should be diminished.' In a similar vein, Getnet Aseffa from Ethiopia's iCog Labs lamented the very high digital gender divide in Ethiopia and noted the need for the education system to do more to address gender equality.

Women as affected groups, as stakeholders, rights holders, experts on gender equality and women's empowerment, and as creators and policy makers should therefore be included and present in these capacities to generate awareness, inform, implement, and monitor *all* of the principles.

## MISSING ELEMENTS

Finally, it is worth undertaking a review with gender equality experts and affected groups on where there are gaps and if there are either broader or sub-principles, or specific language that should be included. Alternatively, is it enough that any gaps will be filled through interpretation, contextualization, and the operationalization phase?

# RECOMMENDATIONS FOR INTEGRATING GE INTO AI PRINCIPLES

**In order to effectively integrate gender equality into AI principles, it will be important to take a rights-based and comprehensive gender equality approach. This means giving substance, meaning, nuance, and specificity to principles. It requires filling gaps and addressing challenges at a systemic level. It also means being more anticipatory, considering future scenarios, and being proactive. The risk of not integrating these measures means that abstract concepts may be ignored, or necessary interventions may be absent all together. The following are further considerations and needs.**

## PROCESS OF PRINCIPLE DEVELOPMENT

Ensure gender equality experts and women's participation in the initial process of principle formulation and in their ongoing interpretation, application, monitoring, reporting, and recalibration. This is important whether being done at an intergovernmental level, within sectors, or at the institutional level.

## CONTENT OF PRINCIPLES

### Where Gender is Located: gender equality as a stand-alone principle

Lessons from past efforts on gender and technology are worth considering. When gender is made explicit, and when it is left as implicit, should be deliberate and thoughtful. It also matters where the explicit references are made. For instance, is gender equality in a long list, is it in a preamble, is it in the main body of principles and in their implementation? Vague, overarching references and placement where there is little 'teeth' or direct focus, or conversely, where it is everywhere and therefore nowhere, risk that gender receives less attention and follow through, accountability or stricter monitoring. Notably, UNESCO states that 'concern for *gender* equality is sometimes bundled as a subset of a general concern for bias in the development of algorithms, in the datasets used for their training, and in their use in decision-making. However, gender is a much larger concern, which includes, among others, women's empowerment linked with their representation of women among developers, researchers and company leaders, as well as access to initial education and training opportunities for women, which all require the consideration of gender equality as a stand- alone principle.'[47]

### Whole of Society, Systems, and Lifecycle Approach

Issues of gender equality and AI need to be considered with a lifecycle approach in terms of technology development and implementation, and a systems approach in terms of how AI is contextualized in structural issues and in the broader technology and society discussion. This is not just about specific algorithms or datasets. One aspect of this is recognizing relationships and power, between actors, from local to larger scales, around private and public sectors, and

within society – between men and women in all spheres.[48] The Harvard study notes that some approaches are 'technochauvinistic', where solutions are found in the data and technology itself. Rather, it recommends the approach of the Toronto Declaration: 'All actors, public and private, must prevent and mitigate against discrimination risks in the design, development and application of machine learning technologies. They must also ensure that there are mechanisms allowing for access to effective remedy in place before deployment and throughout a system's lifecycle.'[49]

## Addressing Gender Equality in AI: Avoiding Harm, Increasing Visibility, and Contributing to Empowerment

**Avoiding Harm:** Much of the standard accounting for gender equality is around avoiding or mitigating harm. While some examples may be obvious, avoiding harm still requires considerable nuance, 'un-packing', and the input of gender equality experts broadly and in specific domains. This requires proactive mitigation, monitoring, and bringing to light negative real-world impacts, as well as accepting that some things may not be able to be fixed and therefore should not be done at all, or should ultimately be abandoned (e.g. the example of Amazon's hiring algorithm which remained biased after multiple attempts to fix it). Moreover, while eliminating or reducing harm is of utmost importance, it is too reductive to equate gender equality with this one objective.

**Making the Invisible Visible:** Issues of gender equality and women's empowerment can also be simply omitted and passed over. Although this is often discussed in the context of the unrepresentativeness of data, women and gender issues can also be invisible depending on which questions are being addressed, by whom, and where priorities for AI lie. Omission and failure to consider gender equality can also come through ignorance (willful or not), and unconscious bias. In the UNESCO Dialogue, a number of experts noted how they had seen instances of people making incorrect assumptions about gender, even amongst those that were trying to account for it, leading to missed opportunities, or negative impacts.

**Empowerment:** Finally, there is a need to better understand and activate AI for emancipation and women's empowerment, for challenging norms and stereotypes, power dynamics, and so on. Can AI be used to create more empowering narratives and realities for girls and women? Can it be used to root out instances and patterns of discrimination and negative norms? Can it be used affirmatively to elevate girls and women? How can it be used to address positive goals around gender equality, like combatting gender-based violence, access to education, health, economic empowerment, political participation, peace and security, environmental justice?

A few of the UNESCO Dialogue participants insisted that we should be cautious when thinking about emancipatory or transformative AI (AI systems that would actively redress gender inequalities). In this respect, Gardner, co-Founder of Women Leading in AI, said that it is important to 'understand the limitation of what can be achieved' with technology.

Smith argued that an 'emancipatory AI' could be dangerous if it means that gender biases are hidden behind good intentions. She wrote: 'Many Machine Learning (ML) systems seek to "do good" by tackling existing gender issues/inequalities through use of AI. This is great, however, they are still at risk of perpetuating harmful bias. As Ruha Benjamin, author of *Race After Technology* notes, this "techno-benevolence" can cause significant harm as prejudice can be even harder to detect when the stated purpose is to override human or do good.'

Finally, representatives of pymetrics, a technology company that has developed an emancipatory AI-powered hiring tool, stressed that although emancipatory AI tools can have a positive and powerful impact for greater gender equality, they may also influence people in thinking that technology is a silver bullet that can solve problems of gender inequality on its own. The risk is that people become complacent and leave social justice issues to technology. In the words of Sara Kassir from pymetrics: 'stakeholders may sometimes overestimate the extent to which an AI tool can fully address a complex issue like equity in the workforce. As the first stage in the talent pipeline, pymetrics takes its role in the talent selection process very seriously, but the simple reality is that an assessment platform is only one part of a very large ecosystem. Organizations that are committed to promoting gender equality need to recognize the importance of multi-pronged solutions. Fair AI evaluations

are not enough; disparities in education, training, retention, and promotion need to be addressed, and inclusive cultures need to be fostered.' She adds: 'technology is an important tool to mitigate bias, but it is not a comprehensive solution to social problems.'

Across these three dimensions of avoiding harm, increasing visibility and contributing to empowerment, it would be useful to take the strongest elements of existing frameworks that respond to these imperatives and use them to inform action.

## Provision for Differentiated Impact on Women, Girls and Points of Intersectionality (and multiple forms of discrimination).

Whether or not gender and women are explicitly mentioned, there is still a need to further unpack the differentiated experience and impacts around AI for women and girls. These include things like redress for harm, where centuries of lessons, and more recent experiences of cyber-violence, show that access to justice can be challenging for women to realize. The Harvard analysis of AI frameworks likewise called for 'additional mapping projects that might illustrate narrower or different versions of the themes with regard to particular geographies or stakeholder groups.'[50]

## Provision for participation of Women/ Girls and Gender Equality Experts

These affected groups (at the macro societal level and micro level with specific use cases), should be included as developers of technology, as beneficiaries (if women and gender equality practitioners in various sectors use AI within their work), as informed and informing citizens and public (understanding women's implications and ability to engage in public comment, conversations and policy decisions), and as monitors and within accountability mechanisms (impact audits, public bodies, etc.).

## Prioritization and Tradeoffs

There is a need to understand the complexity and tradeoffs between issues of representative data collection versus privacy, monitoring versus censorship, data ownership and exploitation and women's and gender equality advocates' personal security. For example, Buolamwini suggests that 'a common response to uncovering severe data imbalances is to collect more data; however, how data is collected, categorized, and distributed presents ethical challenges around consent and privacy along with societal challenges with the politics of classifications where social categories like race and gender become reified into the technical systems that increasingly shape society.'[51]

What is the process for considering gender issues in 'negotiation' of different principles and their realization? How do we ensure that women are not deprioritized?[52]

# GENDER TRANSFORMATIVE OPERATIONALIZATION OF AI PRINCIPLES

**While participants in UNESCO's Dialogue acknowledged the importance of principles such as fairness and transparency, they lamented the lack of guidelines regarding their concrete implementation (Smith, Schiebinger, Chow). For example, Christine Chow asked: 'How do we operationalize these principles at the industry-level?' In her opinion, there is an urgent need to translate the 'rhetoric and big statements' into 'action'. Tannenbaum et al. also state: 'Numerous groups have articulated "principles" for human-centred AI. (…) What we lack are mechanisms for technologists to put these principles into practice.'[53]**

Many others have also noted that there is a gap in putting principles into actions and that this needs to be remedied. How we operationalize gender responsive principles will result – or not – in gender transformative AI. The following are illustrative actions that can be taken to begin to develop an effective AI for Gender Equality ecosystem. A comprehensive account of possibilities and steps should be developed.

## AWARENESS, EDUCATION AND SKILLS

'The public is largely unaware of the ways in which AI shapes their lives, and there are few regulations that require disclosure about the use of the technology. Without awareness about the uses, risks, and limitations of AI, we remain at the mercy of entities that benefit from opaque AI systems, even when they propagate structural inequalities and violate civil rights and liberties.' Joy Buolamwini[54]

In order for society at large, girls and women and the AI industry to truly understand, reap the benefits of and prevent negative impacts of AI, a robust approach is required to raise awareness and literacy, develop technical and ethical education and skills development, and build capacities for AI application. Most UNESCO Dialogue participants made the point that more training and awareness raising was needed in order to educate people about AI, what it is, how it works and how it can impact gender equality.

The following are some suggested steps to address gender inequalities in AI applications:

- **Shift the narrative**, language and framing of AI to improve public understanding and discourse. Media, journalism and academia play critical roles in demystifying and promoting accuracy, an evidence base and accountability.[55]

- **Improve education on AI and society, bias, and ethics**, for technical and non-technical audiences, particularly among professions and in areas where there are gaps – e.g. engineers and ethics/social sciences. As UNESCO notes, 'global engineering education today is largely focused on scientific and technological courses that are not intrinsically related to the analysis of human values nor are overtly designed to positively increase human and environmental wellbeing. It is most important to address

this issue and to educate future engineers and computer scientists to develop ethically aligned design of AI systems. This requires an explicit awareness of the potential societal and ethical implications and consequences of the technology-in-design, and of its potential misuse.'[56] It is similarly important to provide opportunities for gender equality advocates and practitioners to learn about and develop understanding on what AI is, how it works, and entry points for influence over its development, and its use to advance their goals.

- **Create programmes for development of AI Literacy**, or 'algorithm awareness', for individual girls and women to critically assess personal and societal risk and benefit from the impacts of AI, as well as to empower them as public advocates and agents in its development.

- Enable gender equality advocates, practitioners and women in different sectors to be able to identify opportunities and **build adequate skills to integrate and apply AI in different situations**, for emancipatory purposes and on a lifelong basis.

- **Improve access to quality STEM education and careers** for girls and women. There remains significant horizontal and vertical gender segregation[57] in STEM education and careers, in access to technology for learning and productive purposes. There are barriers across the STEM ecosystem preventing women from taking advantage of, and contributing to, opportunities in technology. AI is the latest example of these trends that must be reversed. See UNESCO's *Cracking the Code* and *I'd Blush if I Could* reports for more detail on the gender divide in STEM and digital skills.

- **Adopt a forward-looking view**, so that girls and women are not playing catch-up in understanding and skills and are not relegated to a reactive position.

## PRIVATE SECTOR/INDUSTRY

Some of the following suggestions are 'generic' and cut across a number of ethical implications and interest groups, beyond women. However, what is key is that there is a 'gender lens' adopted where appropriate. Some of these topics are just suggestions, others are being implemented in places. There is, however, further thought and discussion required in order to respond to many of the issues raised above.

### THE BIGGER PICTURE OF ETHICS

There is a clear imperative for the AI industry to embrace robust and gender transformative ethical principles. To this end, Mozilla suggests that this requires shifting the conversation from putting the onus on individuals and 'personal behavior' to system change.[58] In addition, Buolamwini notes the importance of reducing conflicts of interest between profit and AI ethics work.[59]

At the corporation level, Christine Chow, who is Director, Global Tech Lead and Head of Asia and Emerging Markets at Federated Hermes, emphasized the importance of a coordinated top-down/bottom-up process for the drafting of AI principles. She pointed to Intel, IBM and Microsoft as good examples of using such an approach. She also highlighted the importance of being forward-looking. She added that principles could not chase the problems of yesterday, but should rather anticipate what technology will be developed tomorrow, how it will be deployed and who will use it.

Collett and Dillon make the point that AI gender-specific guidelines should be research-informed and as specific and relevant as possible to AI systems.[60] Guidelines need to be context specific: either about AI in crime and policing, or AI and health, or financial services, etc.

Finally, UNESCO's role in establishing global recommendations regarding the ethics of AI was welcomed by participants of the Dialogue (Axente and Braga). Axente, for example, noted that UNESCO's value added was its global mandate.

### DATA

- **Increase awareness** within the industry and address issues of lack of representativeness and bias within training data. Data that is representative by one standard (e.g. equal numbers of women) may still be biased by another standard (e.g. those equal numbers of women are stereotyped in a negative way).

- Increase collection and use of **intersectional data**, especially from under-represented groups. For example, DefinedCrowd provides custom-built training data for its clients and

the Algorithmic Justice League, founded by Joy Buolamwini, provides such data sets for research.

- **Correcting for Data Bias:** Where possible, commit to using data sets that have been developed with a gender equality lens (despite potential cost increase), detecting and correcting for data bias and ensuring that data sets (including training data) represent the populations a machine learning application will affect, including by consulting with domain experts.[61] Dillon and Collett call for context-specific and gender-specific guidelines for best practice regarding data.[62] Guidelines would cover data collection, data handling and subject-specific trade-offs.

- Address **issues of consent and confirmation of ethical use of data**, privacy and security for women and girls (addressing parents as well) and understand the specific vulnerabilities which they face (Gardner).

- **Data quality control** and development of **common standards for assessing the adequacy of training data** and its potential bias through a multi-stakeholder approach.[63] Please also refer to the discussion below on Standards and Policies, and on Technical Solutions and Transparency.

- Address **socio-cultural dimensions of data selection, classification, Natural Language Processing (NLP), labels**, among other.

- **Support gender equality advocates in their broader work** around women in STEM, digital literacy (including content development), and gender equality across all disciplines that inform AI data training.

## STANDARDS AND POLICIES

**Codes of Conduct:** Codes for professional conduct, ethics and establishing responsibility need to be created for AI and software developers, similar to the medical profession, lawyers, journalists or the aviation industry.[64]

**Human Rights:** Create standards that address human rights risks in Machine Learning system design.[65]

**Certification:** According to Gardner, a certification in trained AI implementation should have these main components:

- Definition of data streams, and techniques to remove bias;
- Compliance with existing equality laws and regulations;
- Ethical use of data with confirmation agreement if required;
- Confidence level measurement through testing and deployment usage;
- Corrective algorithms to counteract imbalance of predictive capability; and
- Conscience algorithms to counteract heteronomy (actions that are influenced by a force outside the individual), to flag abuse, and to remove overstimulation within the AI (through adversarial behaviours).

As an example of certification, Gardner points to the IEEE standard P7003 on Algorithmic Bias Considerations that is currently being developed.

**Policies:** In his contribution to UNESCO's Dialogue, McCawley from IBM put the onus on the AI industry itself in helping policymakers develop AI policies: 'the industry must work extra hard to make AI accessible and explainable to regulators to help them develop policy' and 'get better at governing AI'. Finally, he emphasized the importance for governments to get it right, saying he feared for either too much or too little AI regulation.

**Gender Scorecard or Seal:** This could be developed by a body with deep expertise on gender equality that addresses a range of criteria leading to a systems level approach.

## INTERNAL MITIGATION AT COMPANY LEVEL

### CORPORATE GOVERNANCE

- Create or enhance company governance models and mechanisms for ethical compliance that include reflections on gender equality and the involvement of women and gender advocates.[66] Examples include creating an "Ethical Use Advisory Council" with experts and civil society groups, which integrates gender equality concerns. For example, Microsoft launched an internal high-level group on AI and Ethics in Engineering and Research (AETHER) which examines how its software should, or should not, be used.

- Create highest-level policies that signal management support for advancing gender equality through corporate products and services.

- Enhance and leverage existing processes and mechanisms to increase attention to gender equality rather than creating additional layers. 'What can be done more readily, make as easy as possible to pick up'.[67] On the other hand, harder steps could be taken, as this is necessary for culture change.

- Create incentives for non-biased products including by linking this to job promotions and opportunities to participate in development teams (IBM does this based on participation in bias training).

**Sponsors and Supporters:** Identify sponsors and supporters throughout the organization and across teams.[68] This is not the same as Affinity Groups (which typically focus on workforce development – an important factor in gender transformative AI – but just one dimension of what needs to be done) but rather is about getting support in the area of gender responsive products and services at all different stages of the AI life cycle.

## EDUCATE, PROMOTE, CONTEXTUALIZE

Create education and training opportunities for the workforce to understand the issues at play. This should be contextualized to the work being done by the company so it resonates more deeply. 'Give broad issues but then contextualize for your own work. Demonstrate how principles, and policies apply to their role, work, product'.[69] Axente suggests that an effective argument is that reducing bias reduces risk (which companies do not like) and can lead to important reputational gains and therefore increased competitiveness.

In the UNESCO Dialogue, Smith argued that the education of technical teams is required in order for them to design and implement AI systems that ensure gender equality. Braga also said that the technology sector needs a lot more education. According to her, 'people who are building these AI systems are trained engineers that were never made aware of [questions surrounding bias and gender equality] during their trainings and careers. Managers come from different fields.' She says: 'If you are an expert you don't go through an ethical training. If you're a manager, you did an MBA. Maybe you've heard about bias but don't know what to do about it.'

Axente also pointed out the need for 'a continuous educational process explaining why [gender equality in AI] is important'. She says: 'We need to raise awareness, make sure they understand the value of fairness, and gender parity.' In their field of AI, pymetrics is also of the opinion that it is 'crucial to educate the entire ecosystem around Human Resource technology on the importance of auditing talent selection for fairness.' Chow also pointed out that those writing AI principles within companies need to be adequately trained. She mentioned that at times, those in charge of drafting AI principles within a company do not fully understand how AI technology is developed and deployed. When this is the case, those drafting AI principles may not recognize the different types of AI technology that exist and could be deployed and may therefore narrowly focus their principles on one specific application of AI rather than on AI as a whole.

Mike McCawley explained that IBM makes time and creates incentives for staff to pursue training: each staff member needs to set aside a minimum of 30 hours per year for training. Accumulating training accomplishments is rewarded by giving staff opportunities to participate in new projects, and for promotions. According to McCawley, training is important for:

- Staff within the technology industry to build and design better AI;
- AI customers to understand how good the AI that they are using is;
- Government policymakers to get better at governing AI and developing policy; and
- The public at large to understand what AI is, how it works, and how it impacts their lives.

Educating and raising awareness could be done through:

- Technology companies providing ethics training for their staff;
- Affinity groups such as Women in Machine Learning and the Black in AI initiative that organize awareness raising events and push towards adequate inclusion of all women within the AI research ecosystem (Abebe); and
- UNESCO, which reaches the public at large through conferences, reports, videos and could

possibly develop curricula for schools and journalists about AI and about AI's impacts on gender equality.

## DIVERSE DEVELOPMENT TEAMS AND DESIGN PROCESS

'Privileged Ignorance: The vast majority of researchers, practitioners, and educators in the field are shielded or removed from the harm that can result in the use of AI systems leading to undervaluation, deprioritization, and ignorance of problems along with decontextualized solutions. The communities most likely to be harmed by AI systems are least likely to be involved in the teaching, design, development, deployment, and governance of AI; even when underrepresented individuals enter previously inaccessible spaces, we face existing practices, norms, and standards that require system-wide not just individual change.' Joy Buolamwini[70]

- Substantially increase and bring to positions of parity women coders, developers and decision-makers, with intersectionality in mind. This is not a matter of numbers, but also a matter of culture and power, with women actually having the ability to exert influence. See more in the discussion below on Women in Tech;
- Build multi-disciplinary teams, including gender experts, representatives of the social sciences and other domain experts;
- Institute residency programmes (engineers to visit non-profit organizations; end users, domain experts, non-profits to visit companies).

The multidisciplinary nature of AI was discussed by many Dialogue participants (Aseffa, Braga, Axente, Smith). Considering the many different fields of AI, such as dialogue systems, image processing, social media filtration, big data analysis, and perception synthesizers, hiring multidisciplinary teams seems to be a prerequisite. Braga illustrates this point. She came from the fields of linguistics and literature, and then shifted to engineering and voice design. She wrote her PhD about the need for multidisciplinarity in AI and then founded her company DefinedCrowd, which is in her words 'all about multidisciplinarity'.

According to Burnett, for Explainable AI to work in a truly effective manner, there is a need for 'integrating contributions from AI foundations with contributions from foundational human sciences'. In relation to gender equality, Smith argued that hiring diverse and multidisciplinary teams would help ensure that technology companies design and implement AI systems that integrate gender equality considerations. She explicitly referred to the 'need to integrate the knowledge of social scientists and gender experts into the teams developing, managing and using AI systems'. By mentioning 'gender experts', Smith makes the implicit case that hiring more women in technolgy is not sufficient because women are not de facto gender experts.

Keeping a human in the loop is also often recommended, in order to ensure there is human oversight.[71] But, it is not just about keeping 'a' human in the loop but about keeping the right human in the loop, one with empathy,[72] and one with an understanding of the need to account for gender equality. As noted by UNESCO however, 'even having a human "in the loop" to moderate a machine decision may not be sufficient to produce a "good" decision'.[73]

Auditing the design process and reflecting good practices in 'gendered innovations' is also essential. This refers to employing methods of sex and gender analysis as a resource to create new knowledge and stimulate novel design. Within technology, this applies to hardware and software components and is a lens through which to undertake all work. Saralyn Mark, Founder and President of iGiant, a non-profit accelerator for gender equal technologies, provides a seal of approval for companies that use a gender lens when designing technology. iGiant assesses the design process, rather than the resulting design itself. Companies can find commercial value in this. It may encourage them to become more committed to gender equality. Smith was also in favour of using auditing tools right from the beginning of the design process. She said: 'There are also some great qualitative tools that could be adapted to be gender specific – such as fairness checklists that should be completed before an AI system is developed and at certain points in the development process. These are currently quite broad, and could be adapted towards a justice and equality perspective for gender.'

While AI systems are composed of algorithms and training data contained in software, they can

also be integrated into hardware such as robots or smart speakers, like Amazon's Alexa. When AI hardware exists, it is important to closely examine its design, whether the user experience of people varies according to their sex or gender, and whether the design of the AI hardware participates in spreading harmful stereotypes about a specific gender. Tannenbaum et al. point out, for example, that robots can be gendered with their names, colour, appearance, voice, personality, and whether the robot is programmed to complete gender-stereotypical tasks (for example a female looking/sounding robot designed to clean).[74] In line with the analysis provided in *I'd Blush if I Could*, Gardner pointed out that: 'Robots are generally created to appear human, and pervasively female form, this stereotype must be removed, robots should not have a gender or failing that, not have a gender-specific role, for example robots as care assistants, reinforces that women should only have subservient roles.' In their guidelines for human-AI interaction design, Microsoft also tackles the issue of gender biases.[75]

## ETHICS PANELS AND IMPACT MAPPING

Companies engineering and/or implementing machine learning systems have a responsibility to map human rights risks before and during the life cycle of the product – from development to deployment and use.[76] High-risk areas should be subject to review.[77] The responsible technology think tank Doteveryone calls for Consequence Scanning Workshops during the feature planning stage to think about unintended consequences of features and plan mitigation strategies. UNESCO suggests the use of ethical impact assessments.[78] Similar calls have been made for holding ethics panels that review and assess work. Others suggest creating checklists (though this can risk a superficial approach). The AI Now Institute, an interdisciplinary research center dedicated to understanding the social implications of artificial intelligence, calls for rigorous testing in sensitive domains.[79]

In her contribution to UNESCO's Dialogue, Gardner states that she wishes to see the development of Algorithmic Impact Assessments (AIAs) as a more holistic way of regulating and auditing AI. In her view: 'A component of these AIAs would be a requirement for publishing the gender balance of the project team (and roles), whether data has been tested for gender bias and what methods were used to address this, [or] fairness tests of algorithmic outcomes using available tools such as [IBM's AI Fairness 360 toolkit] for example. Likewise, evaluating outcomes and impact on stakeholders [that are] subject to the outcome plus mitigations. Impact on the workforce and its skills, and appropriate mitigations, from a gender impact viewpoint, should also be listed. Governance is key, so if the AIA notes a significant area of risk then approval must be sought from the regulator(s) before deployment.'

As noted above, a level of ignorance and failure to check assumptions on gender can lead to incomplete risk assessments. A process for breaking down assumptions is therefore also essential. This requires the participation of gender equality and women's rights experts in panels and exercises (going beyond gender balance in teams).

## CONSULTATION WITH AFFECTED END USERS

Affected users should be involved to mitigate potential for unanticipated or unintended harms through community review boards.[80] Additionally, Gardner notes that the role and authority of an ethics panel and/or a citizens' panel should be required and publicized. These panels should have meaningful authority, including stating 'No Go'.

## TECHNICAL APPROACHES

In her contribution to UNESCO's Dialogue, Gardner emphasized that rigorous evaluations of AI were needed every step of the way: in design, training, testing, implementation/deployment and maintenance (i.e. post implementation). Participants in the UNESCO Dialogue mentioned a variety of existing and easily accessible tools, as well as tools currently being developed, to remove bias from AI datasets, algorithms, designs and design processes. Some also emphasized the importance of making these tools open-sourced through platforms like GitHub, so that as many people as possible can access and use them (although GitHub has its own history of gender bias[81]). For example, IBM's AI Fairness 360 toolkit is an open-sourced structured set of tools and algorithms designed for the technology community. The toolkit provides prescriptions (specific dance steps) to follow, that will reveal how fair or robust an AI system is. It enables people to find out whether their AI has sufficient

coverage or hidden biases, and thus seeks to increase AI fairness.

The following are existing or suggested tools to remove bias in AI systems:

- **Data checks**: Various standard methods are available to check the quality of AI data and training data.[82] Getnet Aseffa mentioned the need to use a 'gender detection software' in order to check whether data is balanced in terms of gender. He also said that the people cleaning and organizing data have to make sure that gender equality is taken into account.

- **Fairness audits may compare an AI's results to vetted standards found in the real world**. For example, in her contribution to UNESCO's Dialogue, Sara Kassir explained how pymetrics conducts fairness audits of their AI powered hiring tool (open-sourced on GitHub): 'When a hiring tool does create systematic disadvantage (i.e. favoring men over women), the consequence is known as "adverse impact". Pymetrics audits our models for evidence of adverse impact as a means of ensuring that we are sufficiently promoting gender equality in hiring. Our audits are conducted using a specific standard found in U.S. employment regulations called "the 4/5th rule." For example, assume a hiring assessment is used to evaluate 200 people, 100 men and 100 women. If the assessment "passes" 50 men, it must pass no less than 40 women, indicating an adverse impact ratio of no lower than 0.8 or 4/5ths.'

- **Counter-factual analyses examine the extent to which variables such as 'gender' impact the decision made by an AI.** Tannenbaum et al. explain this in the following way: 'Consider Google Search, in which men are five times more likely than women to be offered ads for high-paying executive jobs. The algorithm that decides which ad to show inputs features about the individual making the query and outputs a set of ads predicted to be relevant. The counterfactual would test the algorithm in silico[83] by changing the gender of each individual in the data and then studying how predictions change. If simply changing an individual from "woman" to "man" systematically leads to higher paying job ads, then the predictor is—indeed—biased.'[84]

- **Multi-accuracy auditing improves the transparency of the AI systems by quantifying how its performance varies across race, age, sex and intersections of these attributes.** Tannenbaum et al. describe this auditing technique as follows: 'In multi-accuracy, the goal is to ensure that the algorithm achieves good performance not only in the aggregate but also for specific subpopulations—for example, "elderly Asian man" or "Native American woman". The multi-accuracy auditor takes a complex machine-learning algorithm and systematically identifies whether the current algorithm makes more mistakes for any subpopulation. In a recent paper, the neural network used for facial recognition was audited and specific combinations of artificial neurons that responded to the images of darker-skinned women were identified that are responsible for the misclassifications.'[85]

- **Auditing user Interfaces for gender biases**: Burnett's research reveals that there are gender differences in how people interact with, use and learn from technology. According to her findings, men and women differ in their motivations (why they use technology), their information processing style, their self-confidence in their digital skills, their aversion to risk, and their learning style (whether they will enjoy tinkering to find a solution or whether they just want to complete a task quickly). Her research also reveals that if you make technology easier to use for people at the margins, you end up improving the user experience of everyone. She therefore developed the Gender Inclusiveness Magnifier, called GenderMag or GMAG, which is a method for identifying gender biases in user interfaces. It works for apps, websites and all kinds of software but can also be applied to AI software. GenderMag enables people to solve 'gender inclusiveness bugs' in their software or in the tools designed to help people create software.[86]

- **Other tools** exist to test the fairness of AI systems, such as Word Embedding Association Tests, which measure whether removing part of the training data affects the biases of a word embedding software.[87] Another way to increase the transparency and accountability of an AI system is to use White-box Automatic Machine Learning models, which are becoming

increasingly possible and available. White-box models are interpretable models: 'one can explain how they behave, how they produce predictions and what the influencing variables are'.[88] Finally, companies should carry out rigorous pre-release trials of their AI systems to check for biases.[89] Google for example uses 'adversarial testing' in an attempt to remove bias from its products before their launch.[90] In January 2020, Margi Murphy reported in *The Telegraph* that in order to prepare for the launch of its smart speaker and smartphone, Google asked its workers to stress-test the devices by repeatedly insulting them. Homophobia, racism and sexism were actively encouraged. Annie Jean-Baptiste, Head of Product Inclusion at Google explains that with this adversarial testing technique, they are 'trying to break the product before it launches.'[91] While such pre-release trials can reveal the more obvious cases of bias, they may not be able to expose implicit and unconscious biases in a systematic way. It is therefore important to carry them out concurrently with other auditing and testing techniques at all stages of the AI life cycle.

While a number of tools already exist to test the fairness of AI systems, more research is needed. In her Dialogue contribution, Christine Chow stressed that evidenced-based research about how to actually remove gender biases in AI is still missing. While she acknowledged that companies like Google and IBM had developed tools to help raise the issue, she argued that there was still a gap between our understanding of the issue and the availability of the tools to fix the issue.

## TRANSPARENCY

- **Data/Fact Sheets:** In order to increase the reliability and fairness of AI systems, scholars point to the necessity of having a standard document that defines 'what [data] needs to be collected at a minimum so stakeholders can make informed decisions'.[92] For example, Tannenbaum et al. note that 'researchers have designed "nutrition labels" to capture metadata about how the dataset was collected and annotated. Useful metadata should summarize statistics on, for example, the sex, gender, ethnicity and geographical location of the participants in the dataset. In many machine-learning studies, the training labels are collected through crowdsourcing, and it is also useful to provide metadata about the demographics of crowd labellers.'[93] This echoes the paper 'Datasheets for Datasets' by Timnit Gebru et al. in which they describe in detail what information should be contained in such 'datasheets'.[94]

- Creating **Model Cards** can also be done for model performance. These should reflect gender representativeness of data and any potential bias risks in models. IBM, for instance, envisions such documents to contain purpose, performance, safety, security, and provenance of information and to be completed by AI service providers for examination by consumers. Salesforce also creates model cards for customers and end users.

## TRACKING

- Tracking **where AI systems are used** and for what purposes.[95]

- **Auditing the auditing tools:** While some auditing tools exist, there are still no standardized processes or rules regarding AI auditing. Some questions require standardized answers. For example, who decides what needs to be audited? How should it be audited? By whom? And do we need to audit the auditing tools and processes? If so, how? Who should do this?

- **Maintenance auditing:** In her UNESCO Dialogue contribution, Gardner expressed the need to audit AI systems post implementation through regular maintenance checks. She suggests: 'it should be built into the maintenance schedule to periodically carry out regression testing using the test data (not training data) to ensure that algorithm weights are still providing the expected outcomes and no algorithm creep or distortion has occurred.'

## EXTERNAL MITIGATION

**Community Advisory Boards / Panels:** While these should be involved proactively (as noted above), they should also be involved in monitoring and accountability mechanisms. These may be independent of a particular company.

**Audits and Algorithmic Impact Assessments:** While these are undertaken by companies, they should also (preferably?) be undertaken by neutral third parties and in real time. Results of audits should be made available to the public

together with responses from the company.[96] The Algorithmic Justice League is one organization that will undertake such algorithmic audits.

## DIVERSITY IN AI INDUSTRY – WOMEN IN TECH

When asked about gender equality in AI, most UNESCO Dialogue participants started by discussing gender parity in the workplace and equal opportunities for men and women. All participants seem to agree that hiring more women in AI is a necessary – albeit not sufficient – step for greater gender equality in this field. On an individual level, Daniela Braga, the Founder and CEO of DefinedCrowd, is probably the one having achieved most success on this front, with her company having 42% of women employees.

Some, like Gardner, argue that hiring more women is not enough. The real objective is to make sure that women are hired in core roles such as development and coding: 'One thing that is missing is a requirement for diverse development teams (that are not just ethics-washing with women in peripheral roles but actually part of the development and coding).'

When discussing the underrepresentation of women in the field of AI, some pointed to the importance of going beyond the gender binary, in order to include those that identify as genderqueer or gender non-conforming, as well as those that intersect with multiple identities based on their race, geography, ability, sexual orientation, socio-economic status etc. For example, Abebe stated: 'We need adequate representation of women and genderqueer and non-binary individuals in positions where they can inform and guide the end-to-end cycle of AI.'

In their analysis of gender equality in AI, participants also went further than just looking at hiring practices in the field. For example, Braga, Smith and Axente pointed to how the culture of an organization can either empower or disenfranchise women. To describe work environments in the technology industry that are toxic for women, Smith used the word 'brotopia' – a term coined by Emily Chang in her 2018 book **Brotopia: Breaking Up the Boys' Club of Silicon Valley**. In a similar vein, the term 'brogrammer culture' is also used to describe a male-dominated and macho work environment that pushes women away from the technology industry. [97]Axente noted that solving this issue requires that we go beyond just bringing more women in the industry. In her opinion, technology companies should provide mentoring and coaching opportunities to make the culture more female friendly and create space for women's contributions. Ultimately, she argues that achieving gender equality is linked to changing power structures within organizations.

In order to rebalance unequal power structures and promote gender equality in the technology sector, Gardner points to positive action as a potential solution. Positive action are measures that can be lawfully taken to enable or encourage members of underrepresented groups to overcome their disadvantage, meet their needs or participate in an activity. Speaking on behalf of the Women Leading in AI network – a think thank that addresses gender biases in AI, Gardner declares: 'We want to emphasize that there is a difference between fairness and equality. We view equality not as equal treatment in the spirit of fairness but in battling the inequality and lack of representation [women] currently face within the field. For a future where there is true equality we must address the issue of the obstacles women face in entering and continuing within the field. In order to overcome [this], it may be that we need to inject positive action to redress the problems.'

Sara Kassir, on behalf of pymetrics, also made the point that de-biasing hiring practices is only part of the solution: 'Organizations that are committed to promoting gender equality need to recognize the importance of multi-pronged solutions. Fair AI evaluations are not enough; disparities in education, training, retention, and promotion need to be addressed, and inclusive cultures need to be fostered.'

There is a long and evolving body of work and lessons on increasing gender diversity in the workplace. This includes tools like the Women's Empowerment Principles, which were established by UN Global Compact and UN Women and offer guidance to businesses on how to promote gender equality in the workplace.[98] The National Center for Women in Technology also offers a plethora of research, tools and initiatives on this topic. These resources respond to a number of gaps identified by Dialogue participants and could be tailored to the unique needs of the AI industry.

Generally, the literature and practices on promoting women in business, and inclusion in the technology industry emphasize the following (illustrative):

### INCREASE FUNDING FOR:

- AI startups founded by women;
- Scholarships for women studying in the field of AI;
- Women AI researchers;
- Participation of women's groups and gender equality experts as affected AI users and stakeholders;
- Corporate Social Responsibility gender equality initiatives.

### HIGHEST LEVEL SUPPORT, GOALS AND POLICIES:

- Establish organizational strategies, policies and budgets that are gender transformative;
- Put clear management commitment, messaging and accountability structures in place;
- Address and correct for power asymmetries;[99]
- Eliminate unfair policies, creating gender equality related policies (e.g. equal pay, harassment, family leave/needs);
- Put in place transparent, independent and clear processes for remedial action.

### HIRING, RETENTION, PROMOTION TO DECISION-MAKING:

- Review hiring and promotion practices for bias;
- Address intersectionality;
- Review task assignments (who is given visibility and stretch goals?);
- Create mentoring and professional development programmes.

### SUPPLY CHAIN:

- Support women-owned businesses and gender responsive companies through supply chain management.

### CULTURAL CHANGE:

- Provide training on bias;
- Recognize that training on bias alone will not create transformative workplaces. Deeply seated norms and unconscious bias cannot be undone with training;
- Provide support for ally and affinity groups BUT without making the burden of change fall on them or individual women;
- Refuse to participate in, or hold 'manels' and 'manferences' – male-dominated panels and conferences – and advance women as organizational speakers and representatives;[100]
- Provide childcare at events.

### MEASURING AND REPORTING:

- Publish data on workforce positions, compensation and rates of promotions and actively address identified gaps.

## OTHER STAKEHOLDERS

Although the UNESCO Dialogue and this report are primarily focused on the private sector, achieving ethical and gender transformative AI unequivocally depends on the deep leadership and engagement of multiple stakeholder groups. A proper treatment of the roles and responsibilities of each stakeholder group and recommendations on actions is warranted and some of the literature has started to define these aspects (e.g. Buolamwini's 2019 testimony to US Congress on the role of the government vis-à-vis AI and the World Wide Web Foundation's 2018 recommendations to the G20 on gender and AI). The following merely highlights a few entry points.

### GENDER EQUALITY ADVOCATES / CIVIL SOCIETY

**Learn:** Develop opportunities for learning about AI and its implications for gender equality experts, practitioners and those working on girls and women's empowerment.

**Advocate:** Get involved in advocacy efforts around AI and ethics, including at the global and national level. Similarly, more actively reflect AI implications within the deliberations and work of gender equality bodies – e.g. the UN Commission on the Status of Women, the Beijing Platform for Action – and initiatives. Demand AI that is accountable to women – what it should look like must come from the gender equality community.

**Apply/Incorporate:** Support the ability of gender equality advocates, women's groups, and women themselves to apply and incorporate good AI technologies to support girls' and women's empowerment.

**Contribute and Monitor:** Make available gender equality experts, practitioners, and representatives from affected groups to participate in AI development and monitoring.

## GOVERNMENT

**Priorities and Practices:** Set priorities for AI development and application that respond to the needs of girls and women, and that contribute to social good more broadly. Ensure that government use of AI is not driving gender inequality and the perpetuation/exacerbation of bias and explore affirmative actions to counter bias.[101]

**Policy, Legislation and Regulation:** Develop appropriate policies, legislation and regulation that respond to gender equality concerns (e.g. require disclosure). According to Tannenbaum et al., to ensure that AI does not cause harm we need 'stringent review processes': 'Since the Second World War, medical research has been submitted to stringent review processes aimed at protecting participants from harm. AI, which has the potential to influence human life at scale, has yet to be so carefully examined.'[102] When explaining what legislative or supervisory bodies need to be put in place to regulate AI, UNESCO Dialogue participants drew parallels with the review processes of the pharmaceutical industry, the Food and Drugs Administration, and the role of the World Health Organization. Braga for example said there is a need for an ISO certification for AI, stating that companies that want to be trustworthy will be incentivized to obtain the certification. Any certification should explicitly address bias and harm around gender.

**Funding and Investment:** Devote public funding to support AI for women and women in AI, including through investment in research and development, scholarships and training, and allocating budgets within agencies/ministries (gender-responsive budgeting). Create an Accountability Fund to support critical research, provide redress for harm (e.g. through a government tax on AI firms) and explore the impacts of AI and ML on women.[103] Support investment in open and public models.

**Procurement, Hiring and Appointments, Requirements for Publicly Traded Companies:** Ensure that women are represented, including in decision-making positions, in government and on any of its advisory bodies and citizen panels. Review hiring, retaining, and promotion practices and other gender related policies. Utilize the power of the purse and government procurement to secure gender equality commitments and practices from business through Request for Proposal (RFP) processes and requirements for publicly traded companies. For example, Smith suggested that governments make it mandatory for companies that respond to their AI-related RFPs to have gender-specific AI regulation in place. In her words: 'Since governments are big users of AI tech for public purposes (e.g. education, health, etc.) it would be great to have mandates for completion of [gender-specific qualitative] tools in RFP proposals from companies.'

**Infrastructure and Education:** Build the digital data (e.g. collection of sex and gender-disaggregated data), and educational infrastructure that supports girls' and women's access to, use of, benefits from, and contributions towards, the digital society and AI.

## EDUCATION SYSTEM / ACADEMIA

**Cross-pollination of Ethics and Technology:** Review pedagogy and educational curricula for opportunities to better integrate AI, ethics, social sciences and gender equality studies.

**Research:** 'Promote deeper collaborations between AI researchers and organizations that work most closely with communities that are most harmed by algorithmic inequality. Fund university/community partnerships both to study AI harm on marginalized groups, and also to do participatory design of AI that is rooted in the needs of marginalized communities.'[104]

**Data Training Sets:** Develop gender responsive data training sets (e.g. Cornell University).

**Include AI-relevant programmes into public interest technology clinics** at degree granting institutions, similar to how law schools provide students with hands-on experience via public interest law clinics.[105] Such clinics would help produce graduates who will engage in public interest activities as AI professionals.

**Incorporate AI and ethics into STEM education** and computer science classes at the secondary school level and within **digital literacy** initiatives in formal and informal educational settings.

**Be forward looking** and incorporate future scenario practices and thinking.

# ACTION PLAN AND NEXT STEPS

**In the UNESCO Dialogue, Getnet Aseffa reminded us that once AI has boomed it will be too late to address gender equality issues. The window of opportunity is now. Therefore, it is recommended that a dedicated, cross-disciplinary initiative that can help catalyze deep thinking and action on gender equality and AI be swiftly created.**

## AWARENESS

It is crucial to generate awareness regarding the importance and urgency of these issues, what gender responsive AI would look like, and what is required to develop and deploy such AI systems.

- **Development of primers for different audiences** in accessible language with real world examples, and create briefs for the gender equality community on the implications of AI and how it impacts their work. Create briefs also for the AI industry on the dimensions of gender equality and how gender biases can potentially manifest themselves through AI. Eventually, briefs should be created at a more micro level, e.g. on access to justice, education, employment, etc.
- **Create Outreach Materials and Campaigns** to develop a shared sense of understanding and collective response, and create core reinforcing messages and other outreach materials. Launching creative campaigns to draw attention to these issues, including for the public, should be considered.
- **Visualization** - develop a map of the AI ecosystem and life cycle from a gender equality perspective, which will show 'at a glance' the different levels of challenges and opportunities, the actors, and the various levers of influence and action.
- **Engage influencers** - identify and recruit trusted influencers in key communities to broadcast messages and encourage engagement.

## FRAMEWORK

As AI ethics principles and frameworks are developed and implemented, it is critical that women and gender equality issues are represented.

- **Contribution to Broader AI and Ethics Dialogue**: Ensure that gender equality issues are continually fed into the process of developing AI ethics principles and frameworks, bringing in a more impactful approach than has been taken to date. This could be done by providing substantive recommendations and methodologies for adapting or contextualizing recommendations, and advocating for establishing a gender responsive process for their implementation.
- **Develop an Action Plan for gender equality and AI:** At the next level of detail, create a detailed Gender Equality and AI Action Plan that can inform a broader initiative and help guide different stakeholders. This could outline key priorities, timelines, stakeholder roles, existing and needed organizational expertise, and identify and secure funding.

## COALITION BUILDING

'The complexity of the ethical issues surrounding AI requires equally complex responses that necessitate the cooperation of multiple stakeholders across the various levels and sectors of the international, regional and national communities. Global cooperation on the ethics of AI and global inter-cultural dialogue are therefore indispensable for arriving at complex solutions.' UNESCO [106]

This statement holds equally true for multi-stakeholder cooperation on gender equality and AI. Many experts have called for the development of a broad coalition to address the layers and intricacies of these issues. Its features and functions could include:

- **Multi-stakeholder participation:** Include stakeholders from different sectors (academia, government, civil society, private sector), disciplines (technology, data, gender equality, ethics, human rights), and engage the full spectrum of those involved in gender equality from the women in the technology workforce, to the girls and women in STEM movements, to gender equality and women's empowerment experts and feminists, to those working on gender equality and women' and girls' empowerment in specific contexts (access to justice, economic empowerment, ending violence against women, education, health, etc.).

  In the UNESCO Dialogue, Chow was of the opinion that the AI industry needs stronger ties with academia. She said that 'academic collaborations' and 'more authentic and fundamental research' would enable the AI industry to address the root causes of gender biases in AI rather than just fixing symptoms. Similarly, Axente talked about the need to 'connect the dots' between the gender experts and the technologists. She also said that the AI industry is in need of both women data scientists and ethicists. She suggested UNESCO could have a crucial role to play in this respect thanks to its unique convening power.

  The need to 'connect the dots' echoes the findings of Collett and Dillon's 2019 report. In this report, the authors' first recommendation for future research in AI and Gender was to 'bridge gender theory and AI practice'.[107] They recognized the need for a dialogue between gender theorists and technologists, using a quote from Leavy: gender theory and AI practice 'are speaking completely different languages'.[108] The authors also make reference to Gina Neff from the University of Oxford: 'Gina Neff highlights this problem of the growing distance between those who are designing and deploying these systems, and those who are affected by these systems'.[109] In an attempt to bridge this gap, Gina Neff partnered with the Women's Forum for the Economy and Society to establish a new doctoral research opportunity in Gender and AI at the University of Oxford;[110]

- **Intergenerational Participation:** Engage young people, including at the secondary school level, as well as older adults to create intergenerational approaches, tapping into differing perspectives, thinking, and lived experiences. Young people who will be living in the future being created now should have an important voice around ethics and the purpose of AI;

- **Relationship Building and Dialogue:** Develop critical relationships and break down silos across groups and unpack and generate a baseline understanding of AI and gender equality. Creating a baseline understanding also means breaking down one's assumptions, acknowledging what you do not know, and identifying where collaboration with discipline experts should be ongoing and built into approaches. This is particularly necessary in order to embrace a global approach that includes the whole of society. While women in the technology workforce – and their allies – are critical protagonists and part of the solution, they should not be treated as proxies for gender equality experts, particularly when it comes to understanding structural issues and risks; [111]

- **Networking**: Create ongoing opportunities (online and through events/conferences) for conversation, interaction, collaboration, and information sharing;

- **Collective Impact and Action Oriented**: While developing an action plan and its implementation framework, seek ways to build collective action, to fill gaps in action, to connect to and leverage the work of others, and to join forces for greater impact. Fundamentally, a coalition should be action oriented, a 'think' and 'do' shop.

## CAPACITY BUILDING, TECHNICAL ASSISTANCE AND FUNDING

Practical mechanisms for translating principles into reality, to enable cooperation, and to drive concrete action are of highest priority. They should together form a package that enables stakeholders to make transformative rather than marginal or superficial change:[112]

- **Translators and Experts:** Explore the possibility of creating gender equality and AI 'translators' who are able to navigate both worlds and find common language between gender equality experts, ethicists and technologists. Explore the possibility of providing training and/or of creating a roster of technical advisors available to stakeholders on critical issues, policy and AI applications;
- **Resource Base and Tools:** Collect and further build a resource base on gender equality and AI. Where they do not exist, develop a suite of tools and guidance for the AI industry, as well as for civil society and gender equality advocates. Also, participate in external tool development bringing in a gender equality component;
- **Policy Dialogues:** Create or participate in policy dialogues on AI (e.g. through the work of the UN, the OECD or the EU) and on Gender Equality (e.g. through the UN Commission on the Status of Women);
- **Capacity Building:** Develop capacity building opportunities on gender and AI for different stakeholder groups, e.g. workshops or other projects that provide rich learning opportunities;
- **Seal:** Explore the creation of a 'Gender Equality AI Seal' that adopts a holistic system/society approach and considers process, content, expertise, funding, etc. Critically, this would provide monitoring and accountability mechanisms and be awarded based on effective action, not merely a pledge;
- **Fund:** Participate in the establishment and implementation of a fund that would support concrete and sustained action on gender equality and AI. Adequate resources are necessary for achieving these goals and should be budgeted at institutional/organizational levels, as well as for collaborative initiatives, and to support the participation of gender equality practitioners who are under-resourced.

## RESEARCH, MONITORING AND LEARNING

As society moves rapidly into new horizons with greater complexity and many unknowns, research, monitoring, learning and evaluation become ever more critical:

- **Learning from Recent Efforts:** What can we learn and apply from past and current gender and technology efforts? What has failed (over-emphasis on the 'pipeline',[113] piecemeal efforts, 'corporate feminism',[114] vague language, siloed efforts, hype and under-estimation of deeply rooted norms and power asymmetries, disinterest in gender on the side of technologists and for technology on the side of gender experts, etc.)? What efforts have led to success, and what is fundamentally different about AI?
- **Defining and Undertaking a Research Agenda:** Create partnerships to develop and undertake a robust research agenda.[115]
- **Monitoring and Evidence:** Conduct quantitative and qualitative analyses regarding the impacts of AI on gender equality, the status of funding and investment, the roles of women in decision-making in AI, and other metrics. Participate in, and hold others accountable for monitoring, transparency and reporting. Develop bottom-up feedback loops and more real-time monitoring possibilities;
- **Visibility:** Ensure that the implications of AI for gender equality are presented in real world cases and accessible languages for the public.

# SUMMARY OF RECOMMENDATIONS

## Adequately Frame the Overarching Landscape of AI and Society and the Imperatives of Gender Equality

Shift the narrative of AI as something 'external' or technologically deterministic, to something 'human' that is not happening to us but is created, directed, controlled by human beings and reflective of society. Establish a baseline understanding of AI, where and how it should be used for advancing a humanist framework and societal goals, and develop mechanisms for its governance. Similarly, establish a baseline understanding of the imperatives of gender equality and how AI as a socio-technical system can reinforce or challenge inequality.

## Significantly Strengthen Gender Equality in AI Principles

- Integrate feminist theory and frameworks rather than merely adding in 'women' as a target group. Treat gender equality as a way of thinking, a lens, an ethos and a constant – not a checklist.
- Ensure participation of gender equality advocates and practitioners, affected groups, and organizations working on specific discipline areas in principle development and implementation.
- Make gender equality more explicit and position gender equality principles in a way that provides for greater accountability for their prioritization and realization.
- Take systemic approaches that take into account society as a whole, the entirety of the AI ecosystem, and address structural gender equality.
- Unpack and substantiate what vague, undefined, or potentially contradictory principles mean in concrete terms and with consideration to intersectionality and compounded discrimination – particularly for principles that are implicit or without obvious gender connections but where there are still differentiated impacts.
- Ensure that gender equality is understood in terms of addressing harm, increasing the visibility of the gendered implications of AI, and encouraging AI's positive applications and beneficial impacts on women's empowerment.

## Operationalize Gender Equality and AI principles

- **Increase Awareness, Education and Skills in Society**
  - *Awareness and Participation*: Increase general awareness within society at large, the gender equality community, and the AI industry, regarding the positive and negative implications of AI for girls, women and gender non-binary people. Greater awareness should enhance public and stakeholder participation in AI governance.
  - *Contributors/Developers/Leaders:* Promote more girls and women in STEM and in AI education, careers and business. This includes creating lifelong learning and skill development opportunities for women in technology and AI, particularly for those displaced by automation.
  - *Personal Literacy:* Expand digital literacy to include AI and 'algorithmic awareness', with attention to the specific impacts for girls and women.*Use and Applications:* Advance women' and girls' access to and ability to use and apply AI to realize gender equality goals.

- **Industry Action**
  - *Advocacy for and commitment to human rights-based and gender responsive ethical frameworks* and support for the creation of industry standards and policies – including in cooperation with governments at the national level – that promote professional codes of conduct, that adhere to certain certifiable components, and that respond to gender equality criteria.
  - *Commitment to use of gender representative data*, mitigating data bias, and advancing data quality and rights.
  - *Engage with external and independent mitigation efforts* such as Community Advisory Boards and External Audits, and Algorithm Impact Assessments that reflect gender equality issues.
  - *Organizational Level Internal Mitigation* including putting in place contextualized and high-level corporate governance mechanisms, policies and incentives for gender responsive product development; building organizational capacity, creating diverse teams and processes (e.g. participation of affected users and gender experts, transparency cards/sheets) to mitigate bias and amplify benefits; and creating or using technical tools before, during and after product development that check for bias and harm.
  - *Ramp up gender diversity in the AI Industry* through good and evolving practices that address corporate policies, hiring, promotion and retention practices, organizational culture and power structures, educational and mentoring programmes, and monitoring and reporting.
- **Actions by Other Stakeholders**
  - *Gender Equality Advocates and Civil Society*: Increase the capacities and engagement of gender advocates and those working on girls and women's empowerment with AI. Scale up advocacy and policy development efforts (in AI *and* women's rights frameworks and bodies) promoting access to and application of AI for the benefit of girls and women, and critically, for monitoring AI's impacts.
  - *Government:* Commit to policies, regulations, and mechanisms (proactively and through redress) that promote gender equality in and through AI; encourage the development of AI applications that do not perpetuate bias or negatively impact girls and women but that rather respond to their needs and experiences; create funding mechanisms for participatory AI, access to AI and AI education for girls and women; promote diversity and equality through hiring, supply chain and related practices; and contribute to the collection of sex-disaggregated data.
  - *Academia and the Education System*: Develop curricula and pedagogy that better integrates cross-disciplinary social sciences, ethics and technology literacy and learning at the secondary and tertiary educational levels. Employ the research and development and other capacities of universities (e.g. public interest clinics) to address gender bias in data, algorithms and other elements of the AI ecosystem.

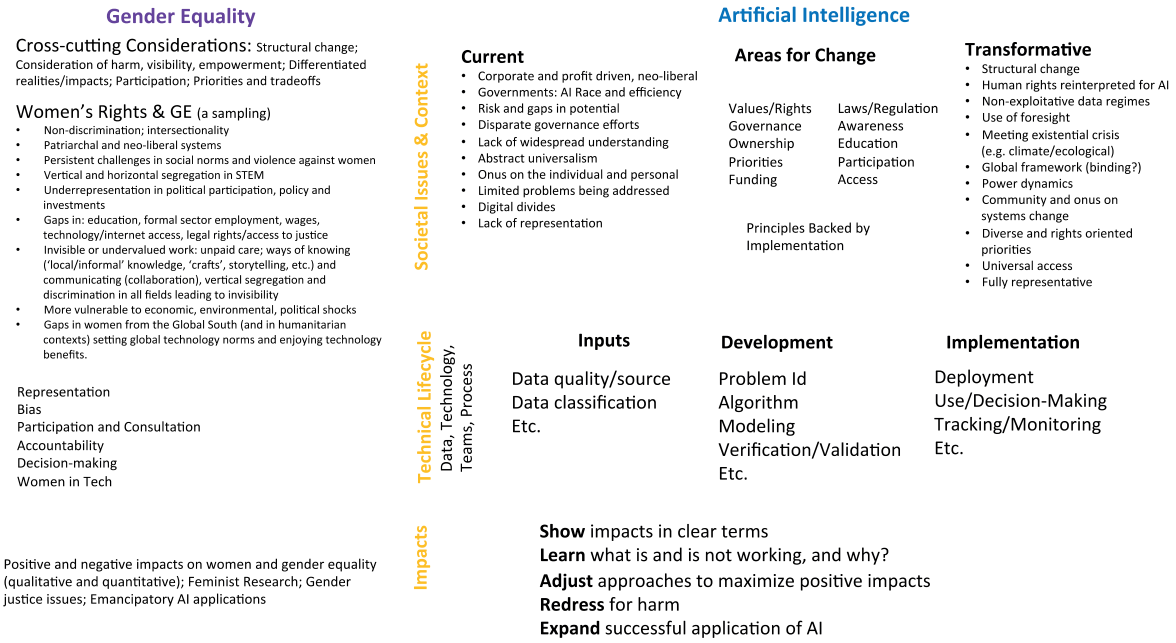
## Create Gender and AI Initiatives: Action Plan and Next Steps

- **Coalition Building**: Establish a multi-disciplinary and inter-generational coalition that builds partnerships across sectors and groups for a holistic society/AI ecosystem approach. Create avenues for dialogue and learning by creating a common understanding and language and through collaborations and collective impact models.
- **Framework Development:** Contribute gender perspectives to the broader development of AI and ethics frameworks and principles and their implementation. Additionally, create a comprehensive Action Plan on Gender Equality and AI that can inform initiatives and guide concrete stakeholder efforts.
- **Awareness Raising**: Develop materials that provide an overview and targeted analyses (e.g. impact areas or for different stakeholders), as well as channels for awareness raising, outreach and engagement.
- **Capacity Building, Technical Assistance and Funding:** Establish mechanisms that enable operationalization of gender and AI principles, including through the development of expertise, resources and tools, capacity building programmes, a fund that can support programming and participation, and explore

the development of a seal that provides criteria for gender equality in and through AI *with* accountability.

- **Research, Monitoring and Learning:** Define a research agenda for gender equality and AI, learn from past experiences – successes, gaps and failures – around gender equality and technology advocacy and initiatives; develop robust monitoring and learning mechanisms; and ensure that gender equality and AI work is publicly accessible and visible.[116]

The following is a 'food for thought' map of a more holistic approach to gender equality and AI.



# ANNEXES

## Annex 1: Explicit references to gender equality in existing AI ethics principles

Below are examples of explicit references to gender in selected texts (emphasis added in italics).

- The Union Network International (UNI) Global Union
  - Principle 3 **Make AI Serve People and Planet**: 'throughout their entire operational process, AI systems [to] remain compatible and increase the principles of human dignity, integrity, freedom, privacy and cultural and *gender diversity*, as well as ... fundamental human rights. In addition, AI systems must protect and even improve our planet's ecosystems and biodiversity.'[117]
  - Principle 4 **Ensure a Genderless, Unbiased AI:** 'In the design and maintenance of AI and artificial systems, it is vital that the system is controlled for negative or harmful human-bias, and that any bias—be it *gender*, race, sexual orientation or age, etc.—is identified and is not propagated by the system.'[118]

- The Organisation for Economic Co-operation and Development (OECD)
  - **Inclusive growth, sustainable development and well-being** (Principle 1.1): Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, *gender* and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.[119]

- The Institute of Electrical and Electronics Engineers (IEEE)
  - Principle 1 **Human Rights**: 'Human benefit is a crucial goal of Autonomous and intelligent systems (A/IS), as is respect for human rights set out in works including, but not limited to: (…) the *Convention on the Elimination of all forms of Discrimination against Women*'[120]
  - Implementing **Well-Being**: '"Well-being" will be defined differently by different groups affected by A/IS. The most relevant indicators of well-being may vary according to country, with concerns of wealthy nations being different than those of low- and middle-income countries. Indicators may vary based on geographical region or unique circumstances. The indicators may also be different across social groups, including *gender*, race, ethnicity, and disability status.'[121]
  - **Affective Computing**: 'Intimate systems must not be designed or deployed in ways that contribute to stereotypes, *gender* or racial inequality, or the exacerbation of human misery.'[122]
  - A/IS for **Sustainable Development**: 'The ethical imperative driving this chapter is that A/IS must be harnessed to benefit humanity, promote equality, and realize the world community's vision of a sustainable future and the SDGs: (...) of universal respect for human rights and human dignity, the rule of law, justice, equality and nondiscrimination;

of respect for race, ethnicity and cultural diversity; and of equal opportunity permitting the full realization of human potential and contributing to shared prosperity. A world which invests in its children and in which every child grows up free from violence and exploitation. A world in which every woman and girl enjoys full gender equality and all legal, social and economic barriers to their empowerment have been removed. A just, equitable, tolerant, open and socially inclusive world in which the needs of the most vulnerable are met.'[123]

- **Embedding Values** into A/IS: 'unanticipated or undetected biases should be further reduced by including members of diverse social groups in both the planning and evaluation of A/IS and integrating community outreach into the evaluation process (...). Behavioral scientists and members of the target populations will be particularly valuable when devising criterion tasks for system evaluation and assessing the success of evaluating the A/IS performance on those tasks. Such tasks would assess, for example, whether the A/IS apply norms in discriminatory ways to different races, ethnicities, *genders*, ages, body shapes, or to people who use wheelchairs or prosthetics, and so on.'[124]
- **Law**: Illustration on sentencing algorithm and bias case (Loomis) in which the defendant claimed that gender (higher risk as a male) was wrongly considered. 'The court reasoned that knowing the inputs and output of the tool, and having access to validating studies of the tool's accuracy, were sufficient to prevent infringement of Loomis' due process. Regarding the *use of gender*—a protected class in the United States—the court said he did not show that there was a reliance on gender in making the output or sentencing decision. Without the ability to interrogate the tool and *know how gender is used*, the court created a paradox with its opinion.'[125]

- Microsoft[126]
  - Fairness: Reduce unfairness rather than making it worse or maintaining status quo. This relates to technical components as well as the societal system in which it is deployed. Responding to socio-technological challenges and realities require greater diversity of AI development and deployment. Increasing fairness requires everyone to think about it, not delegate to others.[127]
  - Inclusiveness: Be intentionally inclusive and diverse. Designing for the 3% can at the same time reach the 97%. Reference to trans women and planning, testing, building with diverse groups.

- The European Commission (EC)
  - Introduction: 'In particular, AI systems can help to facilitate the achievement of the **UN's Sustainable Development Goals**, such as *promoting gender balance*.'[128]
  - (Glossary) **Vulnerable Persons and Groups**: 'What constitutes a vulnerable person or group is often context-specific. (…) factors linked to one's identity (such as *gender*, (...)) or other factors can play a role. The Charter of Fundamental Rights of the EU encompasses under Article 21 on non-discrimination the following grounds, which can be a reference point amongst others: namely *sex (…)* and sexual orientation.'[129]
  - 1.3 **Privacy and Data Protection**: 'Digital records of human behaviour may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, *gender, (…)*'[130]
  - 1.5 **Diversity, Non-discrimination and Fairness**: 'Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, *regardless of their (…) gender (…)*.'[131]
  - 2.1 **Fundamental rights** as a basis for Trustworthy AI: Equality, non-discrimination and solidarity – 'This also requires adequate *respect for potentially vulnerable persons and groups, such as (…) women, (…)*.'[132]
  - 2.2 Non-Technical Methods: **Diversity and Inclusive Design Teams** – 'Ideally, *teams are not only diverse in terms of gender*, culture, age, but also in terms of professional backgrounds and skill sets.'[133]

- The United Nations System Chief Executive Board for Coordination (CEB)
  - **Principle on Inclusiveness**: 'People and particularly those farthest behind*, including women and girls*, should be at the centre of all artificial intelligence-related capacity-building programming and decision-making processes.'[134]
  - 'All artificial intelligence-related **capacity-building programming** by United Nations entities should be *gender transformative. Gender* and age transformative approaches need to be embedded in all artificial intelligence-related capacity-building programming and decision-making processes. The particular effects of artificial intelligence on women and girls, and on the increasing *digital gender and age divide*, should also be taken into account.'[135]
  - In addition to broad principles, the UN CEB focuses on four needed levels of **capacity building**: infrastructure (specifically mentioning the *gender digital divide*); data; human capital and social capabilities (with specific attention to recruiting *girls and women* in STEM); and policy, law and human rights (with specific reference to human rights and *gender equality*).[136]
  - In its Road Map for Action, explicit reference to gender comes in a number of recommended commitments and corresponding measures:
  - 1.3. 'Develop templates and guidelines for **public-private investment** agreements that facilitate greater investments in Internet infrastructure, ensuring that the benefits of such investments are shared widely across society, with a particular focus on those groups that are most likely to be left behind, including *women and girls,* persons with disabilities, migrants and refugees, rural people and indigenous people.'[137]
  - 4.2. 'Promote and support more **inclusive multi-stakeholder participation** in both United Nations-convened and externally organized platforms and organizations related to artificial intelligence. In this regard, launch initiatives to lower the financial, knowledge, accessibility and social barriers to the effective participation of all stakeholders, with a focus on increasing participation from developing countries, as well as increased participation by *women and girls*.'[138]
  - 5.1. 'Build a **repository of artificial intelligence policy challenges and successes** from diverse stakeholders, including the various solutions tried and their impacts, especially those solutions that are focused on the bottom billion and on those at greatest risk of being left behind*, including women and girls.*'[139]
  - 6. 'Increase United Nations System and Member State capacity, particularly in developing countries, to collect, analyse and share open, interoperable **sex-disaggregated data sets**, as well as artificial intelligence tools to support both artificial intelligence innovation and the monitoring of the impacts of artificial intelligence.'[140]
  - 9. 'Maintain strong **ethical and human rights guardrails**, ensuring that artificial intelligence developments do not exceed the capacity to protect society, particularly marginalized, vulnerable and the poorest populations, *including women and girls*.'[141] With a measure, 9.4, that calls to 'Develop, building further on the existing efforts, policy and legal toolkits (with input from diverse stakeholders) that aim to ensure that artificial intelligence systems fully respect human and workers' rights, take into consideration local norms and ethics and do not contribute to, replicating or exacerbating biases including on the basis of *gender*, race, age and nationality, and in areas such as crime prevention.'[142]
- UNESCO and the World Commission on the Ethics of Scientific Knowledge and Technology (COMEST)
  - 'Respect for **human rights and fundamental freedoms** should be considered as one of the foundational values as development, deployment and uptake of AI technologies must occur in accordance with international human rights standards. (…) In order to ensure an inclusive AI, it is crucial that issues such as **discrimination and bias**, *including on the basis of gender*, as well as diversity, digital and knowledge divides are addressed. This is why leaving no one behind could be considered as another foundational value throughout the AI system lifecycle. Thus, the development and use of AI systems must be compatible

with maintaining **social and cultural diversity**, different value systems, take into consideration the specific needs of different age groups, persons with disabilities, *women and girls*, disadvantaged, marginalized and vulnerable populations and must not restrict the scope of lifestyle choices or personal experiences. This also raises concerns about neglecting local knowledge, cultural pluralism and the need to ensure equality. The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity. Particular attention must be paid to the lack of necessary technological infrastructure (…).'[143]

- *'Gender bias* should be avoided in the **development of algorithms, in the datasets used for their training, and in their use in decision-making**.'[144]
- **Value of Justice (Equality):** 'The value of justice is also related to non-discrimination. Roboticists should be sensitised to the *reproduction of gender bias and sexual stereotype in robots*. The issue of discrimination and stigmatisation through data mining collected by robots is not a trivial issue. Adequate measures need to be taken by States.'[145]
- In its summary of possible policy actions to guide reflection, the UNESCO highlights the following explicit recommendations with respect to gender:
- Educating about cost-benefit and inequalities: 'Working to **reduce digital divides**, including *gender divides*, in regard to AI access (…)'[146]
- **Practicing multi-stakeholder governance**: 'Ensuring *gender equality*, linguistic and regional diversity as well as the inclusion of youth and marginalized groups in multi-stakeholder ethical dialogues on AI issues.'[147]
- **Ensuring education**: '*Gender equitable AI and AI for gender equity*';[148] 'Increase artificial intelligence-related human capacity by supporting high quality and inclusive education, learning and training policies and programmes as well as reskilling and retraining of workers, including *women and girls*';[149] and 'Strengthen gender diversity in AI research both in academia and the private sector.'[150]
- Ensuring a **Gender Sensitive Approach**: 'Adopt sustained, varied and life-wide approaches; Establish incentives, targets and quotas; Embed ICT in formal education; Support engaging experiences; Emphasize meaningful use and tangible benefits; Encourage collaborative and peer learning; Create safe spaces and meet women where they are; Examine exclusionary practices and language; Recruit and train gender-sensitive teachers; Promote role models and mentors; Bring parents on board; Leverage community connections and recruit allies; Support technology autonomy and women's digital rights; Use universal service and access funds; Collect and use data, and set actionable indicators and targets.'[151]
- Fighting the **Digital Divide**: 'Work to reduce digital divides, including *gender divides*, in AI access, and establish mechanisms for continuous monitoring of the differences in access. Ensure that individuals, groups and countries that are least likely to have access to AI are active participants in multi-stakeholder dialogues on the digital divide by emphasizing the importance of *gender equality*, linguistic and regional diversity as well as the inclusion of youth and marginalized groups.'[152]

## Annex 2: Selected initiatives that address gender equality in AI

The following initiatives seek to address gender equality in AI and the ethics of AI. They are listed in alphabetical order.

### Gender and AI

- **Catalyst**: Catalyst focuses on creating workplaces that work for women and undertakes analysis on gender and AI. https://www.catalyst.org/
- **Coding Rights**: Coding Rights is an organization bringing an intersectional feminist approach to defend human rights in the development, regulation and use of technologies. The organization acts collectively and in networks, using creativity and hacker knowledge to question the present and reimagine a future based on transfeminist and decolonial values. Their mission is to expose and challenge technologies which reinforce power asymmetries, with focus on gender inequalities and its intersectionalities. https://www.codingrights.org/
- **Equal AI:** Equal AI is an initiative focused on correcting and preventing unconscious bias in the development of artificial intelligence. With leaders across business, technology, and academia, Equal AI is developing guidelines, standards and tools to ensure equal representation in the creation of AI. https://www.equalai.org/
- **Feminist AI:** Founded in 2016, Feminist. AI is a community AI research and design group focused on critical making as a response to hegemonic AI. Rather than simply criticize the lack of diversity in AI design and development, the group proposes an alternative by co-designing intelligent products, experiences and futures from a feminist post humanist approach. They do this by using AI Art and Design projects to create AI products, experiences and systems. https://www.feminist.ai/
- **Gendered Innovations, Stanford University:** The Gendered Innovations project harnesses the creative power of sex and gender analysis for innovation and discovery. Considering gender may add a valuable dimension to research, it may take research in new directions. The peer-reviewed Gendered Innovations project: 1) develops practical methods of sex and gender analysis for scientists and engineers and 2) provides case studies as concrete illustrations of how sex and gender analysis leads to innovation. https://genderedinnovations.stanford.edu/
- **The Center for Gender, Equity and Leadership (EGAL) of the Haas School of Business, UC Berkeley:** EGAL educates Equity Fluent Leaders in order for them to use their power to address barriers, increase access, and drive change for positive impact. EGAL recently published Mitigating Bias in Artificial Intelligence: an Equity Fluent Leadership Playbook. The Playbook provides a framework on how bias manifests in datasets and algorithms, breaking down concepts to remove the gaps between the technical side of the AI field and the business leaders who manage and govern it. After highlighting why bias in AI is a pervasive and critical problem, including the material impacts for businesses, the Playbook outlines evidence-based strategies ("plays") that business leaders can implement to holistically tackle the issue. https://haas.berkeley.edu/equity/
- **The Feminist Internet**: The Feminist Internet is a non-profit organization on a mission to make the internet a more equal space for women and other marginalized groups through creative, critical practice. Projects include creating F'xa, a feminist chatbot designed to provide the general public with a playful guide to AI bias. https://feministinternet.com/
- **Women in Big Data:** The goal of the women in big data forum is to strengthen diversity in the big data field. The aim is to encourage and attract more female talents to the big data and analytics field and help them connect, engage and grow. https://www.womeninbigdata.org/
- **Women in Data Science (WiDS):** WiDS is an initiative that aims to inspire and educate data scientists worldwide, regardless of gender, and to support women in the field. WiDS organizes conferences, datathons, podcast series, education outreach. https://www.widsconference.org/

- **Women in ML and Data Science**: Its mission is to support and promote women and gender minorities who are practicing, studying or interested in the fields of machine learning and data science. It creates opportunities for members to engage in technical and professional conversations in a positive, supportive environment by hosting talks by women and gender minority individuals working in data science or machine learning, as well as hosting technical workshops, networking events and hackathons. The network is inclusive to anyone who supports its cause regardless of gender identity or technical background. http://wimlds.org/

- **Women Leading in AI**: Women Leading in AI is a global 'think tank' for women in AI with the aim to address the bias that can occur within algorithms due to a lack of diversity and inclusivity within the field of Artificial Intelligence. It particularly focuses on the good governance of AI as a means to ensure that the changes in society likely to occur in the 4th industrial revolution will be of benefit to all people and not further embed societal prejudice in our systems. The network also promotes opportunities for women to support other women and men who support women in this field. https://womenleadinginai.org/

## AI, Ethics and Justice

- **ACM Fairness, Accountability and Transparency Conference (ACM FAccT)**: ACM FAccT is an interdisciplinary conference organized by the Association for Computing Machinery dedicated to bringing together a diverse community of scholars from computer science, law, social sciences, and humanities to investigate and tackle issues in this emerging area. It particularly seek to evaluate technical solutions with respect to existing problems, reflecting upon their benefits and risks; to address pivotal questions about economic incentive structures, perverse implications, distribution of power, and redistribution of welfare; and to ground research on fairness, accountability, and transparency in existing legal requirements. https://facctconference.org/

- **AI4All**: AI4ALL is a US-based nonprofit dedicated to increasing diversity and inclusion in AI education, research, development, and policy. AI4ALL organizes summer outreach program particularly for people of colour, young women and low-income high-schoolers, to learn about human-centered AI. https://ai-4-all.org/

- **AI, Ethics and Society Conference (AAAI/ ACM):** The Association for the Advancement of Artificial Intelligence (AAAI) and the Association for Computing Machinery (ACM) joined forces in 2018 to start the annual AAAI/ACM Conference on AI, Ethics, and Society. The aim of these conference is to provide a platform for multi-disciplinary participants to address the ethical concerns and challenges regarding issues such as privacy, safety and security, surveillance, inequality, data handling and bias, personal agency, power relations, effective modes of regulation, accountability, sanctions, and workforce displacement. https://www.aies-conference.com/2020/

- **AI Ethics Lab**: AI Ethics Lab aims to detect and solve ethical issues in building and using AI systems to enhance technology development. In research, the Lab functions as an independent center where multidisciplinary teams of philosophers, computer scientists, legal scholars, and other experts focus on analyzing ethical issues related to AI systems. Its teams work on various projects ranging from research ethics in AI to global guidelines in AI ethics. http://aiethicslab.com/big-picture/

- **AI For People:** AI For People gathers a diverse team of motivated individuals that is dedicated to bring AI Policy to the people, in order to create positive change in society with technology, through and for the public. Its mission is to learn, pose questions and take initiative on how AI technology can be used for the social good. It conducts impact analyses, projects and democratic policies that act at the crossing of AI and society. https://www.aiforpeople.org/

- **Alliance for Inclusive Algorithms**: Organized by two feminist organizations, Women@thetable and Ciudadanía Inteligente, the Alliance for Inclusive Algorithms calls on governments, the private sector, and civil society organizations to take proactive steps to remove biases from AI and increase the representation of women in the field of AI. https://aplusalliance.org/

- **Association for Progressive Communications (APC):** APC is an international network of civil society organizations founded in 1990 dedicated to empowering and supporting people working for peace, human rights, development and protection of the environment, through the strategic use of information and communication technologies (ICTs). Its aim is to build a world in which all people have easy, equal and affordable access to the creative potential of ICTs to improve their lives and create more democratic and egalitarian societies. https://www.apc.org
- **Black in AI:** Black in AI (BAI) is a multi-institutional, transcontinental initiative creating a space for sharing ideas, fostering collaborations, and discussing initiatives to increase the presence of Black individuals in the field of AI. To this end, BAI holds an annual technical workshop series, run mentoring programs, and maintain various forums for fostering partnerships and collaborations. https://blackinai.github.io/
- **Data & Society:** Data & Society studies the social implications of data-centric technologies and automation. It produces original research on topics including AI and automation, the impacts of technology on labor and health, and online disinformation. https://datasociety.net/
- **EQUALS**: The EQUALS Global Partnership for Gender Equality in the Digital Age is a committed group of corporate leaders, governments, businesses, not-for-profit organizations, academic institutions, NGOs and community groups around the world dedicated to promoting gender balance in the technology sector by championing equality of access, skills development and career opportunities for women and men alike. https://www.equals.org/
- **Generation AI, UNICEF:** UNICEF launched Generation AI with a diverse set of partners, including the World Economic Forum, UC Berkeley, Article One, Microsoft and others to set and lead the global agenda on AI and children - outlining the opportunities and challenges, as well as engaging stakeholders to build AI powered solutions that help realize and uphold child rights. https://www.unicef.org/innovation/GenerationAI
- **Global Partnership on Artificial Intelligence:** The OECD will host the Secretariat of the new Global Partnership on AI (GPAI), a coalition that aims at ensuring that Artificial Intelligence is used responsibly, respecting human rights and democratic values. The GPAI brings together experts from industry, government, civil society and academia to conduct research and pilot projects on AI. Its objective is to bridge the gap between theory and practice on AI policy. GPAI will create a strong link between international policy development and technical discourse on AI. http://www.oecd.org/going-digital/ai/OECD-to-host-Secretariat-of-new-Global-Partnership-on-Artificial-Intelligence.htm
- **IT for Change:** IT for Change is an NGO based in Bengaluru, India. It works in the areas of education, gender, governance, community informatics and internet/digital policies to push the boundaries of existing vocabulary and practice, exploring new development and social change frameworks. https://itforchange.net/
- **Leverhulme Centre for the Future of Intelligence**: The Leverhulme Centre for the Future of Intelligence (CFI) builds a new interdisciplinary community of researchers, with strong links to technologists and the policy world, and a clear practical goal: to work together to ensure that humans make the best of the opportunities of artificial intelligence as it develops over coming decades. http://lcfi.ac.uk/
- **OpenAI:** OpenAI is an AI research and deployment company based in San Francisco, California. Its mission is to ensure that artificial general intelligence benefits all of humanity. The OpenAI Charter describes the principles that guide its work. https://openai.com/about/
- **Oxford Internet Institute:** The Oxford Internet Institute is a multidisciplinary research and teaching department of the University of Oxford, dedicated to the social science of the Internet. Students and staff examine gender and AI in projects, courses and research. https://www.oii.ox.ac.uk/

- **Partnership on AI:** The Partnership on AI conducts research, organizes discussions, shares insights, provides thought leadership, consults with relevant third parties, responds to questions from the public and media, and creates educational material that advances the understanding of AI technologies including machine perception, learning, and automated reasoning. It recently initiated "Closing Gaps in Responsible AI" – a multiphase, multi-stakeholder project aimed at surfacing the collective wisdom of the community to identify salient challenges and evaluate potential solutions to operationalizing AI Principles. These insights can in turn inform and empower the change makers, activists, and policymakers working to develop and manifest responsible AI. https://www.partnershiponai.org/about/
- **Teens in AI:** The Teens In AI initiative exists to inspire the next generation of AI researchers, entrepreneurs and leaders who will shape the world of tomorrow. It aims to give young people early exposure to AI being developed and deployed for social good. Through a combination of hackathons, accelerators, and bootcamps, together with expert mentoring, talks, company tours, workshops, and networking opportunities, the program creates the platform for young people aged 12-18 to explore AI, machine learning, and data science. https://www.teensinai.com/
- **The AI Now Institute:** The AI Now Institute at New York University is an interdisciplinary research center dedicated to understanding the social implications of artificial intelligence. Their work focuses on four core domains: rights and liberties; labor and automation; bias and inclusion; safety and critical infrastructure. They have published numerous reports and held discussions on gender equality and bias. Their new initiative 'Data Genesis' is developing new approaches to study and understand the role of training data in the machine learning field. https://ainowinstitute.org/
- **The AI for Good Global Summit**: The AI for Good Global Summit is a leading action-oriented, global and inclusive United Nations platform on AI. The Summit is organized by the ITU with XPRIZE Foundation, in partnership with UN Sister Agencies, Switzerland and ACM. https://aiforgood.itu.int/
- **The Algorithmic Justice League**: The Algorithmic Justice League's mission is to raise awareness about the impacts of AI, equip advocates with empirical research, build the voice and choice of the most impacted communities, and galvanize researchers, policy makers, and industry practitioners to mitigate AI harms and biases. We're building a movement to shift the AI ecosystem towards equitable and accountable AI. https://www.ajlunited.org/
- **The Institute of Electrical and Electronics Engineers' (IEEE) Global Initiative**: The IEEE Global Initiative's mission is 'to ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.' https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

# REFERENCES


1 Brussevich, M., Dabla-Norris, E. and Khalid, S. 2019. Is Technology Widening the Gender Gap? Automation and the Future of Female Employment. *IMF Working Papers*, Working Paper No. 19/91. Washington, DC: International Monetary Fund.

2 Hegewisch, A., Childers, C. and Hartmann, H. 2019. *Women, Automation and the Future of Work*. Washington, DC: The Institute for Women's Policy Research.

3 Buolamwini, J. 2019. Hearing on: Artificial Intelligence: Societal and Ethical Implications. Washington, DC: United States House Committee on Science, Space and Technology. Joy Adowaa Buolamwini is a Ghanaian-American computer scientist and digital activist based at the MIT Media Lab. Her 2018 paper *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* uncovered large racial and gender bias in AI facial recognition systems from companies like Microsoft, IBM, and Amazon. Subsequently, she founded the Algorithmic Justice League, an organization that looks to challenge bias in decision-making software.

4 For more information about Google's AI for Social Good programme visit https://ai.google/social-good/.

5 Gurumurthy, A. and Chami, N. 2019. Radicalising the AI governance agenda. *Global Information Society Watch 2019: Artificial intelligence: Human rights, social justice and development*. Johannesburg, South Africa: APC.

6 Peña, P. and Varon, J. 2019. Decolonising AI: A transfeminist approach to data and social justice. *Global Information Society Watch 2019: Artificial intelligence: Human rights, social justice and development*. Johannesburg, South Africa: APC.

7 Kwet, M. 2019. Digital colonialism is threatening the Global South. Aljazeera, 13 March 2019.

8 ORBIT. 2019. 100 *Brilliant Women in AI and Ethics: Conference Report*. Leicester, UK: Observatory for Responsible Research and Innovation in ICT.
Collett, C. and Dillon, S. 2019. *AI and Gender: Four Proposals for Future Research*. Cambridge, UK: The Leverhulme Centre for the Future of Intelligence.
GIS Watch. 2019. *Global Information Society Watch 2019: Artificial intelligence: Human rights, social justice and development*. Johannesburg, South Africa: APC.

9 Buolamwini, op. cit.
ORBIT, op. cit.

10 Chui, M. et al. 2018. Notes from the AI Frontier: Applying AI for Social Good. *McKinsey Global Institute Discussion Paper,* December 2018. Washington, DC: McKinsey Global Institute.

11 See GIS Watch 2019 for a discussion on radical AI governance.

12 CEDAW: https://www.ohchr.org/en/hrbodies/cedaw/pages/cedawindex.aspx
UN Commission on the Status of Women: https://www.unwomen.org/en/csw

13 SDGs and Women: https://www.unwomen.org/en/news/in-focus/women-and-the-sdgs and https://sdgs.un.org/goals/goal5.

14 Intersectionality refers to the complex, cumulative way in which the effects of multiple forms of discrimination (such as sexism, racism, and classism) combine, overlap, or intersect especially in the experiences of marginalized individuals or groups.

15 In the 2019 GIS Watch report, P. Peña and J. Varon refer to 'transfeminism' as an epistemological tool that aims to re-politicise and de-essentialise global feminist movements that have been used to legitimise policies of exclusion on the basis of gender, migration, miscegenation, race and class.

16 For example, women make up only 17% of the biographies on Wikipedia and 15% of its editors, and there is evidence of gender bias in the language of Wikipedia entries (Wikipedia - https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia), 26% of subjects and sources in mainstream Internet news stories are women (Who Makes the News - http://whomakesthenews.org/gmmp-campaign), there are significant gender gaps across the stages of academic publishing, citation and comment (NIH - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7112170/), and the rather dismal figures go on, particularly looking at historical data. Women's access to and different use of ICT (phones, internet access, digital literacy, etc.) is another significant factor for the representativeness of data. In 2017 there were 250 million fewer women online (EQUALS - https://www.equals.org/post/2018/10/17/beyond-increasing-and-deepening-basic-access-to-ict-for-women).

17 The digital gender gap or divide refers to the differences between men and women and between girls and boys in their ability to access and use digital technologies and participate fully in the online world. Worldwide, women and girls are disproportionately affected by the lack of access to information and communication technologies (ICTs) and lack of digital skills to use them effectively. For more information about the growing digital gender divide see the 2019 UNESCO report *I'd Blush if I Could: closing gender divides in digital skills through education*.

18 Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Cambridge, MA: Berkman Klein Center for Internet & Society.
Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol. 1, pp. 389–399.
AI Ethics Lab. 2020. A Guide to the Dynamics of AI Principles: Our Toolbox for Understanding & Using AI Principles. February 2020.
UNESCO. 2020. Outcome Document: First Version of a Draft Text of a Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/AHEG-AI/2020/4REV. Paris: UNESCO.

19 UNESCO, 2020, Outcome Document, op. cit.

20 The World Commission on Ethics of Scientific Knowledge and Technology (COMEST) is a multidisciplinary scientific advisory body of UNESCO, made up of independent experts. The work of COMEST has built on and complemented work on AI being done within the United Nations system, other international organizations, nongovernmental organizations, academia and others.

21 UNESCO. 2020. Working Document: Toward a Draft Text of a Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/AHEG-AI/2020/3REV. Paris: UNESCO.

22 Amrute, S. 2019. Of Techno-Ethics and Techno-Affects. *Feminist Review*, Vol. 123, No. 1, pp. 56–73.

23 GIS Watch, op. cit.
Amrute, op. cit., p. 70.

24 Ibid., p. 70.

25 Ibid., p. 59.

26 Ibid., p. 57.

27 Ibid., p. 58.

28 Ibid., p. 57.

29 See https://feministinternet.org/

30 E.g.: Bellamy, R. K. E., et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. IBM Journal of Research and Development, Vol. 63, No. 4/5.
Weinberger, D. 2019. Where fairness ends. *Harvard Business School Digital Initiative*, 26 March.

31 Fjeld et al., op. cit., p. 47.

32 Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. and Schiebinger, L. 2019. Sex and gender analysis improves science and engineering. *Nature*, Vol. 575, pp. 137–146.

33 Collett and Dillon, op. cit., p. 23.

34 Tannenbaum et al., op. cit.

35 Fjeld et al., op. cit., p. 50.

36 Ibid., p. 51.

37 Ibid., p. 51.

38 Ibid., p. 52.

39 Ibid., p. 60.

40 Ibid., p. 61.

41 Ibid., p. 64.

42 Jobin et al., op. cit., p. 394.

43 Ibid., p.395.

44 Ibid.

45 Ibid., p. 396.

46 Tannenbaum et al., op. cit., p. 141.

47 UNESCO, 2020, Working document, op. cit.

48 ORBIT, op. cit.

49 Fjeld et al., op. cit., p. 48.

50 Ibid., p. 66.

51 Buolamwini, op. cit.

52 GIS Watch, op. cit.
European Commission High-Level Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI*. Brussels, Belgium: European Commission.
Jobin et al., op. cit.
UNESCO, 2020, Working document, op. cit.

53 Tannenbaum et al., op. cit., p. 4.

54 Buolamwini, op. cit.

55 ORBIT, op. cit.

56 UNESCO. 2019. Preliminary Study on a Possible Standard-Setting Instrument on the Ethics of Artificial Intelligence, 40 C/67. Paris: UNESCO. p. 15.

57 Gender segregation refers to the unequal distribution of men and women in the occupational structure. 'Vertical segregation' describes the clustering of men at the top of occupational hierarchies and of women at the bottom; 'horizontal segregation' describes the fact that at the same occupational level (that is within occupational classes, or even occupations themselves) men and women have different job tasks. See: Gender Segregation (in employment). A Dictionary of Sociology. *Encyclopedia.com*. 11 August 2020. https://www.encyclopedia.com/social-sciences/dictionaries-thesauruses-pictures-and-press-releases/gender-segregation-employment

58 Mozilla Foundation. 2020. Internet Health: Trustworthy Artificial Intelligence.

59 Buolamwini, op. cit.

60 Collett and Dillon, op. cit.

61 World Economic Forum Global Future Council on Human Rights 2016-2018. 2018. How to Prevent Discriminatory Outcomes in Machine Learning. *White Paper*, March 2018. Geneva: World Economic Forum.

62 Collett and Dillon, op. cit.

63 World Economic Forum Global Future Council on Human Rights 2016-2018, op. cit.

64 ORBIT, op. cit.

65 World Economic Forum Global Future Council on Human Rights 2016-2018, op. cit.

66 Ibid.

67 Baxter, K. 2020. AI Ethics at Salesforce. *Women in Big Data Webinar on AI and Ethics*, 21 May.

68 Ibid.

69 Ibid.

70 Buolamwini, op. cit.

71 ORBIT, op. cit.

72 Herrera, C. 2020. TIBCO: The good, the bad, and the human sides of AI. *Women in Big Data Webinar on AI and Ethics*, 21 May.

73 UNESCO, 2019, op. cit., p. 14.

74 Tannenbaum et al., op. cit.

75 Amershi, S. et al. 2019. Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, Paper 3, pp. 1–13.

76 World Economic Forum Global Future Council on Human Rights 2016-2018, op. cit.

77 Baxter, op. cit.

78 UNESCO, 2020, Working document, op. cit.

79 West, S. M., Whittaker, M. and Crawford, K. 2019. *Discriminating Systems: Gender, Race and Power in AI*. New York: AI Now Institute.

80 Buolamwini, op. cit.
World Economic Forum Global Future Council on Human Rights 2016-2018, op. cit.
AI Ethics Lab, op. cit.

81 Wong, J. C. 2016. Women considered better coders – but only if they hide their gender. *The Guardian*, 12 February.

82 E.g. Pang, W. 2019. How to Ensure Data Quality for AI. *insideBIGDATA*, 17 November.

83 In silico (pseudo-Latin for 'in silicon', alluding to the mass use of silicon for computer chips) is an expression meaning 'performed on computer or via computer simulation'.

84 Tannenbaum et al., op. cit., p. 5.

85 Ibid.

86 Burnett, M. 2020. Doing Inclusive Design: From GenderMag to InclusiveMag. Human-Computer Interaction Institute Seminar Series. Pittsburgh, PA: Carnegie Mellon University.

87 Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D. M., Kalai, A. T. 2019. What are the biases in my word embedding? Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), Honolulu, HI, USA, 2019.

88 Sciforce. 2020. Introduction to the White-Box AI: the Concept of Interpretability. *Medium*, 31 January.

89 West et al., op. cit.

90 Murphy, M. 2020. Inside Google's struggle to control its 'racist' and 'sexist' AI. *The Telegraph*, 16 January.

91 Ibid.

92 Buolamwini, op. cit.

93 Tannenbaum et al., op. cit., p. 5.

94 Gebru, T. et al. 2018. Datasheets for Datasets. Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, 2018.

95 West et al., op. cit.

96 World Economic Forum Global Future Council on Human Rights 2016-2018, op. cit.

97 Litt, M. 2018. I accidentally built a brogrammer culture. Now we're undoing it. *Fast Company*, 2 February.

98 See https://www.weps.org/.

99 West et al., op. cit.

100 Else, H. 2019. How to banish manels and manferences from scientific meetings. Nature, Vol. 573, pp. 184–186.

101 Avila, R., Brandusescu, A., Ortiz Freuler, J., Thakur, D. 2018. *Policy Brief W20 Argentina: Artificial Intelligence: Open Questions about Gender Inclusion*. Geneva: World Wide Web Foundation.

102 Tannenbaum et al., op. cit., p. 4.

103 Buolamwini, op. cit.t.
Avila et al., op. cit.

104 Buolamwini, op. cit.

105 Ibid.

106 UNESCO, 2020, Working document, op. cit.

107 Collett and Dillon, op. cit., p. 8.

108 Ibid., p. 10.

109 Ibid.

110 See https://www.oii.ox.ac.uk/blog/doctoral-research-opportunity-in-gender-and-ai-applications-open/

111 To be avoided is a scenario where a woman in technology is treated as a spokesperson and source for all issues related to gender. The technology sector needs to dig deeper. Similarly, gender equality practitioners cannot outsource conversations on technology to women in the technology workforce and those working to advance women in STEM. Rather, early on, they must play a fundamental role in shaping understanding on gender and become very conversant themselves on the implications of AI.

112 Learning from current efforts on gender and ICT, we see that piecemeal approaches, simplistic checklists, one-off trainings, or efforts lacking accountability, have been insufficient in creating meaningful change.

113 The pipeline argument postulates that the gross underrepresentation of women in the technology sector is due to a low female pool of talent in STEM fields. This argument tends to ignore the fact that the underrepresentation of women in the technology sector is also due to technology companies failing to attract and retain female talent.

114 The term 'corporate feminism' is used by some to refer to a specific brand of feminism that encourages assimilation into the corporate mainstream rather than a complete deconstruction (or at least re-thinking) of the system as a whole. It puts the onus on the individual to break the glass ceiling, rather than addressing systemic mechanisms of oppression (based on class, sexual orientation, race, religion, ability, etc). Sheryl Sandberg, author of the book *Lean In: Women, Work and the Will to Lead,* is often cited as one of the figureheads of corporate feminism.

115 E.g. Collett and Dillon, op. cit., for some questions that have already been asked.

116 Ibid.

117 UNI Global Union. 2017. *Top 10 Principles For Ethical Artificial Intelligence*. Nyon, Switzerland: UNI Global Union. p. 7.

118 Ibid., p. 8.

119 OECD. 2020. Recommendation of the Council on Artificial Intelligence. *OECD Legal Instruments*, OECD/LEGAL/0449. Paris: OECD. p.7.

120 IEEE. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. First Edition. New York: IEEE. p. 19.

121 Ibid., p. 82.

122 Ibid., p. 96.

123 Ibid., p. 140.

124 Ibid., p. 188.

125 Ibid., p. 241–242.

126 Microsoft. 2020. Microsoft AI Principles.

127 While gender and women were not explicitly mentioned, there were images of women provided in the video discussion of this principle.

128 European Commission High-Level Group on Artificial Intelligence, op. cit., p. 4.

129 Ibid., p. 38.

130 Ibid., p. 17.

131 Ibid., p. 18.

132 Ibid., p. 11.

133 Ibid., p. 23.

134 United Nations System Chief Executive Board for Coordination (CEB). 2019. A United Nations system-wide strategic approach and road map for supporting artificial intelligence capacity development on artificial intelligence. CEB/2019/1/Add.3. Geneva: CEB. p. 3.

135 Ibid.

136 Ibid., p. 5.

137 Ibid., p. 7.

138 Ibid., p. 9.

139 Ibid.

140 Ibid.

141 Ibid., p. 11.

142 Ibid.

143 UNESCO, 2020, Working Document, op. cit., p. 11.

144 COMEST. 2019. Preliminary study on the Ethics of Artificial Intelligence. SHS/COMEST/EXTWG-ETHICS-AI/2019/1. Paris: UNESCO. p. 25.

145 COMEST. 2017. Report of COMEST on Robotics Ethics. SHS/YES/COMEST-10/17/2 REV. Paris: UNESCO. p. 52.

146 UNESCO, 2020, Working Document, op. cit., p. 39.

147 Ibid., p. 40.

148 Ibid., p. 42.

149 Ibid., p. 43.

150 Ibid., p. 44.

151 UNESCO and EQUALS. 2019. *I'd Blush if I Could: Closing Gender Divides in Digital Skills Through Education*. Geneva: EQUALS. p. 9.

152 UNESCO, 2020, Working Document, op. cit., p. 48.

# BIBLIOGRAPHY


AI Ethics Lab. 2020. A Guide to the Dynamics of AI Principles: Our Toolbox for Understanding & Using AI Principles. https://aiethicslab.com/big-picture/

Amershi, S. et al. 2019. Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, Paper 3, pp. 1–13. DOI: 10.1145/3290605.3300233

Amrute, S. 2019. Of Techno-Ethics and Techno-Affects. Feminist Review, Vol. 123, No. 1, pp. 56–73. DOI: 10.1177/0141778919879744.

Avila, R., Brandusescu, A., Ortiz Freuler, J., Thakur, D. 2018. Policy Brief W20 Argentina: *Artificial Intelligence: Open Questions about Gender Inclusion*. Geneva: World Wide Web Foundation. http://webfoundation.org/docs/2018/06/AI-Gender.pdf

Baxter, K. 2020. AI Ethics at Salesforce. Women in Big Data Webinar on AI and Ethics, 21 May. https://www.youtube.com/watch?v=qLFIfHoRe_E

Bellamy, R. K. E., et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *IBM Journal of Research and Development*, Vol. 63, No. 4/5. https://arxiv.org/abs/1810.01943

Brussevich, M., Dabla-Norris, E. and Khalid, S. 2019. Is Technology Widening the Gender Gap? Automation and the Future of Female Employment. *IMF Working Papers*, Working Paper No. 19/91. Washington, DC: International Monetary Fund. https://www.imf.org/en/Publications/WP/Issues/2019/05/06/Is-Technology-Widening-the-Gender-Gap-Automation-and-the-Future-of-Female-Employment-46684.

Buolamwini, J. 2019. Hearing on: Artificial Intelligence: Societal and Ethical Implications. Washington, DC: United States House Committee on Science, Space and Technology. https://science.house.gov/imo/media/doc/Buolamwini%20Testimony.pdf

Burnett, M. 2020. Doing Inclusive Design: From GenderMag to InclusiveMag. Human-Computer Interaction Institute Seminar Series. Pittsburgh, PA: Carnegie Mellon University. https://scs.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1fa67161-215d-45e8-b228-ab3e00f844ce

Chui, M. et al. 2018. Notes from the AI Frontier: Applying AI for Social Good. *McKinsey Global Institute Discussion Paper*, December 2018. Washington, DC: McKinsey Global Institute. https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good#

Collett, C. and Dillon, S. 2019. *AI and Gender: Four Proposals for Future Research*. Cambridge, UK: The Leverhulme Centre for the Future of Intelligence. https://www.repository.cam.ac.uk/handle/1810/294360

COMEST. 2017. Report of COMEST on Robotics Ethics. SHS/YES/COMEST-10/17/2 REV. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000253952

COMEST. 2019. Preliminary study on the Ethics of Artificial Intelligence. SHS/COMEST/EXTWG-ETHICS-AI/2019/1. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367823

Else, H. 2019. How to banish manels and manferences from scientific meetings. Nature, Vol. 573, pp. 184–186. https://www.nature.com/articles/d41586-019-02658-6

European Commission High-Level Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI*. Brussels, Belgium: European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Cambridge, MA: Berkman Klein Center for Internet & Society. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

Gebru, T. et al. 2018. Datasheets for Datasets. Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, 2018. https://arxiv.org/abs/1803.09010

GIS Watch. 2019. *Global Information Society Watch 2019: Artificial intelligence: Human rights, social justice and development*. Johannesburg, South Africa: APC. https://www.apc.org/en/pubs/global-information-society-watch-2019-artificial-intelligence-human-rights-social-justice-and

Gurumurthy, A. and Chami, N. 2019. Radicalising the AI governance agenda. *Global Information Society Watch 2019: Artificial intelligence: Human rights, social justice and development*. Johannesburg, South Africa: APC. https://www.giswatch.org/2019-artificial-intelligence-human-rights-social-justice-and-development

Hegewisch, A., Childers, C. and Hartmann, H. 2019. *Women, Automation and the Future of Work*. Washington, DC: The Institute for Women's Policy Research. https://iwpr.org/iwpr-issues/employment-and-earnings/women-automation-and-the-future-of-work/

Herrera, C. 2020. TIBCO: The good, the bad, and the human sides of AI. *Women in Big Data Webinar on AI and Ethics*, 21 May. https://www.youtube.com/watch?v=qLFIfHoRe_E

IEEE. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. First Edition. New York: IEEE. https://ethicsinaction.ieee.org/

Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol. 1, pp. 389–399. DOI: 10.1038/s42256-019-0088-2.

Kwet, M. 2019. Digital colonialism is threatening the Global South. *Aljazeera*, 13 March 2019. https://www.aljazeera.com/indepth/opinion/digital-colonialism-threatening-global-south-190129140828809.html

Litt, M. 2018. I accidentally built a brogrammer culture. Now we're undoing it. *Fast Company*, 2 February. https://www.fastcompany.com/40524955/i-accidentally-built-a-brogrammer-culture-now-were-undoing-it

Microsoft. 2020. Microsoft AI Principles. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

Mozilla Foundation. 2020. Internet Health: Trustworthy Artificial Intelligence. https://foundation.mozilla.org/en/internet-health/trustworthy-artificial-intelligence/

Murphy, M. 2020. Inside Google's struggle to control its 'racist' and 'sexist' AI. *The Telegraph*, 16 January. https://www.telegraph.co.uk/technology/2020/01/16/google-asking-employees-hurl-insults-ai/

OECD. 2020. Recommendation of the Council on Artificial Intelligence. *OECD Legal Instruments*, OECD/LEGAL/0449. Paris: OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

ORBIT. 2019. 100 *Brilliant Women in AI and Ethics: Conference Report*. Leicester, UK: Observatory for Responsible Research and Innovation in ICT. https://www.orbit-rri.org/100-plus-women/

Pang, W. 2019. How to Ensure Data Quality for AI. *insideBIGDATA*, 17 November. https://insidebigdata.com/2019/11/17/how-to-ensure-data-quality-for-ai/

Peña, P. and Varon, J. 2019. Decolonising AI: A transfeminist approach to data and social justice. *Global Information Society Watch 2019: Artificial intelligence: Human rights, social justice and development*. Johannesburg, South Africa: APC. https://www.giswatch.org/2019-artificial-intelligence-human-rights-social-justice-and-development

Sciforce. 2020. Introduction to the White-Box AI: the Concept of Interpretability. *Medium*, 31 January. https://medium.com/sciforce/introduction-to-the-white-box-ai-the-concept-of-interpretability-5a31e1058611

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D. M., Kalai, A. T. 2019. What are the biases in my word embedding? Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), Honolulu, HI, USA, 2019. https://arxiv.org/pdf/1812.08769.pdf

Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. and Schiebinger, L. 2019. Sex and gender analysis improves science and engineering. *Nature*, Vol. 575, pp. 137–146. DOI: 10.1038/s41586-019-1657-6.

UNESCO and EQUALS. 2019. *I'd Blush if I Could*: *Closing Gender Divides in Digital Skills Through Education*. Geneva: EQUALS. https://unesdoc.unesco.org/ark:/48223/pf0000367416

UNESCO. 2019. Preliminary Study on a Possible Standard-Setting Instrument on the Ethics of Artificial Intelligence, 40 C/67. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367823

UNESCO. 2020. Outcome Document: First Version of a Draft Text of a Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/AHEG-AI/2020/4REV. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000373434

UNESCO. 2020. Working Document: Toward a Draft Text of a Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/AHEG-AI/2020/3REV. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000373199

UNI Global Union. 2017. *Top 10 Principles For Ethical Artificial Intelligence*. Nyon, Switzerland: UNI Global Union. http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/

United Nations System Chief Executive Board for Coordination (CEB). 2019. A United Nations system-wide strategic approach and road map for supporting artificial intelligence capacity development on artificial intelligence. CEB/2019/1/Add.3. Geneva: CEB. https://undocs.org/en/CEB/2019/1/Add.3

Weinberger, D. 2019. Where fairness ends. *Harvard Business School Digital Initiative*, 26 March. https://digital.hbs.edu/artificial-intelligence-machine-learning/where-fairness-ends/

West, S. M., Whittaker, M. and Crawford, K. 2019. *Discriminating Systems: Gender, Race and Power in AI*. New York: AI Now Institute. https://ainowinstitute.org/discriminatingsystems.html

Wong, J. C. 2016. Women considered better coders – but only if they hide their gender. *The Guardian,* 12 February. https://www.theguardian.com/technology/2016/feb/12/women-considered-better-coders-hide-gender-github

World Economic Forum Global Future Council on Human Rights 2016-2018. 2018. How to Prevent Discriminatory Outcomes in Machine Learning. *White Paper*, March 2018. Geneva: World Economic Forum. https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning