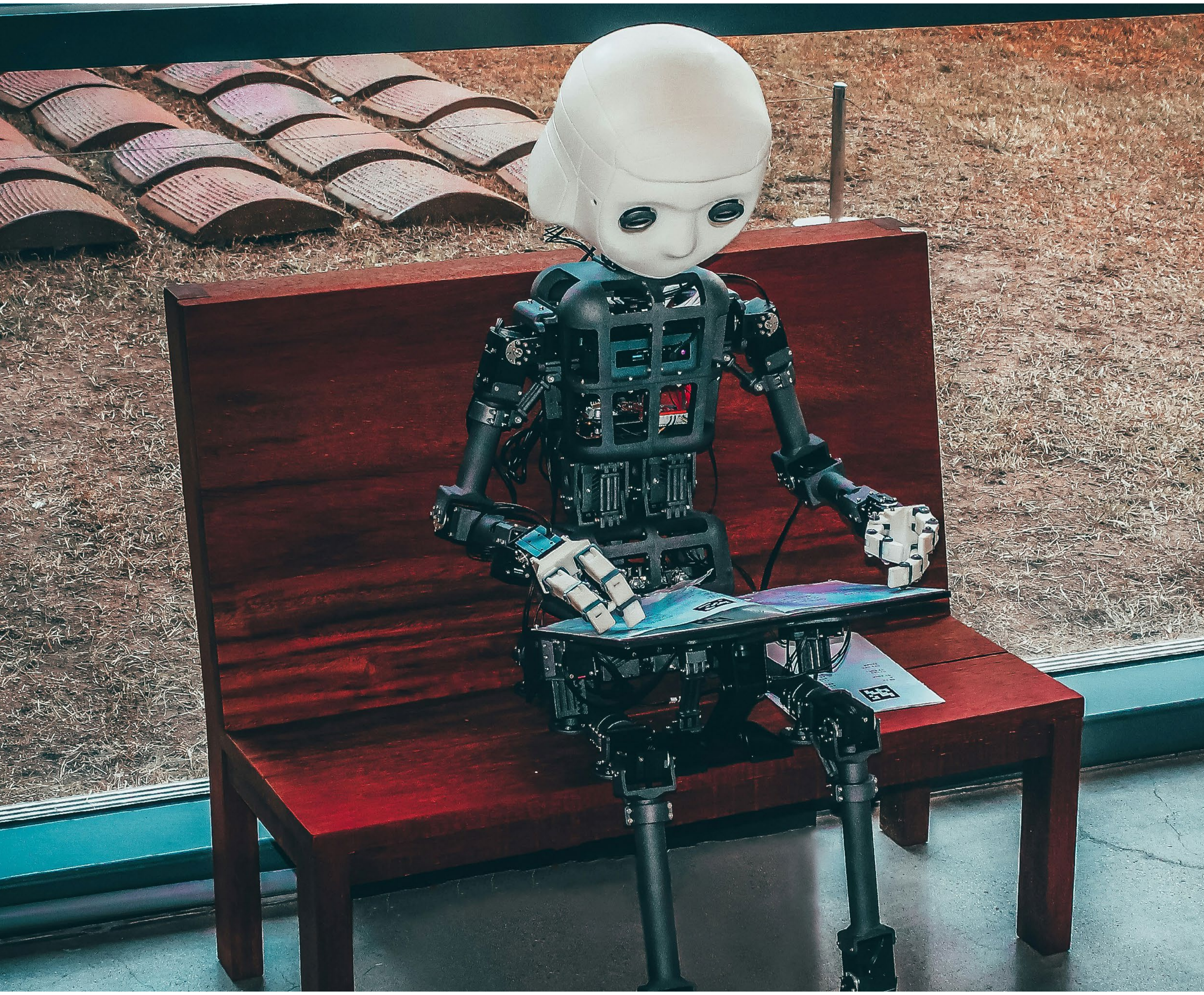
NETHOPE



NetHope's Center for the Digital Nonprofit

GENDER EQUITABLE AI TOOLKIT

Table of **Contents**

Acknowledgements

Gratitude is extended to the support and inspiration behind this research. This includes all those who have participated in the key informant interviews as well as the initiators of this project including Bo Percival, Leila Toplic, and Synne Marion Olsen, whose contributions these individuals made during their time at NetHope, and their respective organizations is the epitome of collective action. Appreciation is also extended to supporters of the NetHope Collective Impact Fund as well as the founders and partners of NetHope's Center for the Digital Nonprofit.

THE PATTERSON FOUNDATION

Andy Pokladowski
Akhtar Badshah
Cory Evans
Craig Molyneaux
Dipak & Radha Basu
Edward Happ
Liz Bronder
Macon Philips
Priscilla Chomba
Sibel Berzeg
Randy Pond
Ted Beatie

Partners of NetHope's Center for the Digital Nonprofit

Microsoft
avanade
servicenow
box
okta

NetHope Members

The successful completion of this research would not have been possible without the contributions and support of NetHope's esteemed members. We extend our deepest gratitude for their dedication to leveraging technology for social impact.

actionaid
americares
care
CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE
CRS CATHOLIC RELIEF SERVICES
CGIAR
ChildFund International
children international
christian aid
Compassion
Concern worldwide
CORUS
DCA actalliance
DRC DANISH REFUGEE COUNCIL
Direct Relief
fhi360
FOOD FOR THE HUNGRY
FFTP FOOD FOR THE POOR
Global Communities
GOAL
Habitat for Humanity
HEIFER INTERNATIONAL
HELVETAS
HIAS
Humanitarian OpenStreetMap Team
IFRC
International Medical Corps
RESCUE
Islamic Relief Worldwide
JSI
msh
MEDAIR
MEDECINS SANS FRONTIERES
MEDICAL TEAMS International
Mennonite Central Committee
MERCY CORPS
Mercy Ships
MSI REPRODUCTIVE CHOICES
NRC NORWEGIAN REFUGEE COUNCIL
orbis
OXFAM
PLAN INTERNATIONAL
pro mujer
RELIEF INTERNATIONAL
RIGHT TO PLAY
THE SALVATION ARMY
Samaritan's Purse
Save the Children
simprints
SOS CHILDREN'S VILLAGES
TEAM RUBICON
The Asia Foundation
THE CARTER CENTER
TREK MEDICS INTERNATIONAL
Trócaire
ICC international computing centre
Ushahidi
VSO
WAR child HOLLAND
WaterAid
WELT HUNGER HILFE
Wildlife Conservation Society
WINROCK INTERNATIONAL
World Vision
WWF

Contact the Research Team

Questions, comments, and collaboration interests regarding this research are welcomed. To pursue further exchange or provide input around this topic, please contact the study authors and administrators using the information provided below.

NETHOPE

THE CENTER FOR THE DIGITAL NONPROFIT

LinkedIn:
Daniela Weber
Jean-Louis Ecochard
Nicholas Kerastas

Email:
Daniela.weber@nethope.org
Jean-louis.ecochard@nethope.org
Nicholas.kerastas@nethope.org

Schedule:
Schedule a meeting with the team here.

Toolkit Overview

INTRODUCTION

With the rapid development of functionality in machine learning (ML) and artificial intelligence (AI) using these innovations for the humanitarian sector has the potential to address complex humanitarian and development challenges. However, without intentional efforts, this same technology can become a mechanism for extended inequities and harm, particularly from the lens of gender.

The intention of this toolkit is to inform humanitarian and development professionals on best practices for gender equitable AI and ML solutions. In doing so, the toolkit enables nonprofit organizations to frame the deployment of innovative technologies through an ethical lens – focused on gender – that deliver positive impact for people in need.

The Gender Equitable AI Toolkit is a product delivered through funds from NetHope's Collective Action Fund. The work behind this toolkit was organized by community leaders of NetHope's AI Workstream. Research began in late 2021 and drew on input from over 30 consultations with representatives from the NetHope Member and Partner ecosystem. As a result, this research represents an aggregation of best practices and resources from a diverse array of subject matter experts.

USAGE

This toolkit focuses on leveraging and fine-tuning AI solutions for humanitarian and development services. The content of the toolkit can be accessed in two ways:

- An e-learning course through NetHope's Digital Leadership Institute (NDLI) that serves as a user-friendly learning repository for the toolkit and emerging best practices, subject matter expertise, and practical exercises for building capability.
- A downloadable PDF document containing the toolkit content which can be used if the NDLI platform is inaccessible to users.

The toolkit offers practical guidance on the entire lifecycle of deploying gender equitable AI solutions. It encompasses topics such as problem statement identification, data collection and preprocessing, algorithmic design, model evaluation, user engagement, and continuous learning and improvement.

AUDIENCE

This toolkit will provide readers with principles and actionable steps to implement gender equity within the AI/ML development lifecycle. It includes an introduction to AI/ML ethics, and an overview of the lifecycle of development behind an AI/ML solution. Accordingly, the primary and secondary audience is as follows:

- Primary: non-technical staff (i.e. program staff)
- Secondary: technical staff (i.e. analytics, MEAL, or IT)

The toolkit content is designed as an entry point for humanitarian and development professionals seeking to deploy AI and ML solutions in a gender equitable capacity. Knowledge gained here can be combined with existing technical expertise from across organizations thereby enabling nonprofit professionals to make meaningful contributions to equitable adoption of AI and ML solutions.

Important Concepts

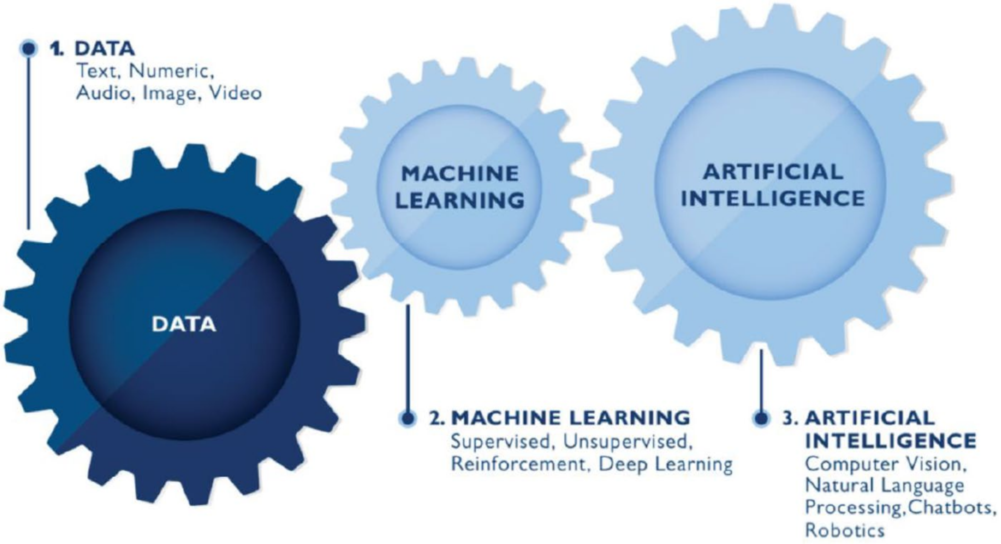
Machine Learning (ML): is a set of methods for getting computers to recognize patterns in data and use these patterns to make future predictions. For shorthand, one can think of ML as "data- driven predictions."

Big Data: A set of technologies developed to handle data sources that are "big" in terms of volume, velocity, or variety. While the term "Big Data" emphasizes data management more than learning and predictions, many former Big Data companies have rebranded themselves as AI companies, and there is broad overlap in tools and techniques. Big Data provides the extensive datasets needed for ML and AI, enabling data-driven insights and decision-making.

Deep Learning: Deep Learning is an advanced subset of Machine Learning (ML) that revolves around the utilization of neural networks with multiple layers, known as deep neural networks. These networks are designed to automatically extract intricate and hierarchical features from vast datasets. Deep Learning significantly enhances the capabilities of Artificial Intelligence (AI) systems by enabling them to process complex data representations, making it especially effective in tasks like image recognition, speech recognition, and natural language understanding.

Artificial Intelligence (AI): Uses computers for automated decision-making that is meant to mimic human-like intelligence. Automated decisions might be directly implemented (e.g., in robotics) or suggested to a human decision-maker (e.g., product recommendations); the most important thing for our purpose is that some decision process is being automated. For shorthand, you can think of AI as "smart automation."

Figure A) Conceptualizing AI/ML concepts.


1. DATA
Text, Numeric, Audio, Image, Video
DATA
MACHINE LEARNING
ARTIFICIAL INTELLIGENCE
2. MACHINE LEARNING
Supervised, Unsupervised, Reinforcement, Deep Learning
3. ARTIFICIAL INTELLIGENCE
Computer Vision, Natural Language Processing,Chatbots, Robotics

Source: NetHope's AI Ethics Workshop for Nonprofits

Natural Language Processing (NLP): Natural Language Processing is a specialized capability of AI that focuses on the interaction between computers and human language. NLP equips AI systems with the ability to understand, interpret, and generate human language text or speech. It involves the application of ML and deep learning techniques to tasks such as language translation, sentiment analysis, chatbots, and text summarization. NLP plays a vital role in AI applications where communication with humans in a natural and intuitive manner is essential.

Generative AI: Represents an advanced concept within AI and ML, building upon the foundations of deep learning. It pertains to the development of models and algorithms capable of generating new data samples based on patterns learned from existing data. These models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), enable generative AI systems to produce novel content, including images, text, music, or even entire datasets, simulating human-like creativity in various domains.

Examples of Gender Bias in AI / ML

The following examples illustrate the issues that can arise if AI solutions are not developed with a gender equitable lens.

Language Embeddings in Google Translate

Research shows that translations by Google Translate between gender-inflected languages and non-gender-inflected languages led to embedded bias. Female occupations were often incorrectly translated into their male forms. The core issue here is the reliance on large bilingual datasets for training machine translation models. These datasets are often skewed towards certain languages and lack comprehensive coverage of gender-inflected languages. Consequently, machine translation systems like Google Translate may struggle to generalize the intricate gender-related rules and exceptions present in these languages.

Bias in Amazon's Assisted Job Matching

Amazon's AI powered recruiting tool recommended fewer women for technical positions. This bias was attributed to the design of the system, which relied on historical hiring data that reflected gender imbalances in the tech industry. The implicit preference for male candidates in the tech industry was then carried over in evaluating new candidates, thus exacerbating the gender bias in the industry. This case underscores the critical importance of considering the quality and fairness of training data when developing AI systems. In this instance, the reliance on historical data not only failed to address the gender imbalance but also amplified it.

Gendering of AI Voice Assistants

Examples of outputs from digital voice assistants such as Amazon's Alexa and Apple's Siri technology, shows how women are impacted by sexist norms that society continues to embrace. The major concern is the way in which these voice assistants respond to gender-related queries or comments. They often default to using female voices, reinforcing the historical association of women with servitude and subservience. When these technologies embody and propagate sexist biases, they not only perpetuate harmful stereotypes but also undermine collective efforts to advance gender equality.

Examples of Nonprofit Gender Equitable AI / ML

By highlighting these cases of equitable AI implementations, it becomes apparent that practitioners in the NGO sector can utilize this technology to successfully achieve impact.

Representative Data with CyberSight AI

Utilizes algorithms to analyze retinal images, thereby identifying signs of diabetic retinopathy. This pilot project has set a precedent for AI-supported diabetic retinopathy screenings, with broader applicability across other healthcare settings. The solution leveraged multi-modal input from images and textual data to enhance the diagnosis and referral process for eye care concerns. By using multi-modal input, representative data was available for recognizing health patterns darker skin tones as well as solution features that assist in bridging language barriers across cultural contexts.

Gender Differentiating Credit Scoring

UC Berkeley is building a machine learning model for RappiCard in Mexico that conducts differentiated credit scoring by gender to expand women's access. They compare the model to Rappi's performance to assess if a gender-specific algorithm leads to fairer approvals. Addressing inequity in credit scoring aims to increase financial access and economic opportunity for women that are frequently denied credit based on traditional scoring models.

Findings from the NetHope Community

During the development of this toolkit, NetHope performed key informant interviews with subject matter experts on artificial intelligence and machine learning in the nonprofit and humanitarian sector. This direct input from community practitioners served as the guide for the project team during toolkit development.

Approach	Format	Number of Participants
Key Informant Interviews	Standardized Interview Protocol	37 Professionals

Below are some of the key findings and input received from the community grouped into thematic areas that were identified through analysis of key informant input:

Data	Design	Scale
The input for AI / ML	The creation of AI / ML	Expanding Access to AI / ML
The digital universe is estimated to be about 44 zettabytes in 2020.	Neural networks can now have billions of parameters (e.g., OpenAI's GPT-4 has around 175 billion parameters).	AI is projected to contribute up to $15.7 trillion to the global economy by 2030.

Data for AI / ML

Data is the bedrock of AI and ML, influencing how models perceive and interact with the world. For nonprofits, lessons learned from the NetHope community around data collection and governance underscore the importance of ensuring representation of gender identity as well as other socioeconomic variables.

- The data collected by NGOs inherently reflects societal inequities, creating a biased foundation for AI / ML initiatives. Methodologies based on existing community norms can reinforce biased gender norms, favoring men in powerful positions.

- Current NGO standards lack emphasis on sex disaggregated data, hindering data-driven decisions related to gender. Ineffective training data, often from elite northern institutions, perpetuates biases, posing a disadvantage for gender equity.

- The governance of NGO data lifecycles have shifted from an individualist to a collective approach, fostering intersectionality around gender-related issues, by securing and utilizing collected data responsibly.

- Defined labeling rules, gender-specific requirements, and multiple annotators from diverse backgrounds are key to generating unbiased solutions. Evaluation metrics, like Gender Parity Index, enable gender audits for bias mitigations.

- Actively involving community members in data collection fosters a more inclusive and representative dataset, enhancing the effectiveness of AI solutions. Weighing segmented data samples, and foresight exercises contribute to enriched gender datasets.

Design for AI / ML

Design is not merely about user friendly interfaces or unique app functionalities. Rather, it encompasses the ethical considerations of the solutions being created. For nonprofits, lessons learned from the NetHope community highlight the need for participatory design and purposeful engagement with concepts of gender identity to ensure AI / ML solutions meet the needs of people.

- Acknowledging gender is crucial for designing technologies that meet unique needs and address historical underrepresentation.
- The deep-rooted issue of women's underrepresentation in AI leads to biased feature engineering, necessitating interdisciplinary project teams.
- Gender equity in design extends beyond hiring, emphasizing consideration of how specific gender groups access the internet and practice their rights.
- Adopting gender-inclusive design principles, such as participatory design, is key for developing AI solutions that address specific gender outcomes and build trust
- Open-source feminist design tools, user-centered policy design maps, and ethical AI principles offer starting points for practitioners to build gender-specific AI / ML solutions.
- Participatory design approaches center solutions around the lived experiences of women, fostering trust, safety, and privacy from the outset.
- Iterative design and data lifecycle evaluation, including algorithmic audits and bias remediation techniques, are essential for building trust in AI solutions.

Scale for AI / ML

Scaling solutions presents unique challenges and opportunities for nonprofits. Lessons learned from the NetHope community emphasize the importance of maintaining ethical standards and gender inclusiveness as solutions reach a wider audience. As AI technologies scale, so does their impact, making it crucial for organizations to ensure that biases are not amplified and that the benefits are equitably distributed across genders.

- The open sourcing of AI models and cross-sector collaborations, such as ethical hackathons, play a crucial role in expanding resources for gender equity tools and fostering shared evaluations and adoption of solutions.
- Public APIs and strategic partnerships with technology vendors enhance the scalability of solutions, enabling access to large gender inclusive datasets and cloud-based infrastructures for more effective models.
- As AI solutions scale the inherent biases may amplify. Digital skill gaps and organizational capabilities posing significant barriers to implementing gender equity principles effectively.
- While scaling solutions, it's vital to localize to meet the specific needs and contexts across regions. This ensures that the solutions are not only scalable but also inclusive and effective.

Principles for Gender Equitable AI/ML

Overview of Principles

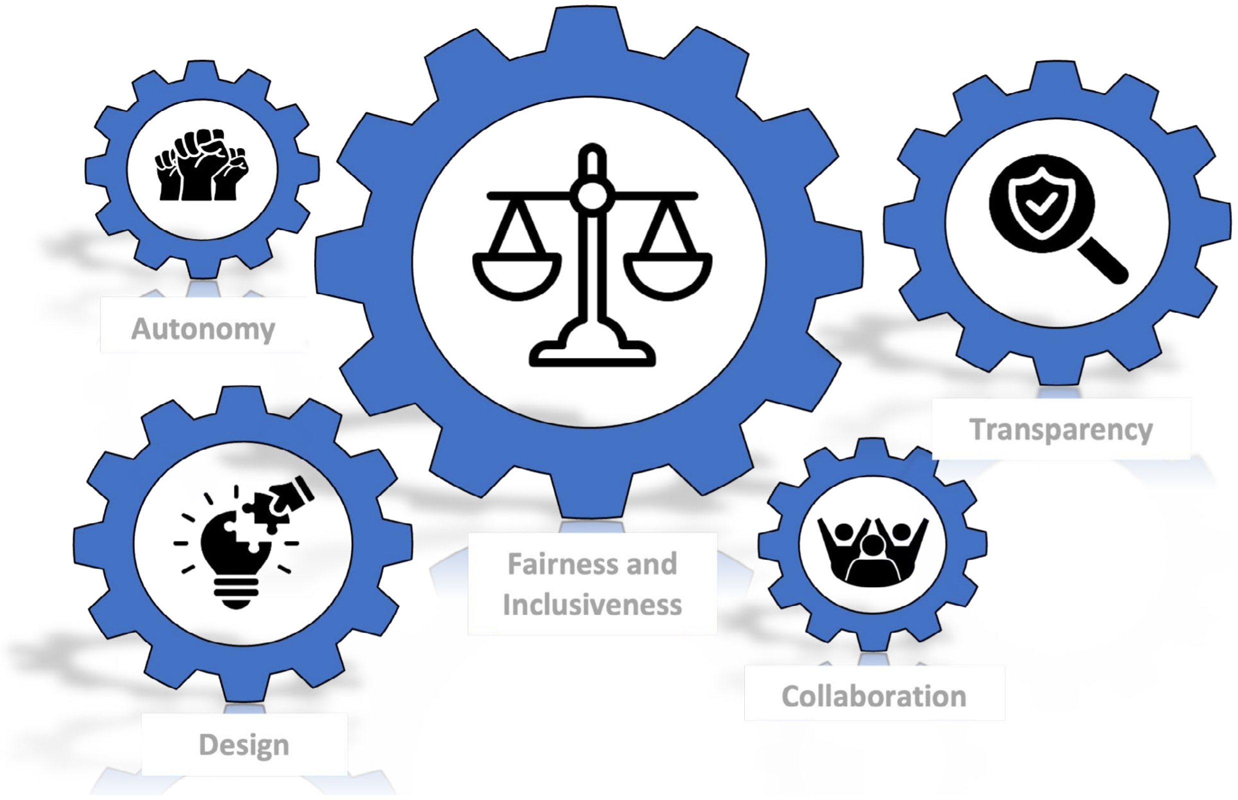
AI/ML based solutions hold a promise of becoming a transformative force in multiple facets of humanitarian and development work. However, alongside this lies the potential of negative implications in the form of furthering entrenched biases and exacerbating inequality. The principles have been developed to provide a framework for ensuring gender equitable AI.

The principles laid out in this toolkit are sourced from a synthesis of input from the Member and Partner community. Each principle represents a response to the challenges identified in consultations with NetHope Members and Partners. Through these collaborations, NetHope identified 5 sets of principles for gender equitable AI development in the following areas:

- **Principles related to fairness and inclusivity.**
- **Principles related to transparency.**
- **Principles related to design and development.**
- **Principles related to governance and autonomy.**
- **Principles related to collaboration and capacity building.**

To mitigate risk, it is critical that implementing organizations adopt principles - like the above - and thereby champion gender equity. At the core of these intertwined principles is the concept of inclusiveness and each following principle serves to promote the visibility of underrepresented gender groups (Figure B). Moreover, each principle is matrixed and thus embedded into the various stages in the development of an AI/ML solution.

Figure B) A representation of the above principles for gender equity in AI/ML



Principles for Gender Equitable AI/ML

Principles related to fairness and inclusivity

Embracing fairness and inclusivity principles enable organizations to tackle biases, combat discrimination, and advance inclusive technology solutions. This entails considering the impact of AI on priority gender groups, involving diverse perspectives in decision-making, and ensuring for representative data, algorithms, and models. Inclusive AI solutions are characterized by empathy, cultural sensitivity, and a contextualized understanding of human complexities. They should also aim to bridge existing digital divides, empower marginalized communities, and foster social justice. By championing inclusivity, organizations foster a digital ecosystem that distributes AI benefits in an equitable manner, and that no one is left behind.

Accessibility: AI initiatives must tackle challenges related to equitable access to digital resources. This involves bridging the gap for individuals across the globe, especially those in underserved regions. The overarching goal is to protect the rights of individuals and prevent harm in the design and deployment of AI solutions.

Gender-Centricity: AI initiatives should have a dual objective to first support existing program work as well as to actively address existing inequities by considering historical disparities in humanitarian aid. By establishing comprehensive metrics and objectives that effectively advance both facets, organizations can systematically measure the impact of AI on programmatic effectiveness and gender equity, thereby fostering fairness in the conception and execution of AI technologies.

Bias Mitigation: AI initiatives should acknowledge the inherent biases present within culture and shared languages thus dictating the creation of data sources or algorithms which can perpetuate and reinforce gender stereotypes. It underscores the need to take proactive measures in countering an inherently biased world. Recognizing that language itself can carry historical and societal biases, the principle emphasizes the significance of addressing the potential biases ingrained in the data used to train AI models.

Figure C) A Dalle-2 generated image on "Henri Matisse impression of accessibility."

Principles related to Transparency

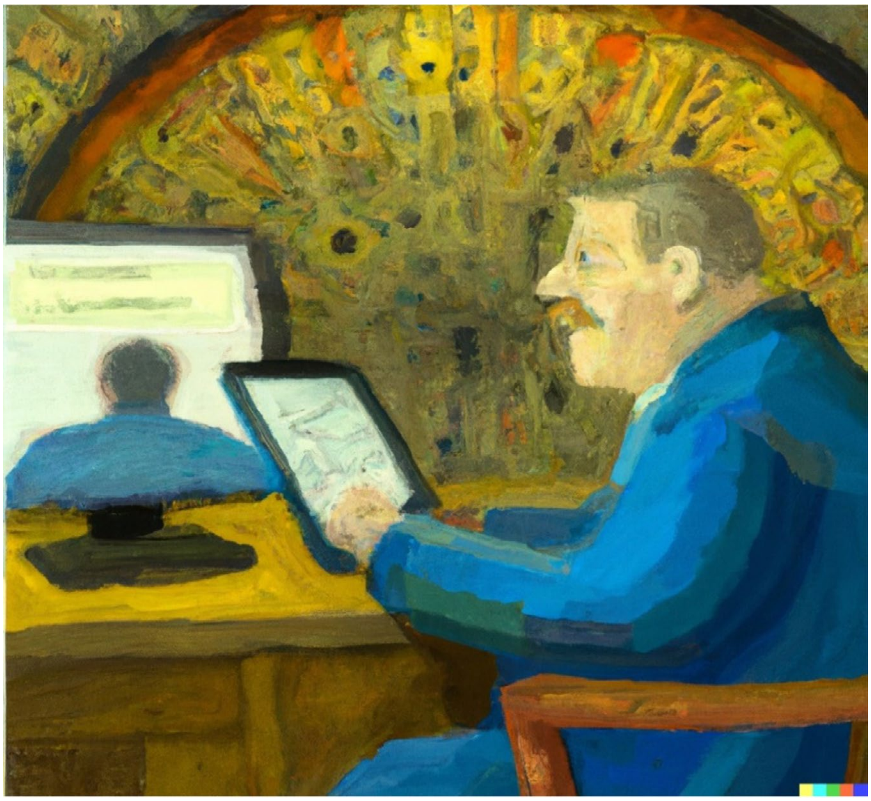
Transparency in AI ensures openness and the demystification of AI systems thereby enabling a deeper understanding of outcomes and potential biases. By embracing transparency, organizations also champion the values of inclusivity, ensuring that AI technologies are developed and deployed in a manner that promotes fairness and the equitable representation of non-expert perspectives. Transparent practices also encourage the collaboration and feedback necessary for continuous improvement, enhancing the overall impact of gender equitable AI solutions.

Accountability: AI initiatives must embrace responsibility for the 'Do No Harm' principle alongside the significance of human autonomy in the development of AI solutions, safeguarding human agency and control. Those involved in the development, deployment, and governance of AI systems should be accountable for their impact, and there should be mechanisms for addressing any harm caused. Organizations can uphold accountability by establishing clear lines of responsibility, ensuring that AI solutions are fit for purpose, and monitoring the impact of AI solutions.

Explainability: AI initiatives should ensure solutions are demystified for non- experts. It is critical to provide clear documentation that explains how a given model works and that design decisions are rationally explained. By demystifying complex algorithms and offering simple explanations, organizations empower everyone regardless of technical background. Achieving explainability not only fosters understanding but also promotes the sustainability of AI solutions by enabling long-term usability, maintenance, and stakeholder engagement.

Value Maintenance: AI initiatives should emphasize the need for donors to recognize the equal importance of maintenance and care for critical systems alongside innovation. Organizations should advocate that public funding develops AI solutions that are a public good and accordingly prioritize open-source methods. By dedicating resources to value maintenance, organizations ensure the longevity and sustained impact of AI initiatives. This principle underscores that innovation is just the beginning; consistent and responsive upkeep is essential to meet evolving needs of different gender groups.

Figure D) A Dalle-2 generated image on "Van Gogh's impression of Digital Accountability."



Principles related to design and development

Design and development principles guide the creation of AI systems that are accessible, unbiased, and responsive to the needs and aspirations of diverse individuals and communities. By embracing these principles, organizations can ensure that AI technologies are designed with empathy, cultural sensitivity, and the deliberate inclusion of diverse perspectives. This approach fosters the development of AI systems that are impactful and capable of addressing societal challenges while leaving no one behind.

Gender Centered Design: AI initiatives must center solutions around the needs, perspectives, and leadership of priority gender groups at the local level. Participatory design, in this context, involves ensuring that the voices and experiences of gender are represented in decision-making processes. Organizations should actively engage with communities, seeking their input, feedback, and co-creation throughout the design and development of AI technologies. By embracing participatory design, AI initiatives empower communities to influence the technology landscape resulting in solutions that genuinely meet their aspirations and needs.

Intersectionality: AI initiatives should recognize that gender is a critical aspect of an individual's identity, and that like any human group, deserves dignity and visibility in digital ecosystems. This principle emphasizes moving beyond a monolithic understanding of gender and pushes practitioners to consider gender identity during data collection and curation. This entails intentionally implementing data collection methods that seek out diverse perspectives and experiences, ensuring comprehensive representation. By embracing an intersectional lens in data collection organizations can ensure that AI models are built upon an equitable foundation of representative data sources.

Model Awareness: AI initiatives should emphasize the importance of training models in ways that raise awareness of fairness, inclusivity, and equal representation of diverse gender groups. They should acknowledge the potential biases in data and algorithms that can perpetuate gender inequalities and take steps to mitigate these biases during the training process. By incorporating conscious model awareness, AI initiatives enhance the ethical and social responsibility of their systems, ensuring that AI technologies contribute positively to gender equity and broader social progress.

Figure E) A Dalle-2 generated image on "Frida Kahlo's impression of Female Centered Design."

Principles related to governance and autonomy

Governance principles provide the foundation for ensuring ethical, legal, and responsible decision-making in the development, deployment, and use of AI technologies. They emphasize the establishment of clear policies, guidelines, and oversight mechanisms to promote accountability, transparency, and the protection of individual rights and societal well-being. By embracing these principles, organizations can foster an ecosystem of responsible AI use, where ethical considerations, fairness, and human values are at the forefront. This approach ensures that AI technologies are harnessed in a manner that aligns with the broader public interest, upholds fundamental rights, and promotes the common good.

Autonomy: AI initiatives must recognize an individuals' rights of ownership over personal information and identities. Any AI enabled solution needs to implement informed consent mechanisms when collecting and using the personal information of any individual. This principle seeks to challenge the notion of viewing people as consumers or users by promoting respect for an individual's right to autonomy and privacy.

Human Oversight: AI initiatives should uphold human autonomy by safeguarding human agency and control. This involves integrating human participation, critical discernment, and moral assessment in AI implementation, ensuring that the technology functions as a supportive instrument guided by human supervision. Place paramount importance on ethical management of data and models to counteract risks and prevent the amplification of vulnerabilities within marginalized groups. This overarching principle underscores the necessity for rigorous supervision and answerability across the entire AI lifecycle.

Privacy by Default: AI initiatives should surpass regulatory compliance and proactively safeguard communities, particularly marginalized gender groups. This principle demands comprehensive data governance, transparent communication, and empowering individuals to control their data, ensuring not just legal compliance but the prevention of harm in the digital realm. By prioritizing "pervasive privacy and security," NGOs embrace their role as custodians of data and champions of digital well-being, setting a higher standard for responsible AI implementation.

Figure F) A Dalle-2 generated image on "Banksy's impression of human autonomy."

Principles related to collaboration and capacity building

Collaboration principles emphasize the importance of fostering partnerships and knowledge exchange among stakeholders, including academia, industry, civil society, and government entities. By working together, organizations can pool resources, expertise, and diverse perspectives to address complex challenges and drive innovation in AI. Capacity building principles focus on enhancing the knowledge, skills, and capabilities of individuals and organizations in leveraging AI technologies. This includes promoting education, training, and professional development programs to empower stakeholders with the necessary skills and competencies to develop, deploy, and govern AI systems responsibly.

Localization: AI initiatives must foster collaborations between local affiliates, governments, nonprofits, and technical experts. These forms of collaborations aim to co-create policies, guidelines, and frameworks that promote inclusive and ethical AI practices. Additionally, they prioritize efforts in designing solutions that purposefully consider the perspectives of all gender identities. By working together, these stakeholders contribute to the development of responsible AI solutions that account for bias at the outset and thus enhance the overall transparency of AI technologies.

Inclusivity: AI initiatives should recognize the importance of assembling teams with a diverse composition. This includes considering factors such as backgrounds, perspectives, skills, experiences, and demographics that align with the specific use case and performance requirements of the AI system.

Capacity Building: AI initiatives should incorporate capacity building efforts that go beyond inclusivity, considering factors such as infrastructure and connectivity, access to devices and hardware, as well as digital literacy. Moreover, to ensure sustained growth and learning, organizations should actively seek to establish communities of practice. These communities provide a platform for stakeholders to collaboratively share knowledge and best practices thus fostering a supportive environment for continuous learning.

Figure G) A Midjourney generated image on "Loïs Mailou Jones' impression of inclusivity."

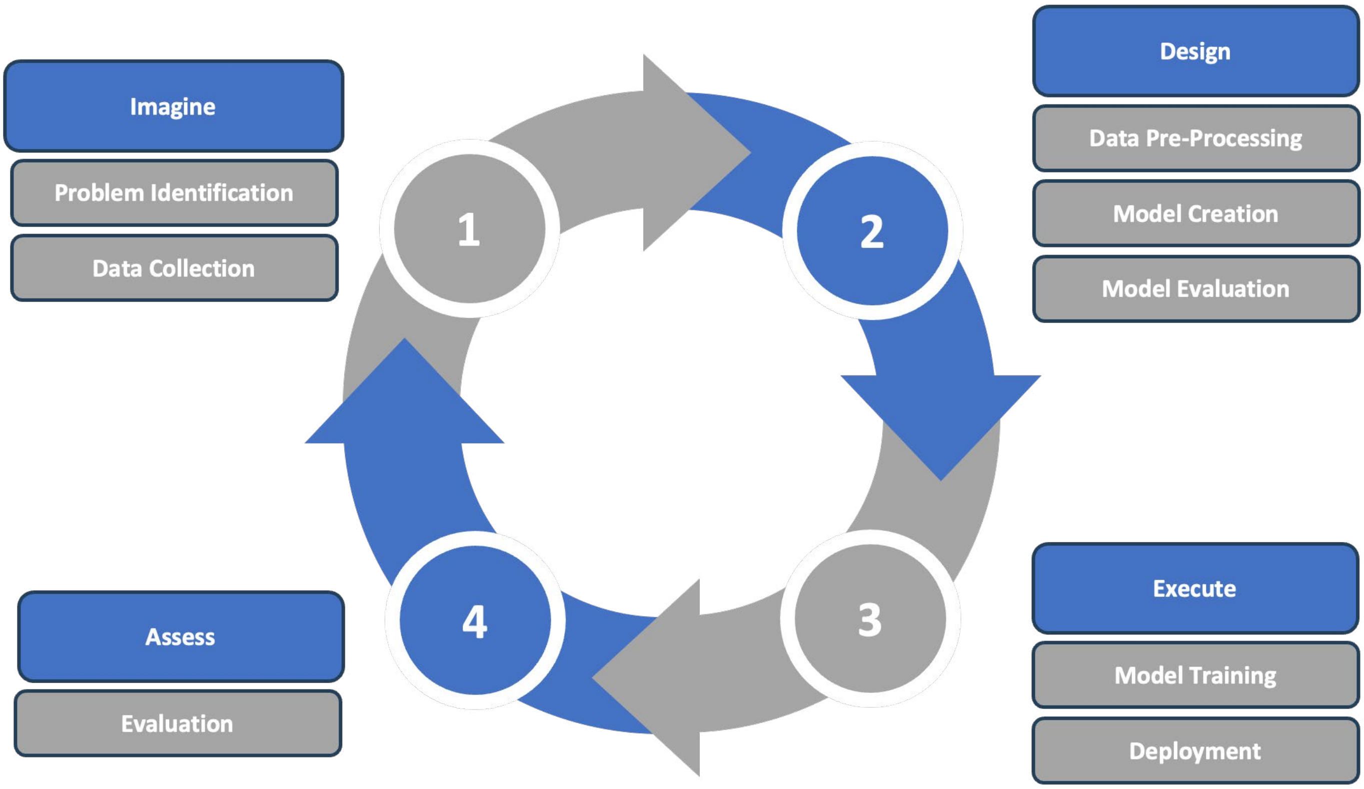
Key Stages in Solution Development

Below are the key stages in the development and deployment of an AI/ML enabled solution. These stages are based on previous research the community has done in the advancement of the NetHope AI Ethics Toolkit. Thus, with minor revisions, they are consistent with previous community language around AI/ML. The stages are as follows:

1. **Problem identification** is where the team defines the objectives and the data needed to address the objectives
2. **Data collection** is where the team consolidates different data, either collected internally or from external sources.
3. **Data pre-processing** is where the team cleans data and labels data (in supervised learning), preparing it to be used by model.
4. **Model creation** is the core technical step where the team selects and develops potential models using the preprocessed data. Data is split into a training set for building the models and test set for validating the models.
5. **Model evaluation** is where models are tested against predefined criteria (such as accuracy, performance, etc) to determine which approach is best suited for the problem.
6. **Model tuning and training** is where appropriate threshold values and hyperparameters are set to optimize the model's performance.
7. **Deployment** is where the model is used in real-world applications, ideally starting with a smaller beta testing phase.
8. **Evaluation** involves constantly checking the model to make sure it works as intended, revisiting earlier phases and/or retraining the models when new data comes in.

These stages also align with the four phases of NetHope's IDEA Journey design methodology which can be utilized by nonprofits to support a user centric approach to accelerated digital transformation.

Figure H) A representation of the development behind an AI/ML Solution (NetHope IDEA Journey)



Problem identification is where the team defines the objectives and the data needed to address the underlying problem.

Questions to ask:

1. What is the problem that we are trying to solve and how does gender play a role?
2. How do existing processes and technologies currently deal with the problem?
3. Is an AI/ML solution necessary?

Principles to follow:

- Accessibility
- Accountability
- Gender Centered Design
- Autonomy
- Localization

Checklist:

□ **Mitigate project bias** by establishing gender balanced selection process that evaluates project applicants strictly on skills and experience using standardized criteria or blind screening techniques. This minimizes the influence of implicit bias in project team decisions during development.

□ **Create an ethics committee** with at least 3 experts together covering domains of AI, program implementation and data privacy to advise on all aspects of gender equity. Multidisciplinary input like this further guards against blind spots in later solution development.

□ **Review laws or regulations** to align with requirements established in existing legal frameworks like the EU AI Act, GDPR, OECD AI Principles etc. related to transparency, explainability, risk assessments, and prohibited practices by data owners.

□ **Complete a landscape analysis** detailing which existing systems fail to resolve gender inequality evidenced in the problem area. This helps build the case for why status quo approaches have fallen short and where an AI / ML solution can value-added.

□ **Evaluate case studies** and document the pros/cons to prepare a selection matrix that assesses the applicability of different AI / ML approaches. This analysis ensures for the reuse of best practices and grounds solution development based on lessons learned from other implementations.

□ **Conduct human centered design workshops with relevant community members and domain experts to gather needs, solidify the core problem, and frame requirements for addressing identified gender biases and related barriers.**

□ **Complete an impact assessment** estimating outcomes along with an evaluation of accessibility to digital resources and participation in development process. This flags potential exclusion and risks associated with gender.

Problem identification (cont.)

Key Step: Human Centered Design for AI / ML

Human-centered design experiences that engage local communities – and priority gender groups – are a critical first step when seeking to build AI solutions for gender equitable outcomes. Facilitating inclusive workshops builds understanding of gendered and intersectional perspectives within client groups. Listening and co-designing with users fosters solutions that earn community trust and uptake. Representativeness in participation and constant feedback loops with clients are vital to developing equitable, unbiased AI that serves its intended purpose.

1. **Define Participant Criteria:** Establish clear participant criteria to ensure gender balance and representation of diverse perspectives, fostering inclusivity in the development of AI solutions.

2. **Identify Problems:** Listen to key stakeholders and communities, identifying participant pain points to inform the development of gender-equitable AI solutions tailored to address local challenges and empower communities with context-specific innovations.

3. **Prioritize Accessibility:** Emphasize accessibility in the design approach, considering factors such as language, literacy levels, and digital literacy, to ensure the inclusivity and usability of AI solutions across diverse demographics.

4. **Develop Personas:** Determine a design methodology rooted in intersectionality, acknowledging, and addressing compounding discrimination to create AI solutions that consider the multifaceted experiences of individuals.

5. **Prepare Design Approach:** Prepare a design approach that incorporates storytelling, journey mapping, and co-design activities, ensuring gender dimensions are central to the development of AI solutions.

6. **Host Design Workshops:** Host design workshops with limited participants, following participatory design best practices, to facilitate candid sharing of gendered perspectives and experiences.

7. **Document Insights:** Document insights ethically with participant consent, recording discussions anonymously to accurately capture gendered inputs and experiences.

8. **Collect Findings:** Collect findings and verify summary insights with participants to ensure alignment with diverse gender perspectives in the development of AI solutions.

9. **Craft a Pitch Deck or Concept Note:** Document insights ethically with participant consent, recording discussions anonymously to accurately capture gendered inputs and experiences.

10. **Implement Concept Testing:** Integrate an iterative testing of developed concepts into the design process, allowing for continuous feedback from participants to refine and enhance the gender-equitable AI solutions based on real-world usage and evolving needs.

Problem identification (cont.)

Resources:

Avanade AI Readiness Report
Artificial Intelligence Governance Toolkit
Artificial Intelligence (AI) Ethics Guide
AI Ethics Lab AI Principle Toolbox
Cisco AI Readiness Assessment
Decision Tree for the Responsible Application of Artificial Intelligence
Diversity and Inclusion Toolkit
IAPP Global AI Legislation Tracker
IDEO Field Guide to Human Centered Design
Map Your AI Use Cases by Opportunity
Microsoft Feminist Design Tool
NetHope AI Suitability Framework
NetHope Center for the Digital Nonprofit IDEA Journey
Participatory AI for humanitarian innovation
Participatory foresight: Preventing an impact gap

Learning Exercises:

Artificial Intelligence and legal issues
AI Fundamentals for Non-Data Scientists
Assess your organization's digital capabilities
Assess your organization's digital skills
Assessing Feasibility of Using ML/AI
Gender Justice: An Introduction
NetHope AI Suitability Framework Flashcards
NetHope How Might We Guide
Perform exploratory data analysis

Data collection is where the team consolidates different data sources, either collected internally or from external sources.

Questions to ask:

1. What internal data sources are available?
2. What external data sources should be integrated, and how reliable are they?
3. How do these data sources need to be processed for use?

Principles to follow:

- Gender-Centricity
- Explainability
- Intersectionality
- Privacy by Default
- Inclusivity

Checklist:

□ **Build capability** for statistical analysis of bias, variance, or test results. This can be done by training staff on data science principles which are fundamental to identifying variables potentially indicative of discriminatory outcomes related to gender.

□ **Invest in data governance** by investing in architecture internally to harmonize data formats, structures, and standards for consistency. See NetHope's Data Governance Toolkit.

□ **Ensure safeguards** to collect, store and label personal data lawfully, transparently, and securely with permission. Such safeguards must clearly explain the purpose of data gathering and types of questions in informed consent procedures.

□ **Create business intelligence process** with tools like Tableau, Power BI, Datalab for visually interactive uncovering of dataset issues. Pair these solutions with tooling and reusable templates for quantifying data set quality through statistical validation techniques.

□ **Assess existing data** from programs and headquarters examining the data integrity through the lens of quality, consistency, and reliability. This can be done by accessing dataset health metrics like provenance, completeness, timeliness, as well as usability and fit.

□ **Develop collection strategies** to gather and augment existing data with times, locations, and technologies convenient for participants from the community. Simultaneously, investigate integration approaches for augmenting existing datasets with newly acquired data.

□ **Recruit data labelers** from target community with local cultural/linguistic competency. Ensure at least half of enumerators are women. Moreover, ensure that respondent metadata and design data schemas link gender, race, location etc. for intersectional analysis.

□ **Establish data cards** that disclose gender composition details in training datasets, evaluation methodologies adhering to standards avoiding opacity. Integrate these into the model cards for the development process.

□ **Conduct gender disaggregated analysis to understand gender dynamics in the program context where AI will be applied. This can be done by establishing minimum thresholds for demographic representation and employing statistical techniques, including intersectional analysis, to uncover gender dynamics within the program and data context.**

□ **Create Feedback Mechanisms** both internally and externally to gather qualitative feedback from affected populations and rectify potential issues related to data collection. Be sure to include statistical validation and bias detection as part of continuous verification practices.

Data collection (cont.)

Key Step: Gender Disaggregated Analysis for AI / ML

Conducting gender disaggregated data analysis during data collection is imperative. It involves separating collected data by gender categories like male/female/unspecified and digging deeper to uncover systemic differences between various groups and identities. For example, analyzing if certain groups have access to more quality data features or if sensitive attributes like income level correlate differently with gender groups. This allows for proactive identification of direct and indirect discriminatory biases against gender groups before models are even trained, when bias mitigation is more straightforward.

1. **Gain Alignment:** Ensure a shared understanding of gender-related taxonomy to prevent confusion among data providers, collectors, and users. Recognize and mitigate gender bias in data collection methods, addressing underreporting or misreporting influenced by gender norms.

2. **Frame the Questions:** Collect information from men and women. Collecting sex-disaggregated data requires not only identifying the right respondents but also asking them questions about themselves and/or others.

3. **Collect the Data:** Tailor data collection methods to fit the specific context. Adjust survey questions to match the unique situation and priority populations. Those gathering and analyzing the data should have a good understanding of gender roles and social dynamics.

4. **Label and Annotate:** Label each data record with gender information (male/female/non-binary) by directly asking individuals to self-identify during surveys, interviews. Also collect gender composition at group levels.

5. **Determine the unit of analysis:** Identifying the appropriate unit of analysis is crucial as it impacts the granularity of insights. The unit of analysis shapes the context and influences the effectiveness of interventions. For instance, understanding individual crop planting choices may require interviewing farmers, while a household-level analysis can provide a comprehensive view of both crop planting and consumption.

6. **Apply the Chosen Statistical Tests:** Apply statistical tests to assess and validate the collected data. Use methods such as hypothesis testing for label consistency, Kappa scores for rater agreement, and other relevant statistical analyses to ensure the robustness of the gender-disaggregated data.

7. **Conduct an Intersectional Analysis:** Perform intersectional analysis to examine correlations of gender with other key attributes such as race, age group, etc. Understand how multiple factors intersect together and affect outcomes within the data.

8. **Preserve Documentation:** Preserve timeline documentation around data cleaning, labeling, analysis etc. to assist future audits. Continually monitor for new gender imbalances.

9. **Share and Report Out:** Share an analysis report with project teams, highlighting biases found, and recommend data augmentation, oversampling, etc., to address identified skews.

Data Collection (cont.)

Resources:

Accion International investing in equitable ai for inclusive finance
CARE Good Practices Framework for Gender Analysis
CARE International Rapid Gender Analysis
CARE Participatory Data Analysis Workshop
CGIAR Standards for collecting sex-disaggregated data for gender analysis
Data on Gender-Equitable Healthcare Accessibility
FHI 360 Gender and ICT Survey Toolkit
IFRC Handbook on data protection in humanitarian action
Mercy Corps Sex disaggregated data collection guide
OECD Mainstreaming and Implementing Gender Equality Toolkit
UNICEF Guide to Conducting Gender Disaggregated Analysis
USAID Tips for Conducting a Gender Analysis at the Activity or Project Level
Solutions to Close Gender Data Gaps

Learning Exercises:

Business Analytics with Excel: Elementary to Advanced
ChatGPT for Beginners: Save time with Microsoft Excel
Data Analysis with Python
Get Started with Microsoft Forms
Use SurveyMonkey to Create a Survey and Analyze Results
Gender-Disaggregated Data Analysis
Practical Statistics (Descriptive Statistics)
Understanding and Applying Text Embeddings

Data preprocessing is where the team cleans data and labels data (in supervised learning), preparing it to be used by the model.

Questions to ask:

1. What are the data sources, how representative are they, and do transformations need to be made?
2. How are we validating that the processed data leads to a model that performs well for all gender groups?
3. Can outsiders to the project understand the decisions made during data processing, including choices related to gender?

Principles to follow:

- Bias Mitigation
- Value Maintenance
- Model Awareness
- Human Oversight
- Localization

Checklist:

□ **Ingest data** from available sources like internal databases, APIs, or web scraping while ensuring collection aligns with problem needs. The quality and relevance of the ingested data directly impacts subsequent data processing steps.

□ **Visualize the data** by computing basic statistics, such as mean and standard deviation thereby assisting in understanding and identifying potential outliers. This is crucial for uncovering gender-specific patterns and ensuring that the data adequately represents local gender scenarios.

□ **Identify quality issues** through statistical methods like z-score analysis, constraint satisfaction checks. Statistical methods identify inconsistencies or inaccuracies related to gender variables, allowing for corrective actions, and promoting fair outcomes.

□ **Clean the data** by handling missing values via imputation, smoothing noise using binning, outlier capping. Addressing missing values ensures that gender-specific information is complete and unbiased, while noise reduction and outlier capping contribute to refining the quality of gender-related features.

□ **Transform the data by methods like min-max scaling for normalization, hot encoding for discrete variables, or dimensionality reduction. Normalization ensures that different gender-related features are treated equally, while encoding discrete variables helps capture gender without introducing bias.**

□ **Reduce the data and follow minimization principles** using techniques like feature selection, discretization, aggregations without losing model performance. Anonymize direct user input throughout planned data sets via techniques like encryption, hashing, masking to protect privacy.

□ **Split the data** into training, validation, and test sets to ensure that the future AI model is robust and generalizes well to diverse gender-related scenarios. A balanced split helps in evaluating model performance across different segments and works to avoid overfitting to specific gender groups.

□ **Record metadata** with a data schema or data dictionaries to inform cleaning approaches. This ensures that everyone involved in the project, including non-technical stakeholders, understands the origin, structure, and limitations of the data.

□ **Incorporate feedback** into cleaning sensitive data through interactive guided workflows. Human reviewers can provide context and nuanced understanding, especially in cases where automated processes might struggle with sensitive gender-related data, preventing unintentional reinforcement of biases.

Data preprocessing (cont.)

Key Step: Data Transformation for AI / ML

An essential precursor to model creation is analysis of the prepared dataset using an intersectional approach to unveil potential gender biases. This involves scrutinizing data distributions and labeling across gender groups while also considering intersections with attributes like race and income. For instance, assess whether specific demographic groups have equitable access to quality data features or experience accurate labeling. Continuous application of techniques such as fairness criteria testing, subgroup performance comparisons, and distribution divergences is imperative to gauge and address biases. Any identified issues should guide data augmentation and cleaning strategies before model creation and training, preventing the inheritance of bias into the models. Early identification of imbalances is crucial to avoid entrenching discriminatory behaviors within the developed models.

1. **Ensure Variable Assignment:** Assign at least three categories (male, female, unspecified) to each data record using surveys or voluntary self-ID mechanisms. Design these mechanisms based on guidance from gender experts to capture diverse gender identities comprehensively.

2. **Localize Gender Concepts:** Account for cultural notions of gender categories by geographically localizing data collection instruments and methodologies based on recommendations from gender experts.

3. **Evaluate Representation:** Use statistical methods to assess the completeness of the gender variable. Test if observed gender category probabilities align with population percentages. This step ensures a representative and inclusive dataset.

4. **Address Imbalances:** Mitigate gender imbalances in labeled datasets by employing augmentation techniques. Oversample minority gender groups to achieve a more balanced representation, enhancing future model fairness.

5. **Identify Key Features:** Profile distributions of key features (e.g., age, income) within each gender grouping. Utilize statistical tests like ANOVA significance test presented in AIF360 toolkit's example bias detection notebooks for auditing imbalances.

6. **Evaluate Proxy Variables:** Calculate gender representation ratios between other identities and features. Analyze whether any skews compared to population demographics are present. Explore if these skews correlate with other attributes, signaling potential bias in the dataset.

7. **Evaluate Metrics for Fairness:** To ensure equitable representation, set maximum permissible divergence thresholds between key features and dataset distributions across gender groups. AIF360 toolkit and FairLearn provide concrete examples, emphasizing metrics such as demographic parity, equalized odds, and equal opportunity.

8. **Create Data Nutrition Labels:** Maintain detailed model cards on entire data preparation process followed using established data ethics report templates outlining key fairness measures established.

Data processing (cont.)

Resources:

A gentle introduction to ML fairness metrics
An Analysis of Data Cleaning Tools
Common fairness metrics from FairLearn
Data featurization in machine learning
How to use AI and personal data appropriately and lawfully
Feature Engineering for AI and ML
NetHope Data Governance Toolkit
Ultimate Guide to Cleaning Data with Excel and Google Sheets
What-if Tool from Google
The Ultimate Guide to Basic Data Cleaning
From Data Governance to AI Governance
IFRC Data Playbook

Learning Exercises:

Analyze User Research Data with Microsoft Forms
Creating Fairness Metrics for Data Evaluation
Data Analysis with OpenAI API: Save time with GenAI
Data Balancing with Gen AI
Data Cleaning Tutorial
Data Cleaning Challenge: Scale and Normalize Data
Data Cleaning Challenge: Handling missing values
Data Science Challenge
FairLearn Walkthrough
Feature Engineering for AI and ML
United Nations Women Guide to Communicating Gender Data
University of Southern California Data Cleaning Guide

Model Creation is the core technical step where the team selects and develops potential models using the preprocessed data. Data is split into a training set for building the models and test set for validating the models.

Questions to ask:

1. Which model architecture / functional capabilities align with the problem domain and gender objectives?
2. What fairness metrics are integrated and measured, such as disparate impact or demographic parity?
3. Is documentation available on decisions made to evaluate gender equity in model development?

Principles to follow:

- Gender Centricity
- Accountability
- Gendered Centered Design
- Privacy by Default
- Capacity Building

Checklist:

□ **Maintain thorough documentation of the model building process, highlighting gender-related considerations. Develop model cards to communicate details, biases, and fairness measures, with dedicated sections on gender equity for transparency.**

□ **Conduct literature analysis** on recent models designed to minimize algorithmic harm, especially in the context of gender bias. Hold focused discussions with experts to gain insights, educate staff, and generate innovative ideas tailored to the problem statement.

□ **Evaluate and select** model architectures that inherently account for gender fairness and equity considerations. Aligning the architecture with the identified problem and promoting unbiased results across gender categories is essential for equitable AI solutions.

□ **Choose foundational models** that have demonstrated fairness in gender-related contexts. Fairness-tested foundational models serve as a solid starting point for the development of AI / ML solutions by ensuring that gender-related biases are minimized.

□ **Implement development practices** aligned with sustainability and open-source resources during the model development phase. Selecting algorithms with fairness and interpretability considerations ensures that the model creation process is iterative and future-proof.

□ **Assess model behavior** across individual and population-level fairness metrics, incorporating gender-specific metrics. Assess gender representation ratios and potential skews related to other attributes and validate results against test data sets.

□ **Perform ethical hacking** with a specific focus on gender considerations or harm. Creating an independent group, known as the "red team," simulates adversarial or critical perspectives to identify vulnerabilities, weaknesses, or overlooked aspects in the created model.

□ **Encode humanitarian principles**, including "do no harm," into model utility functions and constraints.

□ **Ensure automatic optimization** aligns with gender-equitable values, avoiding narrow metrics that might perpetuate gender biases.

□ **Involve staff from field sites** particularly those attuned to gender-related issues in the selection of relevant architectures and foundation models. Participatory approaches refine model behaviors based on diverse experiences and gender-focused perspectives.

Model Creation (cont.)
Key Step: Model Cards for AI / ML

Model cards are essential documents in AI, outlining details from data preprocessing to performance metrics. For gender equity, they play a crucial role by explicitly documenting considerations related to gender biases, fairness metrics, and subgroup performance. By providing transparency and accountability, model cards empower creators to actively monitor, rectify biases, and ensure AI models are equitable across diverse gender categories.

1. Connect to other documentation from the design process, utilizing established reporting processes to log all data preparation, model development, and evaluation steps. This ensures traceability of all decisions made during model development to the core problem.

2. Articulate use cases that explicitly define target demographics and key success metrics aligned with humanitarian objectives. Provide a detailed overview in the model card.

3. Disclose model limitations, emphasizing interpretability, computational requirements, and data limitations that influenced architectural selections, fostering transparency. Aligning use cases with humanitarian objectives and defining target demographics ensures that the model addresses the right problem.

4. Report performance metrics across representative identity subgroups, encompassing gender, income, and geography. Quantify model disparities using bias testing toolkits and be sure to include the methodologies and outcomes in the model card.

5. Document fairness metrics, including specific metrics that measure the model's performance across different gender subgroups. Include these metrics in the version-controlled documentation to provide transparency on how gender equity considerations are actively monitored and addressed in the model's development lifecycle.

6. Periodically publish updates with disaggregated performance benchmarks, including gender and geography-specific metrics, promoting transparent AI principles.

7. Disclose data processing details to encompass cleaning and labeling techniques, fairness criteria thresholds, and acceptance testing measures, ensuring total transparency in the model card.

8. Outline opt-out mechanisms allowing affected individuals to appeal algorithmic decisions, providing reasonable explanations, and ensuring prompt redressal, with details outlined in the model card.

9. Seek external algorithm reviews and public feedback on model cards before deployment, actively listening to marginalized voices to address any overlooked issues and reinforce accountability.

Model Creation (cont.)

Resources:

Algorithms and Audit Basics
AI Fairness Checklist
AI Fairness 360 Toolkit
AI Ethics Lab: Toolbox for Evaluation
Aequitas fairness audit Toolkit
Bias in Algorithms
Eticas Algorithmic Auditing Guide
Catalogue of Tools & Metrics for Trustworthy AI
Microsoft AI learning hub
PwC Responsible AI Toolkit
Model Card Toolkit
TensorFlow Responsible AI Toolkit
Google's The Value of Shared Understanding of AI Models
Microsoft's Fair Learn Whitepaper and Community Space

Learning Exercises:

Creating Model Cards
Developing Model Cards
How to Build an AI: The Chronicles of Jobot
How to Make AI in Python Tutorial
Intro to Machine Learning
Machine Learning Introduction for Everyone
Mini-Project: Make a Machine Learning App

Model Evaluation is where models are tested against predefined criteria (such as accuracy, performance, etc) to determine which approach is best suited for the problem.

Questions to ask:

1. How frequent are algorithmic audits for gender bias?
2. How is model evaluation applied to datasets representing identities, ethnicities, and backgrounds?
3. What method is used to aggregate fairness metrics into a comprehensive fairness scoring system?

Principles to follow:

- Gender-Centricity
- Explainability
- Model Awareness
- Human Oversight
- Capacity Building

Checklist:

□ **Prioritize model architectures** for both fairness and accuracy and leverage multiple evaluations. When possible, opt-for modular and interpretable architectures over opaque Blackbox solutions with little documentation or explainability.

□ **Evaluate available models** trained with algorithmic bias mitigation approaches like adversarial debiasing specifically for efficacy on reducing disparities across gender groups. Well-intentioned techniques still warrant verification in the context of gender identity.

□ **Compare candidate models** not just based on overall accuracy but also key identity variables (e.g., gender, income level, geography) through intersectional analysis, preventing masked harms.

□ **Establish evaluation criteria** such as accuracy, F1 score, precision, and recall, which provide essential benchmarks for assessing the performance of AI models. They offer a quantitative means to measure the model's overall accuracy, its ability to balance precision and recall, and its effectiveness in correctly identifying and mitigating biases, particularly in the context of gender considerations.

□ **Quantify fairness metrics** including variance and uncertainty of computed performance measures. Pay particular attention to gender groups prone to underrepresentation, ensuring robust reproducibility. This comprehensive analysis contributes to the model's trustworthiness and effectiveness in diverse gender contexts.

□ **Quantify accuracy metrics** and favor multiple gender-sensitive indicators. Insist on using representative sample sizes to avoid underpowered experiments. Adopting nuanced evaluation metrics and robust sample sizes ensures a more accurate understanding of the model's performance, particularly in gender-specific contexts.

□ **Apply explainability methods** to establish which models rely on variables potentially relating as proxies to indirect bias. Proxies could provide clues to mitigate issues without having to perform opaque internal diagnoses of models.

□ **Conduct algorithmic audits assessing model fairness from an end-to-end perspective, spanning data collection, processing, and usage. Provide directions for improvement based on audit outcomes to enhance fairness in the model.**

□ **Leverage ethics committee** for review as well as program and field leaders from around the organization. Track results and analysis from the review process and investigate ways to improve model performance and fairness.

Model Evaluation (cont.)

Key Step: Algorithmic Audits for AI / ML

An algorithmic audit is an examination of a model's design, implementation, and outcomes to assess its fairness, transparency, and potential biases. This process involves reviewing the entire lifecycle of the algorithm, from data collection and processing to model training and deployment. It aims to uncover and rectify any disparities or discriminatory impacts on different demographic groups, ensuring equitable results. Algorithmic audits are crucial for fostering transparency, accountability, and trust in artificial intelligence systems, helping organizations identify and address ethical concerns while promoting fairness in algorithmic decision-making.

1. **Select Auditors:** Engage a proficient external auditing firm or specialized consultants with proven ML fairness expertise for an independent algorithmic audit.

2. **Initiate Kickoff:** Launch the audit process with formal kickoff meetings, fostering collaboration among engineering, product, and ethics board members. Clearly define the audit scope, address preliminary queries, and set expectations for information access.

3. **Share Documentation:** Have development teams share pertinent documentation, including model cards, data sheets, and reports showcasing internal evaluations. This provides a comprehensive foundation for the external audit.

4. **Enable Access:** Provide auditors with access credentials and necessary permissions for all infrastructure components, from version control and model registries to CI/CD analytics dashboards used in development workflows.

5. **Schedule Interviews:** Coordinate thorough audit interviews with engineering and operations teams across hierarchy levels, gaining insights into data pipelines, and organizational processes.

6. **Evaluate Results:** Scrutinize detailed audit analysis results, risk rankings, and identified issues. Formulate responses, clarifying constraints and proposing actionable remedies for addressable instances after a cost-benefit assessment.

7. **Document Mitigations:** Work collaboratively with auditors to document mitigation strategies for identified issues. Develop a comprehensive plan outlining steps to address biases, improve fairness, and enhance overall algorithmic performance.

8. **Implement Recommendations:** Track the implementation status of auditor recommendations in periodic governance meetings. Ensure planned software updates and process changes are delivered as expected within agreed timelines.

9. **Integrate Results:** Incorporate major audit recommendations and discovered best practices into organizational policies and model development checklists. Enhance institutional safeguards around commercial AI deployment based on audit learnings.

Model Evaluation (cont.)

Resources:

Algorithms and Audit Basics
Algorithmic Auditing: The Key to Making Machine Learning in the Public Interest
AI Fairness Checklist
A DIY Approach to Algorithmic Audits
A Guide to ICO Audit
Artificial Intelligence Auditing Framework
Aequitas fairness audit Toolkit
Guide to Algorithmic Auditing
Explicability of humanitarian AI: a matter of principles
Evaluating Machine Learning Methods
Eticas' Guide to Algorithmic Auditing
Peer review framework for predictive analytics in humanitarian response
PwC Responsible AI Toolkit
TensorFlow Responsible AI Toolkit
Using Algorithm Audits to Understand AI

Learning Exercises:

AI, Business & the Future of Work
AI Fairness Exercise
Building Trust: Ethics for AI-powered Chatbots
Data Ethics, AI and Responsible Innovation
Data Science Ethics
Reviewing the Danish Refugee Council's Foresight Model
Exploring Algorithmic Bias as a Policy Issue: A Teach-Out
Identifying Bias in AI

Model Tuning and Training is where appropriate threshold values and hyperparameters are set to optimize the model's performance.

Questions to ask:

1. What mechanisms are implemented to balance fairness and accuracy with performance?
2. How is the incorporation of domain expertise into the model's decision-making process being carried out?
3. In what ways are iterative feedback loops available to stakeholders?

Principles to follow:

- Bias Mitigation
- Accountability
- Model Awareness
- Privacy by default
- Capacity Building

Checklist:

□ **Choose a tuning approach** for hyperparameter tuning. Tuning involves adjusting the settings of an AI model, much like custom settings on an iOS. By choosing methods such as grid search, random search, or advanced techniques like Bayesian optimization, developers can optimize the model's performance with a specific focus on promoting gender equity.

□ **Sample Data** to improve model performance on underrepresented groups in imbalanced datasets. Correcting imbalanced data mitigates bias against minority gender groups.

□ **Determine amount of training data by using the "rule of 10" (have 10x data points as model parameters) or analyzing learning curves (model accuracy plotted against dataset size for diminishing returns)**

□ **Input the training data** into model by ensuring adequate representation across demographic groups. Inclusive foundational data enables equitable algorithmic insights.

□ **Assess model accuracy** with test data sets that are unseen by the model, comparing to overall metrics. Independent testing by gender quantifies parity in capability.

□ **Validate model performance** on datasets proportionally representing gender demographics served. Checking accuracy on peer groups affirms gender fair generalization.

□ **Tune model hyperparameters** by using factors like like batch size and learning rate with validation data sets that mirror real-world demographics of program services. Representative validation sets prevent overfitting and improve generalizability across genders.

□ **Implement regularization methods** like dropout and early stopping to reduce overfitting and improve generalizability across demographics. Regularization yields models better suited for varying gender cohorts.

□ **Implement adversarial validation** to proactively search for problematic failure from the model in use cases across demographics. Adversarial testing from humans in the loop unveil hidden gender bias.

□ **Evaluate proxy variables** like geography or socioeconomic status introduce hidden demographic biases into the model. Remove proxy variables if needed. Eliminating proxy variables reduces implicit gender biases.

□ **Develop rollback procedures** for discontinuing model use if unfair biases emerge that cannot be sufficiently resolved through other technical approaches. Rollback protocols enable responding to intractable gender biases.

Model Tuning and Training (cont.)

Key Step: Calculating Training Data for AI / ML

Calculating the amount of training data is vital for developing gender-equitable AI / ML solutions. The proper amount of training data ensures that the model comprehensively learns diverse gender-related patterns, as well as any proxy connections, to minimize bias and improve fairness. Insufficient data may lead to skewed representations and reinforce disparities. Accurate representation in training data thus becomes fundamental to allowing models to make informed predictions across demographic groups. Striking the right balance in training data size not only improves model performance but reinforces ethical development principles promoting inclusivity in the deployment landscape.

Method 1 – Rule of 10

1. **Define Model Complexity:** Assess the model's complexity in terms of parameters and architecture, crucial as more complex models demand larger datasets to capture nuanced gender-related patterns effectively.

2. **Apply Rule of 10:** Set a minimum of 10 times the number of parameters as the target training data size, ensuring sufficient diversity to address the complexities of gender dynamics.

3. **Evaluate and Adjust:** Assess model performance, and if needed, acquire additional data to meet the rule, fostering a robust foundation for accurate and equitable predictions.

Method 2 – Linear Regression and Diminishing Returns

1. **Gather Data and Split:** Collect a dataset and split it into training, validation, and test sets, foundational for comprehensive learning across gender demographics.

2. **Train Model with Varying Sizes:** Train the model using different training set sizes, observing performance on a validation set, enabling an understanding of how data volume impacts gender-equitable model development.

3. **Identify Diminishing Returns:** Analyze the learning curve to identify the point of diminishing returns in model performance, aiding in determining the optimal balance between data volume and model efficacy.

Method 3 – Statistical Power Analysis

1. **Define Significance Levels:** Set significance levels for desired confidence in accuracy and fairness, contributing to the statistical rigor of gender equity assessments.

2. **Determine Effect Size:** Identify the effect size, representing practically significant differences, critical for gauging the model's sensitivity to gender-related nuances.

3. **Use Power Analysis Formula:** Employ the statistical analysis formula based on significance levels, effect size, and power level, providing a quantitative calculation for the data needed in model training.

4. **Calculate Required Sample Size:** Determine the minimum sample size needed to achieve desired statistical power, ensuring adequate representation for gender groups.

Model Tuning and Training (cont.)

Resources:

- 5 steps to train and test your AI algorithm
- Evaluate the performance and accuracy of machine learning models in analytics projects
- Diminishing Returns in Machine Learning
- Fairlearn: A toolkit for assessing and improving fairness in AI
- Fine Tuning with OpenAI
- How to use AI and personal data appropriately and lawfully
- Generative AI is Here: Are you ready?
- How Much Data Do We Need
- Hyperparameter Optimization Explanation
- Is your dataset big enough?
- Telus International's Guide to AI Training Data
- Training Your Own AI Model Is Not As Hard As You (Probably) Think
- The essential guide to AI training data
- The 5 Stages of Machine Learning Validation

Learning Exercises:

- AI Fairness Exercise
- ChatGPT Prompt Engineering for Developers
- Fine-tune a pretrained model
- Hyperparameter tuning a machine learning model
- Hyperparameter Optimization Exercise
- Hyperparameter Tuning tutorial
- Finetuning Large Language Models
- Identifying Bias in AI
- Introduction to Prompt Injection
- Prompt Engineering with GPT: Programming for Custom Content

Deployment is where the model is used in real-world applications, ideally starting with a smaller beta testing phase.

Questions to ask:

1. What measures should be taken to ensure equal access to and participation?
2. How can the deployment process be strategically planned and optimized for efficiency?
3. What approach should be taken to promote cohesion with the program's existing infrastructure?

Principles to follow:

- Gender-Centricity
- Accountability
- Gender Centered Design
- Autonomy
- Inclusivity

Checklist:

☐ **Establish deployment gates** before full production release. Staged gates enforce rigorous standards at each deployment phase, including gender fairness measures.

☐ **Choose deployment locations** based on scalability needs, cost, and data sensitivity. Cloud solutions (e.g., AWS, Azure) offer flexibility and rapid scalability, while on-premises provides more control, crucial for sensitive data or strict regulatory requirements.

☐ **Ensure problem centric deployment that is intricately aligned with the specific challenges faced by country programs and services. Tailor the AI/ML model to directly address the identified issues, promoting solutions that resonate with the programmatic goals and local contexts.**

☐ **Regularly document model changes**, maintain a clear changelog, and use it as a reference for future iterations. Model versioning enables reverting to previous models if needed and promotes team collaboration and understanding.

☐ **Containerize models** with tools like Docker to package models and their dependencies into containers. Utilize container orchestration for simplified deployment, scaling, and operations of application containers across clusters.

☐ **Create data supply agreements** ensuring continued access to necessary third-party data sources post-launch. Supply agreements provide data stability required for upholding gender fairness over time.

☐ **Catalog all inputs and outputs** to define system scope and compliance needs. Defining data flows precisely clarifies regulatory applicability regarding use of potentially sensitive information derived from data on gender identity.

Deployment (cont.)

Key Step: Address Core Porblems with Gender Equitable AI / ML

Leveraging gender-equitable AI solutions enables nonprofit organizations to tackle core problems like climate change, poverty, and food distribution with enhanced efficiency and fairness. By integrating these advanced technologies, nonprofits can identify and implement strategies that address the specific needs and challenges of all genders, ensuring equitable access to resources, decision-making, and benefits. This approach not only fosters inclusivity but also amplifies the impact of their interventions in these critical areas.

1. **Partner with Sector Leads**: Partner with sector leads, such as those in climate, poverty, or food distribution, to garner support for AI / ML solutions addressing their specific problems. By aligning with sector champions organizations can leverage influence to drive awareness and secure resources.

2. **Collaborate with Country Programs:** Engage with country and field programs to unravel core problems by employing methodologies like the '5 Whys?' and identifying shared challenges within the organization. This collaborative approach connects their problems with available models for deployment.

3. **Ensure Deployment Compliance:** Verify that the deployed models comply with local regulations and data privacy laws governing the country programs and services where they will be utilized. Evaluate the ethical implications of deploying the models in different contexts, ensuring fairness, transparency, and accountability in decision-making processes.

4. **Deploy Available Models:** Apply pre-existing models from internal repositories for a quicker response to core problems. By using existing models developed with equity in mind, organizations can address gender biases and promote fair outcomes. This approach saves time and resources while benefiting from proven models that align with the organization's goals.

5. **Develop Operational Documents:** Create operational documents that incorporate model cards, insights from subject matter experts, and other available documentation. These documents, including concise two-page use documents and process flows, serve as vital guidance for effectively deploying AI/ML models within the specific context of program use.

6. **Demonstrate Efficiency:** Demonstrate outcomes without significant financial investments, fostering a cost-effective approach to integrating solutions into programs dedicated to gender equity. Repurposing existing hardware is one example that lowers costs and accelerates the integration of AI into ongoing initiatives.

7. **Empower Local Management:** Empowering local management ensures that individuals, regardless of gender, have the knowledge and skills to engage with AI solutions. This inclusivity fosters gender equity by democratizing access to AI tools and promoting diverse participation in their utilization.

8. **Connect Funding to AI / ML Development:** Establish a direct link between program funding and the management/development of AI/ML models addressing core gender-equitable issues. This ensures sustainable application, aligning financial support with the ongoing evolution and impact of AI solutions.

Deployment (cont.)

Resources:

8 Machine Learning Model Deployment Tools That You Need to Know
15 Best Tools for ML Experiment Tracking and Management
Deploying Artificial Intelligence Whitepaper
Find the Right Pace for Your AI Rollout
Guidelines for Procurement of AI Solutions by the Private Sector
Guidelines for Trustworthy AI
How to use AI and personal data appropriately and lawfully
How we approach responsible AI at Microsoft
Humanitarian AI - The hype, the hope, and the future
IFRC Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks
Infrastructure Strategy to Support your AI Solution
Managing Machine Learning Projects in International Development
Under the Hood of AI

Learning Exercises:

A practical learning exercise for Git
Building and Deploying Machine Learning Models
Code and develop a python app in a container
Machine Learning for Developers

Evaluation involves constantly checking the model to make sure it works as intended, revisiting earlier phases and/or retraining the models when new data comes in..

Questions to ask:

1. Are resources sufficient to maintain and update the AI model over time?
2. How can we identify and mitigate issues arising from concept drift, data shifts, or changes in gender-related patterns?
3. Are there plans to collaborate with peer NGOs and domain experts to gather feedback and scale?

Principles to follow:

- Accessibility
- Value Maintenance
- Model Awareness
- Human Oversight
- Localization

Checklist:

□ **Execute a scaling plan** by incorporating lessons learned from the beta test phase and ensuring scalability for a broader user base. Monitor closely and address any issues during the initial stages of full deployment, with ongoing consideration for gender equity and diverse user experiences.

□ **Continually evaluate** the tradeoffs between achieving model accuracy, parity in performance across genders, and other ethical objectives to determine if technical advances make it feasible to better balance performance.

□ **Budget for regular algorithmic audits** by specialists in identifying algorithmic harms to measure efficacy of group bias, review representation in data sources, and formally verify core model documentation.

□ **Establish a process** for escalation when decline in model performance indicates a need for full retraining. Defined escalation procedures enable prompt responses to address declining metrics before increased gender biases emerge.

□ **Regularly monitor performance** of the model through key metrics disaggregated by gender. Leverage the fairness metrics used during model development to systematically assess and address any emerging biases that disproportionately impact one gender.

□ **Set a minimum frequency** for incorporating new data, such as quarterly. Regular data refreshments help adaptation to change over time before performance declines or bias increases.

□ **Establish simple mechanisms** for end user feedback and report failures or harms. User feedback helps identify real-world failures affecting individuals that may have gender implications.

□ **Review incidents and model failures.** Employ analytical methods to determine whether inequitable impacts are contributing to the identified problems. This analysis is crucial for addressing and rectifying any unintended biases in the model's performance.

□ **Archive versions of data and models** to allow rollback if problems emerge. Archiving supports comparison and rollback for diagnosis of emergent gender-related issues.

□ **Evaluate the feasibility of open sourcing parts of the model or the entire system to encourage collaboration and scrutiny from the wider community. Open-source practices can enhance transparency and contribute to uncovering potential biases or ethical concerns.**

Evaluation (cont.)

Key Step: Evaluate Open Source for Equitable AI / ML

Embracing open-source models in AI / ML development – while not always possible – is imperative for fostering sustainability, scalability, and collaboration. Open-source frameworks not only encourage the sharing of knowledge and resources but also facilitate the reuse of existing solutions thereby minimizing redundancy and maximizing efficiency. This approach promotes sustainable practices by reducing waste because of duplicated efforts. Furthermore, the collaborative nature of open source allows project stakeholders to contribute their expertise to other organizations, leading to more impact across the humanitarian sector. Additionally, the transparency inherent in open-source frameworks fosters trust and accountability, as the codebase is open for inspection and validation by a global community. To execute on open-source efforts, there are several options.

Method 1 – GitHub Repository

1. **Establish Central Repository:** Create a collaborative hub on GitHub by setting up a new repository. Initiate it with README for contextual information.

2. **Upload Models and Documentation:** Ensure accessibility and comprehension by developing detailed documentation. Emphasize the model's purpose, architecture, and usage guidelines.

3. **Cultivate Collaboration:** Facilitate code sharing with version control for continuous enhancements. Upload organized code, including pre-trained weights, and employ Git for version tracking.

Method 2 – TensorFlow and Hugging Face

1. **Leverage Established Platforms:** Maximize visibility and accessibility by signing up on platforms like TensorFlow Hub or Hugging Face.

2. **Upload Models and Documentation:** Expand the model's reach within the community. Follow platform-specific instructions to upload the model, ensuring proper metadata and tagging for discoverability. Provide thorough documentation directly on the platform, ensuring users have all necessary information.

3. **Cultivate Collaboration:** Cultivate a collaborative ecosystem for knowledge exchange and continuous improvement. Engage with the platform's community features, respond to inquiries, and actively participate in discussions.

Method 3 – Dedicated Project Website

1. **Choose a Public Platform:** Provide a seamless platform for professionals by acquiring a domain and choosing hosting for a dedicated website.

2. **Upload Models and Documentation:** Facilitate easy navigation and comprehension of model use and impact. Develop informative sections for documentation, tutorials, case studies, and community discussion.

3. **Cultivate Collaboration:** Empower users with knowledge on model usage and impact. Cultivate collaboration and inquiry on the potential use of the model.

Evaluation (cont.)

Resources:

AI Ethics Lab: Toolbox for Evaluation
Business Model Sustainability Toolkit
Catholic Relief Services and USAID Monitoring and Evaluation Guide
Ethical Explorer Risk Fieldbook
Evaluating Metrics for AI Impact Quantification
Foresight Manual: Empowered Futures for the 2030 Agenda
Google – Perspective on Issues in AI Governance
Introduction to Open-Source Initiatives
Microsoft – Governing AI and a Blueprint for the Future
NIST Artificial Intelligence Risk Management Framework
Ready Set Scale: A primer on Scaling AI for Business Value
Scaling AI to generate better and different outcomes
Scaling AI: Cost and Performance of AI
UNDP Monitoring and Evaluation Training Guide
WEF Empowering AI Leadership Toolkit

Learning Exercises:

Avoiding AI Harm
Developing AI Policy
Exercise: Model Pipelines
Ethical AI in Student Loan Approvals
Managing Machine Learning Projects

NETHOPE

Contact

If you have any questions about this toolkit, its methods or the data used, contact NetHope's Center for the Digital Nonprofit directly at:
nethopecdn@nethope.org

nethope.org