Artificial Intelligence (AI) Ethics
Workshop for Nonprofits

December 2020

Facilitator Guide

About this Workshop

The purpose of Artificial Intelligence (AI) Ethics Workshop for Nonprofits is to provide participants with the information necessary to design and use AI in a responsible and ethical way.

The workshop is focused on the practical application of the concepts and frameworks covered in the first three webinars of the AI Ethics webinar series. Workshop participants will have the opportunity to explore questions surrounding the principle of Fairness in the context of several use cases from the humanitarian and international development sector, including education, health, agriculture, workforce, and humanitarian response.

The workshop was developed and tested by NetHope, USAID, MIT D-Lab, and Plan International.

Learning Objectives

In the workshop, participants will:

- Understand the potential risks of designing and using AI in international development contexts.
- Learn how to practically apply ethical considerations and implement AI responsibly. Practicing asking the key questions and examining ethical considerations related to the principle of Fairness is indeed the main purpose of this workshop.

Target Audience and Delivery

- This workshop is designed to be applicable to any role in international development, with a basic understanding of AI.
- This workshop is designed to be delivered virtually but the facilitator could adapt for in-person delivery. We recommend up to 40 attendees for a virtual workshop (eight participants per breakout).

For Workshop Facilitators

- Facilitators do not need to have a significant AI background. This workshop is meant to serve as a learning journey for the facilitator, too.
- Facilitators should review the workshop materials and the following resources in preparation for the workshop: AI Primer webinar and AI Ethics webinars.
- For the virtual workshop, it's important to have somebody responsible for managing breakouts (ie sending participants to breakouts and bringing them back to the main room) in addition to five facilitators for five breakouts.
- If you are a NetHope Member and would like to continue the conversation about AI and AI ethics after the workshop, please join NetHope's AI Working Group (a group of global NGOs that collaborate on all aspects of AI in the nonprofit sector). To join, please complete this form.

Workshop Materials and Setup for in-person and remote

For workshops delivered remotely:

- Online meeting platform that allows sharing slides, audio and, ideally, video of the facilitator (examples include Zoom, GoToMeeting, Skype, Teams, Webex, UberConference, and many others). A meeting platform with in-built breakout rooms will make managing the workshop easier.
- Create a shared document for each breakout that participants can contribute to in real-time.
- We recommend Miro. There is also Miro Lite.
- We recommend five to ten participants per breakout. Ideal number is eight.

For workshops delivered in-person:

- Sticky notes and pens for participants to capture questions, concerns, resources and ideas.
- Flip-chart or whiteboard labeled "Idea Board" to share sticky notes.
- Printed copies of framework questions to use in breakout group discussions. There should be enough for each person.
- Arrange a room with tables that each seat between five and ten participants and provide enough distance for each group to discuss and collaborate.

Facilitator Materials

In addition to the guide, the following materials form part of the toolkit developed to support facilitators in delivering this workshop:

- Artificial Intelligence (AI) Ethics for Nonprofits - Toolkit
- Workshop deck
- 5 case studies with facilitator notes
- Key AI ethics concepts
- Key AI/ML concepts
- Example Miro board

Participant Materials

This workshop should be accompanied by the following participant materials:

- AI Primer
- AI ethics webinars Please review the recordings and slides for all three webinars
- Key AI ethics concepts
- Key AI/ML concepts

Note: If you are hosting this workshop at a conference where you do not know the list of participants in advance, you may need to follow up with these resources. If you are organizing the workshop yourself, you can send these as a pre-read.

AI Ethics workshop was envisioned and produced by Leila Toplic (NetHope), Amy Paul (USAID), Amit Gandhi (MIT D-Lab), Kendra Leith (MIT D-Lab), and Nora Lindström (Plan International).

AI Ethics Workshop Agenda

TIMING	MODULE	KEY TOPICS	SHOW
[10 mins]	**Introduction**	• Welcome • Introductions. Note: This is to introduce the workshop facilitators. Participant introductions will take place in breakout groups. • Purpose of this workshop/what to expect	Slides 1-6
[25 mins]	**Overview of key AI ethics concepts**	• What is AI ethics? Define Values, Principles, Frameworks. • What is responsible innovation? What is an ethical and humane AI solution? • What are the biggest ethical questions surrounding technology use today? • What are Bias and Fairness? • What are the key considerations relevant to the principle of Fairness? • What can we do to mitigate concerns? • What are some of the other considerations? • Brief overview of a process for developing a Machine Learning project. What are the key questions to ask along the way?	Slides 7-26
[15 mins]	**What can go wrong? How do you address the issues?**	Chatbot case study	Slides 27-32
[10-15 mins]	**Q&A**	Q&A	Slide 33
[5 mins]	**Break**		Slide 34

[10 mins]	**Introduction to breakouts**	Overview of breakouts, introduction to Miro tool, and transition to your breakout room DEMO: Miro	Slides 35-38 For Miro demo, please see Miro demo below
[75 mins]	**Breakouts**	• 10min: Introduction to the breakout (case study, questions) and Miro tool • 5min introduction: 1 sentence per participant/use the sticky notes functionality in Miro • 5-7min: Participants read the case study and ask for any clarifications • 3-5min: Prompt - What are some of your reactions to the case? • 10min: Participants write responses to predetermined questions • 30min: Discussion • 5min: to consolidate feedback and get ready to present *Use sticky notes in Miro for brainstorming throughout the discussion*	Please see Breakouts Guide below
[2 min]	**Transition from breakouts to report-out & introduction of the case studies**	Transition to the main workshop and overview of the next segment Note: This may be a good time for participants to take a quick bio break.	Slide 39-41
[20 mins]	**Report out**	Each group shares a brief summary of their key insights in 2min. Brief discussion	Please show Miro boards for each breakout.
[3 mins]	**Wrap Up**	• Resources • Next steps *Note: You may consider having a longer conclusion for your workshop.*	Slides 42-44

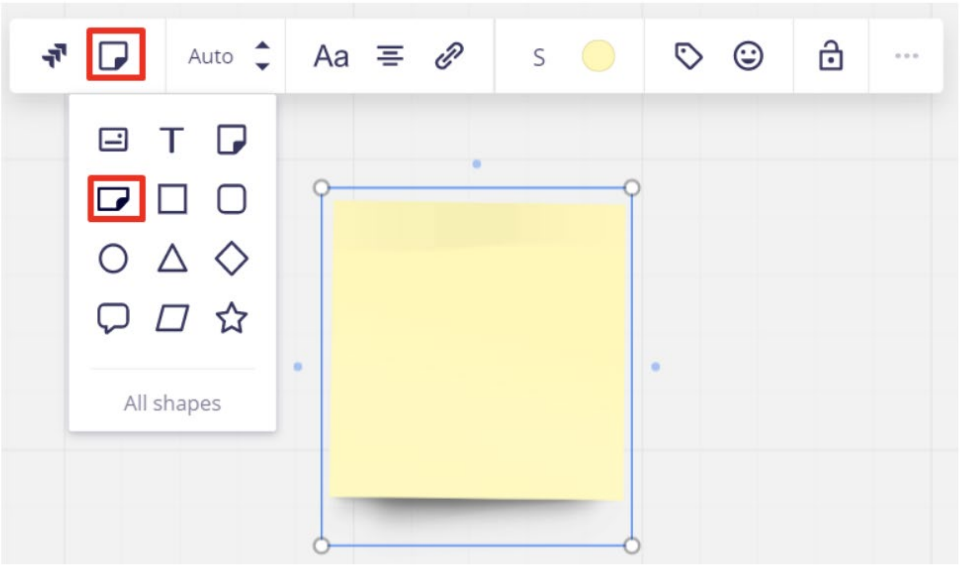
NOTE: Workshop host will provide a demo of Miro prior to sending the participants to breakouts. Breakout facilitators should be able to answer any questions about Miro and make sure that the participants know how to use it.

Miro board

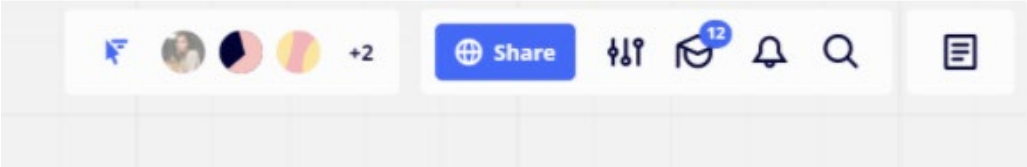
Say: Let me show you all you need to know about Miro to participate in this breakout – how to create a sticky note, how to add your comments, and how to move the sticky note around the board. Don't worry, we'll do a quick exercise in each of the breakouts before we get started with the case studies.

Show how to create a sticky note: Click the sticky note icon on the toolbar or press N on your keyboard to enable the tool. Type some text (your name, organization, role, country, and food) and move the sticky note around on the board. For more information about sticky note functionality, please review this article prior to the session.

*Say: To create a new sticky note, click the sticky note icon on the toolbar or press N on your keyboard to enable the tool. If you use the smallest font size, it is possible to enter about 350 symbols in a square note. You can also select the size for your sticky – I recommend to use either **S**(mall) or **M**(edium) so we can fit many sticky notes in the workspace area (that is section #4).*



Cursors: *Also, as we all start moving around the board, it can get visually very busy - you can click on the cursor icon (to the left of everyone's icons) at the top of the board to disable the viewing of everyone's cursors.*



Zooming: *And, if you'd like to see the whole board or quickly find a different section of the board, you can zoom out using the navigation in the lower right corner of the board. You can also select 'fit to screen' (|<-->| sign to the left of '-' sign) to see the whole board.*

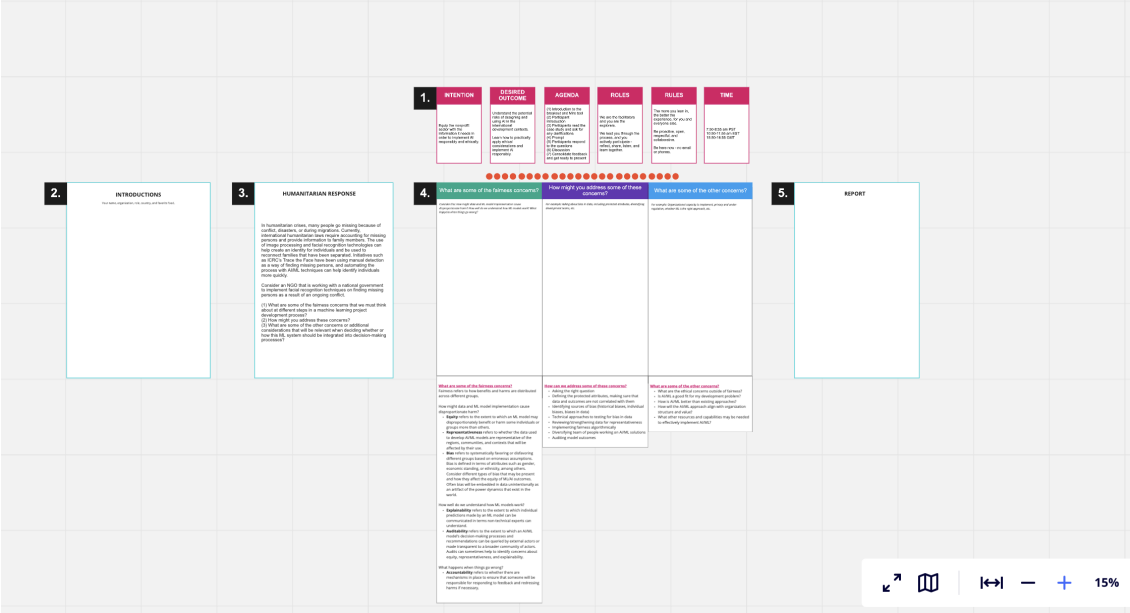
Bring to me: *You'll also see me use 'Bring everyone to' which ensures that we're all looking at the same place on the board.*

All you need to remember is how to create a sticky Note and how to move around the board.

All this will be much clearer once you get started with Miro. And, if you have any questions about using Miro, be sure to ask your breakout facilitator.

AI Ethics Workshop for Nonprofits: Breakouts	
ACTIVITY	**SHOW**
Welcome and Introductions Remember to **turn your video on** during the introduction and to **start recording** the breakout if you planned to record the breakouts. [Click Record in the meeting controls to start a local recording. Keep in mind that local recordings are stored on the computer that initiated them. When you record Zoom meetings locally, each meeting recording is saved on your desktop device to its own folder labeled with the date, time, and meeting name. By default, these folders are inside the Zoom folder, located inside the Documents folder on Windows, macOS, and Linux. You may want to set up a folder where facilitators can upload the recordings.] Welcome the participants and introduce yourself **again**. Remind participants about the purpose of this breakout. <u>Say:</u> *Welcome to the [name of your breakout] breakout. This breakout is designed to provide you with an opportunity to explore the questions surrounding the principle of Fairness in the context of a use case from [focus area of your breakout]. The emphasis here is not on a specific use case or your familiarity with this particular topic (eg agriculture) but on the practical application of the fairness considerations in the nonprofit programs.*	Zoom screen [1min]
Flow Provide a high-level summary of the flow of this breakout session. **Start sharing your screen and show the Miro board. Make sure to zoom out so you can show the whole board.** Note: We recommend to share your screen with the Miro board only when you want everyone to be looking at the board (ie during the discussions and demos), and to stop	Miro board [4 min]

sharing your screen during the quiet time when people are actively working on the board. This helps avoid confusion with multiple windows.



Show section 1 of the Miro board.

<u>Say</u>: We will be using Miro in this breakout as the facilitator shared in the main room. Before you get started with using Miro, I'd like to share with you what we have planned for this breakout and we have that mapped out in this Miro board.

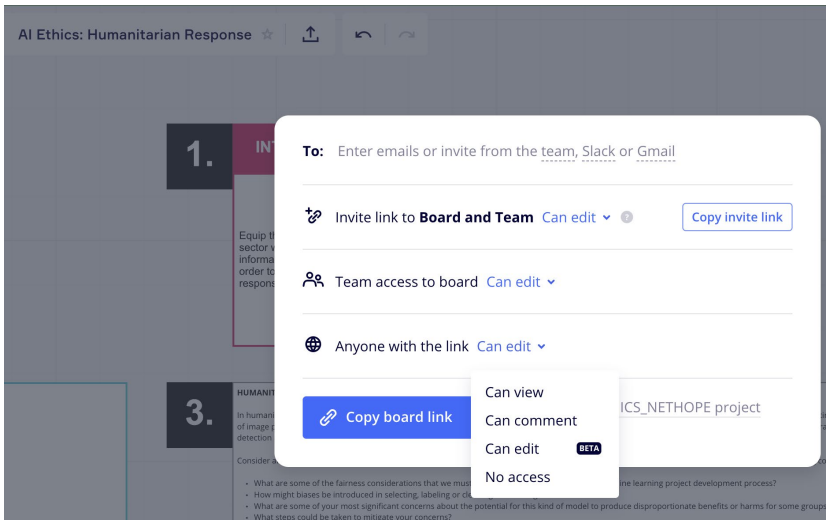
<u>Say:</u> IDOARRT (Intention, Desired Outcomes, Agenda, Roles, Rules, Time) for this breakout is similar to the IDOARRT that XX [the workshop moderator] shared in the introduction to the workshop. Three main differences are the Desired Outcome, Agenda for this breakout, and Timing.

<u>Say:</u> The Desired Outcome for this breakout is for you to learn how to begin to practically apply ethical considerations and implement AI responsibly. We will be focusing on the principle of Fairness.

<u>Say:</u> In terms of the Agenda - here's what you can expect in this breakout.

- *First, we'll take a couple of minutes for introductions. We'll use Miro for that.* **Show section 2 of the Miro board.**
- *Then we'll dive into the case study.* **Show section 3 of the Miro board.**
- *You'll use sticky notes to share your perspective and we'll have 20 minutes for a discussion.* **Show section 4 of the Miro board.**

○ *If at any point in time you need a quick refresher on the Key Concepts like Fairness, Representativeness, or some of the mitigation steps - take a look at the Key Concepts under this workspace. This is the same information you received in email a couple of weeks ago.* **Zoom out to show both section 4 and concepts under the workspace.** • *At the end of the breakout, we'll take a few minutes to consolidate the key insights to be shared in the main room.* **Show section 5 of the Miro board.** ○ *Your observations and ideas shared in this Miro board are the starting point for the discussion that we'll have in this breakout. They will also inform future workshops and training materials. We will consolidate key insights from all breakouts and send in an email as a follow up to the workshop.* *Before we get started, I'd like to ask for a volunteer who will report out key insights from our discussion. Who would like to volunteer? … Thank you very much!*	
Miro board **Please share in the Zoom chat window Guest access to the board for your breakout:** Note: Make sure to share Miro board links with the facilitators prior to the workshop. • Health: **Insert link to Miro board** • Education: **Insert link to Miro board** • Agriculture: **Insert link to Miro board** • Workforce: **Insert link to Miro board** • Humanitarian Response: **Insert link to Miro board** If participants have any issues with accessing the board, you can give them Guest access to the Miro board on your end. Open the Share window and choose the level of access for *Anyone with the link can edit* (or Public link to board). Choose edit access. Please make sure to review this article prior to the workshop.	Miro board (create Sticky) [5min]



<u>Say:</u> Please join me in the Miro board - click on the link in the Zoom chat window. Let me know if you have any issues with accessing the board.

<u>Ask:</u> ***Is everyone in the Miro board? Does anyone need help with getting started with sticky notes?***

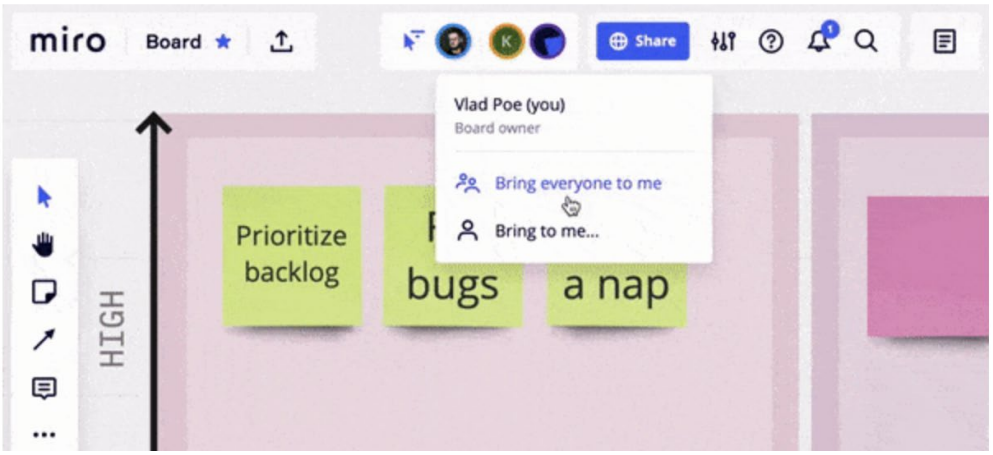
Participants start using Miro

Have participants introduce themselves – their name, organization, role, and country - and respond to an ice-breaker question. (This is especially helpful in mixed sessions where people are not as familiar with one another.)

Bring all participants to the area of the board you want everyone to focus on with *Bring everyone to me* functionality. You can use this functionality throughout the session. For more information, please review this article.

Miro board [Introductions section]

[5 mins]



Say: Please use a sticky note to quickly introduce yourself - your name, organization, role, and country. And, for fun, include the answer to this question: If you could only eat one food for the rest of your life, what would it be? ***Please choose one of the colors and plan to use the same color for the rest of the breakout.***

Say: Introduction should take no more than 1min.You'll each have 20 seconds to share with the group.

Set the timer for 1min - timer is available in the Collaboration Toolbar in the bottom left corner.

Ask for a volunteer - who wants to go first. Call on each participant to share their introduction in 20 seconds. This will give everyone a chance to say something at the beginning and get ready for the discussion later.

Case study

Say: In the next 5-7min, we will:

- *Read through the case study quietly*

Miro board [Case study section and Key Concepts

• *Clarify any questions you might have about what you read.* *Note that this case study is modeled on real-world scenarios from organizations that are either actively implementing or considering how to best implement ML-based interventions in international development.* Set the timer for 3 minutes. *<u>After 3min - Ask</u>: Now that you've read the case study – is there anything that was unclear or confusing?*	section] [5-7 mins]
Prompt question - Optional *<u>Say</u>: Before we dive into the questions, shall we take a few minutes to hear your first impressions of the case - specifically, what are some of the concerns you have about the case?* Ask for a volunteer - who wants to go first? If there are no volunteers or the conversation slows down, you might want to politely call on specific people to participate. Or, to move to the next segment.	Miro board [Case study section] [3-5 mins]
Participants respond to questions using Sticky notes *<u>Say</u>: Thank you for sharing your observations! Now, we'll use an approach to brainstorming that is called 'brain writing' which is a quieter activity with individual work time followed by group discussion and collaboration.* *Please use sticky notes to respond to these 3 questions. The concerns you just identified might overlap with some of these, and if so, that's fine. Feel free to reuse or refine your initial concerns as you consider each question.* ○ *Questions:* ○ ***What are some of the <u>fairness</u> concerns that we must think about at different steps in a machine learning project development process?*** ▪ *Here you might consider some of the following questions: How might data and ML model implementation cause disproportionate harm? How well do we understand how ML models work? What happens when things go wrong?* ○ ***How might you address these concerns?*** ▪ *For example: Asking about bias in data, including protected attributes, diversifying development teams, etc.*	Miro board [Case study, 5 questions section] [10 mins]

○ ***What are some of the other concerns or additional considerations*** *… that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?* ▪ *For example: Organizational capacity to implement, privacy and under-regulation, whether ML is the right approach, etc.* ○ *As mentioned earlier, do consult the Key Concepts document that you received in email or the same information is available under this workspace.* **Make sure to zoom out to show the key concepts under segment #4** ○ *You can respond to the questions in any order.* ○ *You can have as many sticky notes as you'd like.* ○ *You're welcome to work on your sticky notes off to the side and bring them to the board when you're done.* ○ *We'll do this quietly for 9 minutes.* *Ask: Are there any questions before we get started?* **Set the timer for 9 minutes.**	
Discussion *Say: The time is up! Great to see so many interesting observations. Now, before we dive into the discussion, I'd like to encourage you to consider using the orange dots that are located above the segment #4 to 'mark' the comments and observations that you think are important. You can do that during the discussion, and we'll leave 1 minute at the end of the discussion to allocate your dots. This will also help us select key insights to report out in the main room.* ***Note that each person has 3 dots.*** **Set the timer for 30 minutes.** *Say: Who'd like to start by sharing some of the fairness concerns that we must think about at different steps in a machine learning project development process?* *Say: How might you address these concerns?* *Say: What are some of the other concerns or additional considerations… that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?* To keep the conversation flowing, you might want to politely call on specific people to participate. Keep track of participation and if someone hasn't had a chance to say	Miro board [Case study, 5 questions section] [**30** mins]

something yet, encourage them to contribute. Ask specific questions to specific participants. And, encourage participants to keep their remarks brief and to the point. After the time is up: *Say: Now, we'll take 1 minute to mark the most important insights with the orange dots. Just a reminder, each person has 3 dots.*	
Consolidate key insights *Say: We have a few minutes left to consolidate key insights to share with other groups. Reporter - could you take the lead in selecting the top insights based on the number of dots? You can right-click on the sticky note to create a* ***Duplicate*** *and move it over to the Report section.* Help the volunteer breakout reporter consolidate the insights in the Key Insights section of the board. *Ask the group: Is there anything missing from our list of top insights?* *Ask: Are there any questions as we transition into the final few minutes of this breakout?* Before you leave the breakout, please make sure to save the breakout Chat. [To save the meeting chat, click on the 'More' option (three dots) on the right. Select the 'Save chat' option from the menu that pops up. The chat will be saved on your computer as a text file. Please upload the chat to **this folder** immediately after the workshop.]	Miro board [Key Insights section] [10 mins]

Artificial Intelligence (AI) Ethics
Workshop for Nonprofits

December 2020

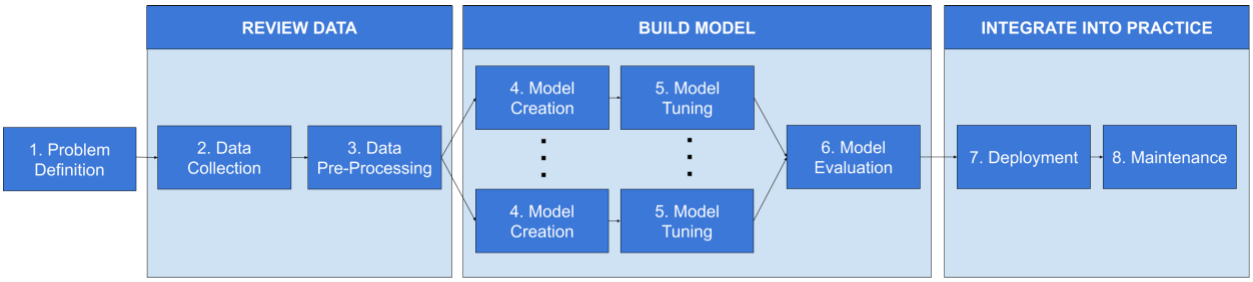
Key AI Ethics Concepts

Overview

- AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies. (Alan Turing Institute)
- Values are broad beliefs held by individuals or groups that reflect concepts of social and cultural importance and norms of appropriate behavior. Ethical values define our moral conduct and help us to determine what is right or what is wrong. An example of a value is Integrity, another one is Respect.
- Principles guide responsible innovation by providing direction for how to embed values in design and use of solutions. Example principles: Fairness, 'Do No Harm".
- Ethical technology solution is a solution that supports individual and collective well-being and enhances our ability to tackle global challenges.
- Responsible Innovation is a transparent, interactive, sustainable process by which organizations proactively evaluate how they can design and use technology in ways that are aligned with their values and missions.

Overview of the Machine Learning Process

The basic steps for the ML process are shown below - in practice, the process is iterative, and steps are revisited until the desired outcome is achieved.



1. **Problem definition** 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 tuning** is where appropriate threshold values and hyperparameters are set to optimize the model's performance.
6. **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.
7. **Deployment** is where the model is used in real-world applications, ideally starting with a smaller beta testing phase.
8. **Maintenance** 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.

Ethical considerations with the focus on principle of Fairness

What are some of the fairness concerns?

How might data and ML model implementation cause disproportionate harm?

- **Equity** refers to the extent to which an ML model may disproportionately benefit or harm some individuals or groups more than others.
- **Representativeness** refers to whether the data used to develop AI/ML models is representative of the regions, communities, and contexts that will be affected by their use.
- **Bias** refers to systematically favoring or disfavoring different groups based on erroneous assumptions. Bias is defined in terms of attributes such as gender, economic standing, or ethnicity, among others. Consider different types of bias that may be present and how they affect the equity of ML/AI outcomes. Often bias will be embedded in data unintentionally as an artifact of the power dynamics that exist in the world.

How well do we understand how ML models work?

- **Explainability** refers to the extent to which individual predictions made by an ML model can be communicated in terms non-technical experts can understand.
- **Auditability** refers to the extent to which an AI/ML model's decision-making processes and recommendations can be queried by external actors or made transparent to a broader community of actors. Audits can sometimes help to identify concerns about equity, representativeness, and explainability.

What happens when things go wrong?

- **Accountability** refers to whether there are mechanisms in place to ensure that someone will be responsible for responding to feedback and redressing harms if necessary.

How can we address some of these concerns?

- Asking the right question
- Defining the protected attributes, making sure that data and outcomes are not correlated with them
- Identifying sources of bias (historical biases, individual biases, biases in data)
- Technical approaches to testing for bias in data
- Reviewing/strengthening data for representativeness
- Implementing fairness algorithmically
- Diversifying team of people working on AI/ML solutions
- Auditing model outcomes

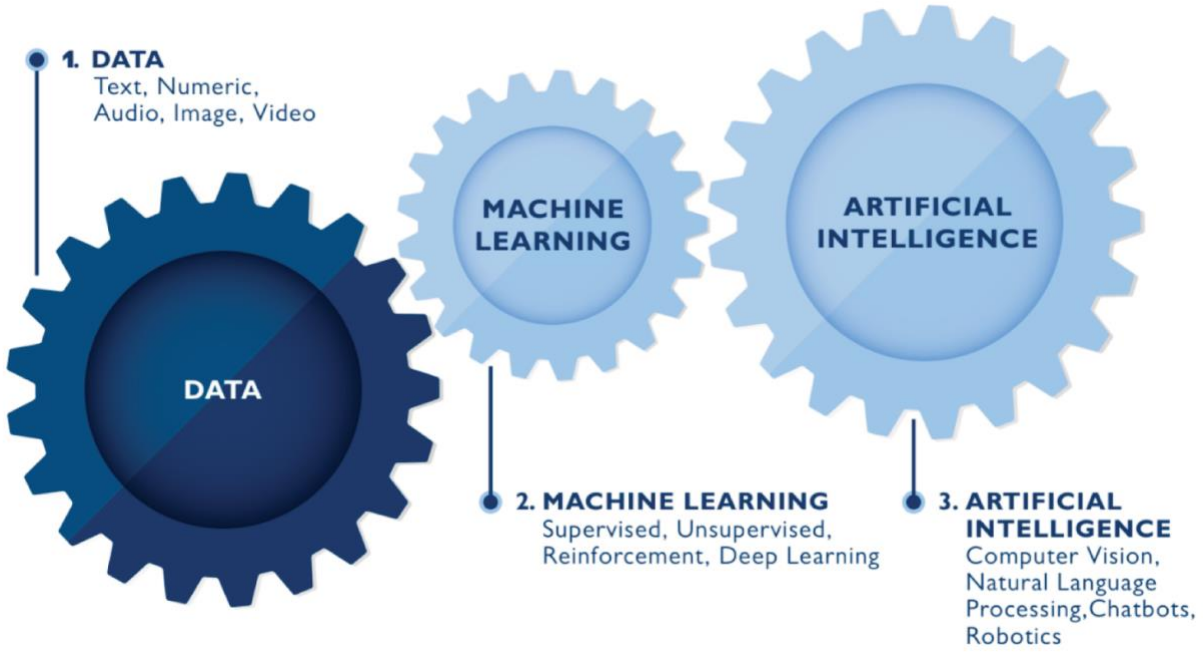
What are some of the other concerns?

- What are the ethical concerns outside of fairness?
- Is AI/ML a good fit for my development problem?
- How is AI/ML better than existing approaches?
- How will the AI/ML approach align with organization structure and value?
- What other resources and capabilities may be needed to effectively implement AI/ML?

Artificial Intelligence (AI) Ethics
Workshop for Nonprofits

December 2020

Artificial Intelligence and Machine Learning: Key Terms and Definitions



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, you could think of ML as "data-driven predictions."

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 in online shopping); 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."

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.

Types of Machine Learning

Supervised learning: Given a set of labeled training data, learn to predict labels for unlabeled data.
Estimate the probability of loan repayment based on financial data from past borrowers.

Unsupervised learning: Find patterns or structure in a dataset
Determine whether potential borrowers comprise several distinct groups, for which different loan products could be designed.

Reinforcement learning: Reward-based training system, maximizing its chances of achieving a well-defined goal
Currently most useful for robotics and autonomous vehicles (and Go)

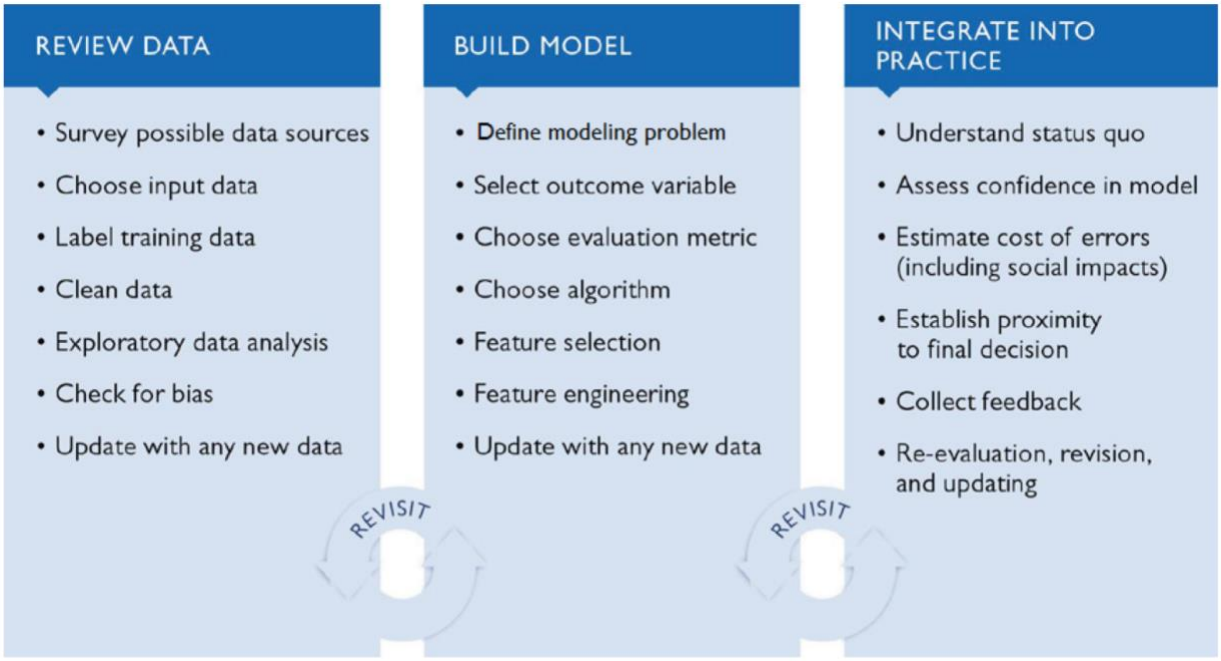
Deep learning is part of a broader family of machine learning methods based on artificial neural networks. **Artificial neural networks (ANN)** are computing systems that are inspired by, but not identical to, biological neural networks. They can be trained to match inputs to specific outputs by adjusting parameters within the neural net.
Can be used for supervised, unsupervised, or reinforcement ML

Key AI Capabilities

- **Natural language processing (NLP)** analyzes or synthesizes "natural" human languages such as English, Spanish, or Arabic.
- **Computer vision** processes images or video in order to identify objects or interpret scenes or events.

- **Speech or audio recognition** analyzes audio files to recognize specific sounds or speech patterns. Speech recognition often relies on NLP to transcribe speech into written text.
- **Advanced Analytics** carries out sophisticated analysis of multiple data sources, structures.
- **Content Generation** creates new text, images, video from understanding of key patterns in training text, images, video.

Developing and using AI / ML: It's a process



Artificial Intelligence (AI) Ethics
Workshop for Nonprofits

December 2020

Case Studies

Table of Contents

Case Study: Health

One significant challenge in controlling HIV/AIDS is ensuring that people living with HIV/AIDS are retained on treatment. A significant portion of people who know they are positive are "lost to follow up" in the first 12 months of their treatment - meaning they don't show up for regular check-ups to monitor adherence to treatment, manage side effects, and address any co-morbidities. Health systems - through case managers or community health workers - spend significant resources tracking down patients lost to follow up. If it were possible to more precisely predict which patients are most likely to be lost to follow up, health system resources could be better targeted to prevent it - concentrating resources where it will have the most impact and improving overall retention.

Consider an organization piloting a machine learning based model to better identify which patients attending their health facility are most likely to be lost to follow up. They plan to use patient and facility level data (eg patient comorbidities, clinic attendance times, behavioral patterns, overall clinic performance metrics), as well as data about events in the surrounding community, to understand the risk factors driving loss to follow up and identifying high-risk patients. They would use the results to know to which patients they should target retention resources, as well as inform the design of other interventions that might reduce root causes of the most significant risk factors.

Questions:

- What are some of the fairness concerns that we must think about at different steps in a machine learning project development process?
- How might you address some of these concerns?
- What are some of the other concerns or additional considerations that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?

Facilitator Notes

How might model design or use cause disproportionate harm?

Prompt: Think through what happens if the model gets it "wrong"? In this case, what would happen if the model identifies someone as high risk for loss-to-follow up who really isn't? Or, misses someone who is?

Equity

- If model results are tied to resource allocation - adherence support -- a false positive result could end up "pestering" those who do not need reminders, causing additional **stigma** of being HIV + (because you'll get a lot of reminders) **or resentment** (of systems that don't let you live your life without reminder/interruption).
- Alternatively, the model could falsely predict someone who needs reminders to be low risk (a false negative), meaning **supportive resources would not be directed to someone who needs them.**

- Depending on how models are used, using the results to be the sole determinant of adherence support resources could end up draining all resources from "mostly adherent" patients in a way that is felt as a harm. Eg "Low risk" patients may not need much support, but value the amount that they do get and as a result of the new ML-approach to targeting adherence support, they receive less support than they

did before the machine learning approach was initiated. It would be up for discussion as to how important that is - it may not result in worse outcomes (the patients may still remain adherent), but their patient experience may not be as satisfactory.

Prompt: How might bias be embedded in training data and lead to model failures? What concerns might you have about data used to develop this model?

Representativeness

- If men are less likely to attend clinics than women, it is possible that data sets have a minority representation of men, and a ML model may not predict loss-to-follow-up (LTFU) as well for them -- meaning that HIV+ men may have worse outcomes than women in terms of treatment adherence and control of the virus. Those incorrectly predicted to be low risk for LTFU (false negative) may subsequently get fewer resources than they would otherwise.
- Risk factors for LTFU may be quite different for men than women, for older people than younger -- may need to make separate models for groups that will have different risk factors for LTFU
- Data quality issues across clinics might misrepresent the adherence patterns of patients altogether

Prompt: Are there ways in which the model might reinforce or create a social bias? What would we need to consider with respect to social implications of a model like this?

Bias

- If model findings are generalized, could create a stereotype of a person who doesn't follow through on treatment and is considered "lazy" or "irresponsible." Would need to be careful that the findings from the model don't create new social bias around people perceived to be at high risk for loss-to-follow-up.

How well do we understand how models work?

- *We don't know much about the specifics of the model here. How important do you think it is to be able to interpret a ML model in this case?* (eg to identify which variables are most contributing to a prediction of high or low risk of LTFU)? Why?
    - If you want to be able to design interventions, it would be important to know what the drivers are and you'd want a high degree of interpretability.
    - On the other hand, if you really just need to save resources and the biggest problem is targeting them to the people most in need, you might trade off some to have a more accurate model. But you would need to consider how much additional accuracy you might get.

What happens when things go wrong? (Redress for possible harms? Feedback mechanisms?)

- *How do you think the organization should respond to errors? Do you see any risks related to their ability to know when mistakes are made and respond when they do?*
    - Relates to Accountability
    - We should ask people getting adherence support how much they value it, why, what would happen if they don't get it
    - Continue to track LTFU data - identify whether you are losing more people than expected, whether there are unintended consequences for those who used to receive support and don't after implementation and consider whether this is useful given the resources you have
    - Evaluate the new ML model against the old way of doing things to evaluate how much value it is adding

What steps should we take to mitigate fairness concerns?

(These are suggestions of actions participants may come up with during discussion. The framing below attempts to align them to specific fairness considerations, but participants may not frame/label them that way and that's ok)

- **Representative data:**
  - Exploratory data analysis to understand who you have data on and who is missing
  - Explore data quality (are those marked as LTFU really lost or just changed to a different clinic) - data quality issues that might lead to bias
- **Unequal model performance:**
  - Measuring performance across groups - is it predicting LTFU better for men or women? certain age groups? geographies?
- **Model failures:**
  - Be intentional in deciding how to optimize the model **-** is it better to have a few false positives (providing support to those who don't need it) so that you don't miss anyone who does need it, or vice versa?
  - Have local community health workers, clinic staff participate in model design process
- **Explainability:**
  - Have a team of health system experts and data science experts work to find the right balance, decide how to use model results
- **Accountability:**
  - Get feedback from people living with HIV/AIDS (those affected by the model's results) on how the model results compare with their experience - do those who are target for reminders find them helpful? Are there people who feel like they are not getting the adherence support they need?
  - Evaluate against prior methods of adherence support

Other concerns

- Data sharing and privacy
- Organizational capacity to continue to implement
- Overall resource allocation questions in HIV response

Case Study: Workforce

Employers may receive dozens, hundreds, or even thousands of resumes for a particular position. It can be a daunting task to go through many resumes and select the correct candidates. Thus, some companies have started to use natural language processing and machine learning techniques to select applicants based on the qualifications that appear in their resumes (such as experience, skills, and degrees). Based on the output of the machine learning model, the applicant may or may not move on to the next round. Machine learning can also be applied to the interview stage to help evaluate candidates based on how they reply to questions and if they display characteristics such as warmth and patriotism, which can help predict whether they will be hired.[1]

Consider an employment organization deciding to implement a machine learning tool that would analyze data from resumes only (eg textual data included in a resume) to determine which candidates should be reviewed and who should be offered interviews.

Questions:

- What are some of the fairness concerns that we must think about at different steps in a machine learning project development process?
- How might you address some of these concerns?
- What are some of the other concerns or additional considerations that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?

Facilitator Notes

How might model design or use cause disproportionate harm?

Prompt: Think through what happens if the model gets it "wrong"? In this case, what would happen if someone who is qualified for the job is not offered an interview? Or someone who is not qualified being offered an interview?

Equity

- False positives (given interview when unqualified) are likely to be addressed later in the hiring process, though they may be representative of hiring biases and could result in false hires. Would be more concerning if the algorithm were misused to directly make hiring decisions.
- False negatives (not given interview when qualified) are more concerning because they remove potential skilled applicants from the job pool.

Prompt: How might bias be embedded in training data and lead to model failures? What concerns might you have about data used to develop this model?

Representativeness

- Past hiring practices could be biased - if specific protected attribute groups are more prevalent in the data (for example, gender, race, etc) then they may be favored by the algorithm.
- Particular wording or phrasing that is region-specific may be favored by algorithms.

[1] Teodorescu, M., Ordabayeva, N., Kokkodis, M., Unnam, A. and Aggarwal, V. 2020. Working Paper June 2020. Human vs. Machine: Biases in Hiring Decisions. Working paper.

Prompt: Are there ways in which the model might reinforce or create a social bias? What would we need to consider with respect to social implications of a model like this?

Bias

- Hiring has elements of subjectivity and is known to be biased in certain contexts. If the algorithm is based on data from a biased process, it is likely to reinforce those biases.

How well do we understand how models work? *We don't know exactly how the algorithm is determining who gets to move on to the next stage of interviews, or which metric is being used for evaluation of candidates. Is the algorithm looking at resumes to see who has been moved on to the interview stage in the past, who has been hired in the past, or who has performed well in the company in the past?*

Auditability

- This approach is difficult to audit because it is hard to collect data on the hiring and performance of individuals that were not given interviews.
- Having individuals at the organization look at resumes manually can be a good way to audit the algorithms.

Explainability

- Job applicants are typically not provided a reason why they weren't hired, so organizations may not look to develop algorithms that are explainable.

What happens when things go wrong? (Redress for possible harms? Feedback mechanisms?)

- Organizations should go through past data and determine if there were any false negatives. If similar positions are still open, reaching out to past applicants to offer interviews could help redress harms.

What steps should we take to mitigate fairness concerns?

- **Representative data:**
  - Assess current hiring practices to see if there are existing biases or inequities
  - Determine if data is representative across key protected attributes.
  - Build a more representative dataset by either balancing the training set across protected attributes or through the use of synthetic data.
- **Unequal model performance:**
  - Does the model favor specific groups? Men v. women? People from a specific region? Race?
- **Model failures:**
  - Should the model have more false positives or false negatives? False positives can burden the interviewers and slow down the hiring process, but false negatives can eliminate skilled applicants.
- **Explainability:**
  - Model decisions would ideally be explainable. Avoid using looks-like algorithms.
- **Accountability:**
  - Track performance of algorithms: do individuals go on to get hired? How do they perform?

Case Study: Agriculture

Some key agricultural NGOs support farmers by providing them with farming inputs including seeds and fertilizers. Farmers typically receive these inputs as loans (either through the NGO or through another financing agent) and repay them at the end of the growing season, depending on the success of the growing season. The NGOs can also help to connect farmers to markets to maximize their earnings. The logistics of where and when to provide seeds and fertilizer and how to effectively connect farmers to markets is a promising area that machine learning can be applied to. By using historical data from farmers, typical growing seasons, and regional market prices for different crops, machine learning algorithms can be used to improve farmer income.

Consider an NGO providing these services to smallholder farmers and cooperatives across Tanzania. Historical market data including prices, crop yield, and plant and harvest dates is available, but data from different areas is of different quality: higher quality data is available in larger, more populous areas, while data from rural areas is of lower quality.

Questions:

- What are some of the fairness concerns that we must think about at different steps in a machine learning project development process?
- How might you address some of these concerns?
- What are some of the other concerns or additional considerations that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?

Facilitator Notes

What are some of the fairness concerns that we must think about at different steps in a machine learning project development process?

How might model design or use cause disproportionate harm?

Prompt: Think through what happens if the model gets it "wrong"? In this case, what would happen if the model gives bad advice to farmers, who end up with additional losses?

Equity

- Could falsely predict when to apply fertilizer, harvest crops, and make connections to markets
  - Could apply fertilizer at the wrong time and could have a neutral or negative effect on the crops
  - Could sell produce too early or too late and not get a good price for the produce, which could limit income
- If accurate data is not readily available in rural areas, then the model could generate benefits for some groups more than others (peri-urban v. rural farmers or men v. women). There could also be variability on data collected by crop with some crops having more information than others, such as crops that are of higher value

Prompt: How might bias be embedded in training data and lead to model failures? What concerns might you have about data used to develop this model?

December 2020

Representativeness

- The model is not likely very representative of rural communities, which means that the model may not accurately predict the yields, harvest dates, and prices in these regions.
- Data may also not be collected in the same way (digital v. paper) and the data may be less consistent with the paper data collection or there could be more errors if data were collected by hand.
- Also, there could be a possibility that some of the data is missing due to loss of data or inability to collect data from certain areas during certain years.
- Also, if it turns out that men are more likely to respond and contribute data than women, then the model may not be representative of female farmers.

Prompt: Are there ways in which the model might reinforce or create a social bias? What would we need to consider with respect to social implications of a model like this?

Bias

- There could be bias in the data collected and bias in how the model was created
- This isn't one we usually talk about, but historical bias -- given climate change, the relationships of the past may not hold. Are they looking at very historical data or relatively recent? What impact might time have?

How well do we understand how models work? *We don't know much about the specifics of the model here. How important do you think it is to be able to interpret a ML model in this case?*

Auditability

- Not clear to the extent to which the model could be audited.
- Could be helpful to have outside group that can evaluate the model for bias issues if possible or compare different models that are trying to predict similar things

Explainability

- Model needs to be used by NGO and farmers to make decisions, so the results of the model need to be fairly easy to understand for a non-technical person. May have to have some capacity at the NGO to be able to interpret the model for the other staff members. Or could work with team of experts

What happens when things go wrong? (Redress for possible harms? Feedback mechanisms?)

- When auditing the model, you discover that there is bias toward peri-urban farmers and women, so NGO needs to work with technical and implementation staff to address these issues, so that model is more accurate and farmers are getting more accurate information that they need.
- If the NGO or other key stakeholders do not understand the model, then it could be challenging to communicate why the model should be trusted and what factors help determine the outcomes.

What steps should we take to mitigate fairness concerns?

- **Representative data:**
  - Identify who is missing through initial data analysis
  - Identify bias that may be occurring as result of the of data collected and model created
    - Data does not seem representative
    - Who created the model? More urban focused researchers? Men v. women? Could we improve the diversity of the team creating the model?
  - Strengthen representativeness of data

December 2020

- **Unequal model performance:**
  - Is the model more accurately predicting outcomes for urban farmers or rural farmers? Men v. women?
- **Model failures:**
  - Be intentional in deciding how to optimize the model **-** is it better to have a few false positives (providing support to those who don't need it) so that you don't miss anyone who does need it, or vice versa?
  - Potentially engage farmers and local NGO staff in designing the model
- **Explainability:**
  - Include someone on the NGO team who can help interpret the results for the NGO staff and farmers
- **Accountability:**
  - Get feedback from the farmers and the NGO staff
  - Evaluate against previous approaches for predicting this type of data

Other concerns

- Data sharing and privacy
- Organizational capacity to continue to implement

Case Study: Education

Educators are using machine learning technologies to automate the process of assessing student performance and adapting material to make it more accessible to individuals. Various machine learning tools can be used for different applications in the assessment process. For example, natural language processing has been used to assess student writing and regression analysis has been used to identify knowledge gaps based on student performance on written exams. Using these insights, the curriculum can be tailored to meet individual students' needs. This can include additional support for students that may be struggling with material or presenting more difficult materials for students who are doing well.

Consider the case where an educational organization is using machine learning to evaluate the quality of student writing for high school students across India. (Quality of writing in this case refers to the thoughts, ideas, and clarity of expression in words, as conveyed in typed format, not handwritten.) In order to ensure high standards, the typed writing samples used to train models are taken from top universities across the country. Looking forward, the organization also plans to use the ML model to make tailored recommendations about additional writing classes for students to take in order to improve their overall writing performance.

Questions:

- What are some of the fairness concerns that we must think about at different steps in a machine learning project development process?
- How might you address some of these concerns?
- What are some of the other concerns or additional considerations that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?

Facilitator Notes

How might model design or use cause disproportionate harm?

Prompt: Think through what happens if the model gets it "wrong"? In this case, what would happen if the model recommends additional classes to students who do not need them? Or, does not recommend them to someone who does?

Equity

- Given that the model's results are tied to the provision of additional writing classes, a false positive result could end up with students who do not need additional support being required to take remedial classes, potentially causing stigma or resentment, while also taking support resources away from those who do need support.
- Alternatively, it could falsely predict someone who needs additional support to be low risk, resulting in them not getting the support they need, or, recommend more difficult materials to students who are not yet at the appropriate level for them.

Prompt: How might bias be embedded in training data and lead to model failures? What concerns might you have about data used to develop this model?

Representativeness

- The training data for the model comes from top universities across the country, yet the model aims to assess writing quality among high school students. The model may therefore assess students for a higher level of writing than would be expected at their educational level. This may skew the results and suggest writing skills are lower than they are expected to be in the target group and that there is a higher need for additional support.
- Is writing at university level necessarily the best example of quality writing? Or does it represent only a particular type of quality writing? Using a more varied dataset of writing samples could improve the model. Ideally, the model would use high quality writing samples from high school students.

Prompt: Are there ways in which the model might reinforce or create a social bias? What would we need to consider with respect to social implications of a model like this?

Bias

- If the model consistently recommends additional support classes to certain groups but not others, based on protected attributes, it could stereotype particular groups as "less gifted" or "stupid" while others are seen as "smarter"
- This could happen if the writing samples used to train the model are not representative of a variety of good quality writing or if the algorithm weighs certain aspects of writing as indicating higher quality and these are more common in particular groups' style of writing than others (there can eg be differences in how girls and boys write).

How well do we understand how models work? *We don't know much about the specifics of the model here. How important do you think it is to be able to interpret a ML model in this case? (eg to identify which variables are most contributing to a prediction of high or low quality of writing)? Why?*

- Given that the model is tied to resource allocation and may lead to stigmatization, implies understanding how the model works and how it defines and assesses high quality writing would be important.
- If the model was used only as one element in assessing student needs, lower accuracy might be acceptable.

What happens when things go wrong?

How do you think we might best stay aware of when mistakes might happen and their consequences? What should we do to mitigate harms if/when they do occur?

- Track the predictions provided by the model and monitor students' progress - use model to assess student writing on a regular basis
- Ask for student and teacher feedback on model predictions
- Evaluate the new ML model against the old way of doing things to evaluate how much value it is adding
- Use model as only one element in assessing student needs, ensure relevant teachers make final decisions on students' needs (human in the loop)

What steps should we take to mitigate fairness concerns?

- **Representative data:**
    - Consider what kinds of writing samples would be most appropriate for assessing quality of writing among high school students specifically and how to ensure diversity across the samples used
- **Unequal model performance:**

    - Measuring performance across groups - is it predicting writing skills better for girls or boys? Certain geographies? Certain schools?
- **Model failures:**
    - Be intentional in deciding how to optimize model **-** is it better to have a few false positives (providing support to those who don't need it) so that you don't miss anyone who does need it, or vice versa?
    - Have teachers participate in model design process
- **Explainability:**
    - Have a team of educational experts and data science experts work to find the right balance, decide how to use model results
- **Accountability:**
    - Get feedback from teachers and educational experts on how the model results compare with their experience
    - Evaluate against prior methods of adherence support

Other concerns

- Data sharing and privacy
- Organizational capacity to provide additional support to students and tailor support to individual students' needs
- Overall resource allocation questions in education

Case Study: Humanitarian Response

In humanitarian crises, many people go missing because of conflict, disasters, or during migrations. Currently, international humanitarian laws require accounting for missing persons and providing information to family members. Image processing and facial recognition technologies can be used to uniquely identify individuals in order to reconnect families that have been separated. Initiatives such as ICRC's Trace the Face have been using manual detection as a way of finding missing persons, and automating the process with AI/ML techniques can help identify individuals more quickly.

Consider an NGO that is working with a national government to implement facial recognition techniques on finding missing persons as a result of an ongoing conflict. They plan to use digital photographs submitted by the family members of missing persons and scraping photographs from public social media sources to identify missing persons.

Questions:

- What are some of the fairness concerns that we must think about at different steps in a machine learning project development process?
- How might you address some of these concerns?
- What are some of the other concerns or additional considerations that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?

Facilitator Notes

How might data and ML model implementation cause disproportionate harm?

Prompt: Think through what happens if the model gets it "wrong"? In this case, what would happen if the service may work better for some (ages/genders/ethnicities) than others - meaning some groups may remain missing more than others.

Equity

- Migrants and their families might not know about the service, or they lack the connectivity required to access the service, or face further barriers of literacy, language, and IT skills.

Prompt: How might bias be embedded in training data and lead to model failures? What concerns might you have about data used to develop this model?

Representativeness

- Data may include photos of adults more than kids, men than women.
- Not everyone will have a digital photo. Not everyone has access to social media

Prompt: Are there ways in which the model might reinforce or create a social bias? What would we need to consider with respect to social implications of a model like this?

Bias

- Facial recognition technology may not work as well for certain skin tones.
- Relying on social media may bias towards those who have access to smartphones or computers to be

have a social media presence -- reinforces bias to extent that those less well off to begin with are more likely to be displaced (in some cases) and less likely to be found by this method

How well do we understand how ML models are working? *Would we recognize bias or inequities when (or before) they occur?*

- **Explainability** - to what extent can the predictions made by ML model be understood in non-technical terms? Can we interpret the relationships underlying the model's predictions?
- **Auditability** - to what extent can outside actors query AI/ML models (eg, to check for bias)?
- **Accountability** - what mechanisms are in place to identify when mistakes are made? To solicit feedback from those affected by the predictions the model makes? To redress possible harms that result from mistakes?
    - What if someone doesn't want to be found? Consent - How will users consent to being identified?
    - What if misidentification occurs?

What steps should we take to mitigate fairness concerns?

- **Representative data:**
    - Consider reviewing/strengthening data for representativeness (skin color, ages)
    - Provide photo scanning service for those who don't have digital photos.
    - Digital literacy and awareness campaign so people know the opportunity and harms
    - Look for additional data sources besides social media.
- **Unequal model performance:**
    - Measuring performance across groups - is it identifying better for men and adults than women and children?
- **Model failures:**
    - Be intentional in deciding how to optimize the model **-** is it better to have a few false positives (identifying 'wrong' people or people who don't want to be found) so that you don't miss anyone who wants to be found, or vice versa?
    - Have case workers participate in model design process
    - Provide alternative (low tech) matching service
- **Explainability:**
    - Have a team of humanitarian experts and data science experts work to find the right balance, decide how to use model results
- **Accountability:**
    - Get feedback from humanitarian staff on how the model results compare with other (eg manual) solutions
    - Provide means to remove data upon request (right to be forgotten)
    - Define adjudication process and triggers

Other concerns

- Data retention, ownership, privacy, security - highly sensitive, biometrics are unchangeable. Who is accountable for privacy and preventing misuse?
- Organizational capacity to manage tech and to support/sustain it
- Is AI 'better' / more appropriate solution than a manual process?
- False positive match raises hopes, causes emotional upheaval, potentially family members incur costs.

NETHOPE

AI Ethics for Nonprofits Workshop

As we get started

Please use the Chat window to say hello, tell us where you're joining from, and what organization you work for.

About the workshop

The Artificial Intelligence (AI) Ethics Workshop for Nonprofits was developed by NetHope, USAID, MIT D-Lab, and Plan International.

The goal of the workshop is to build capacity in the nonprofit sector to design and use AI responsibly and ethically.

NETHOPE

USAID
FROM THE AMERICAN PEOPLE

PLAN INTERNATIONAL

MIT D-Lab
designing for a more equitable world

Purpose of this workshop

Learn how to **practically apply ethical considerations related to the principle of Fairness** in the context of humanitarian and international development use cases.

Workshop Agenda

- Brief overview of the key AI Ethics concepts
- Breakouts: Practical application of ethical considerations
- Report out from breakouts and discussion
- Resources and next steps

Tools for today's workshop

- **Zoom for communication**
  - Video, audio, chat
- **Miro for collaboration**
  - Virtual whiteboarding, with sticky notes

NETHOPE

Ethics Primer

AI in the Nonprofit Sector

- AI can help us do our work better:
  - Reach more people with services and information they need.
  - Make decisions and act faster in emergencies.
  - Predict emergencies before they spread.
  - Amplify human effort and free up limited human resources to focus on high-priority work.
  - Improve outcomes through real-time feedback on the effectiveness of programs and recommendations for improvements.
- Challenge: Responsibly design and use AI technology - maximizing its benefits while minimizing risks and protecting human rights

AI Ethics

"A set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies."

The Alan Turing Institute

What is an ethical AI system?

An AI system that supports individual and collective well-being and enhances our ability to tackle global challenges.

Responsible Innovation

Responsible Innovation is a transparent, interactive, sustainable process by which organizations proactively evaluate how they can design and use technology in ways that are aligned with their values and missions.

What are some of the ethical issues surrounding technology use today?

- Intentional harms such as hate speech, misinformation, weaponization of technologies like AI.
- Infringement on rights and values such as surveillance.
- Unfair outcomes like discrimination and prejudice stemming from bias.

Bias & Fairness

Bias - Systematically favoring one group relative to another. Bias is always defined in terms of specific categories or attributes (eg gender, race, education level).

Fairness - Just and equitable treatment across individuals and/or groups.

Three general questions to ask:

- How might ML model design and implementation cause disproportionate harm?
- How well do we understand how ML models are working? Would we recognize bias or inequities when (or before) they occur?
- What happens when things go wrong?

Key considerations relevant to Fairness and AI

How might data and ML model implementation cause disproportionate harm?

- **Equity** - Does the model work better, or do model failures have significantly worse consequences for one group than another?
- **Representativeness** - To what extent is the training data representative of the population that will be affected by the use of the ML/AI model? To what extent are the people developing the ML/AI model?
- **Bias** - What biases may be embedded in the data? (*consider real-world power dynamics likely to shape what data is available and about whom)*

Where to look for potential disproportionate harm…

Systematic Differences in Failure Rates Between Groups of Interest

	Men	Women
% Accurate predictions		
% Inaccurate predictions		

**Your organization should work collaboratively to identify the groups across which you are concerned about fairness - it could be across more than two groups. The examples above are illustrative only.*

Where to look for potential disproportionate harm…

Systematic Differences in Failure Rates Between Groups of Interest

	Men	Women
% Accurate predictions		
% Inaccurate predictions		

Systematic Differences in Error Types Between Groups of Interest

	Men	Women
FALSE POSITIVE: Given loan, but won't repay		
FALSE NEGATIVE: Denied loan, but would have repaid		

**Your organization should work collaboratively to identify the groups across which you are concerned about fairness - it could be across more than two groups. The examples above are illustrative only.*

Where to look for potential disproportionate harm…

Systematic Differences in Failure Rates Between Groups of Interest

	Men	Women
% Accurate predictions		
% Inaccurate predictions		

Systematic Differences in Error Types Between Groups of Interest

	Men	Women
FALSE POSITIVE: Given loan, but won't repay		
FALSE NEGATIVE: Denied loan, but would have repaid		

Codified Social Bias

**Your organization should work collaboratively to identify the groups across which you are concerned about fairness - it could be across more than two groups. The examples above are illustrative only.*

Key considerations relevant to Fairness and AI

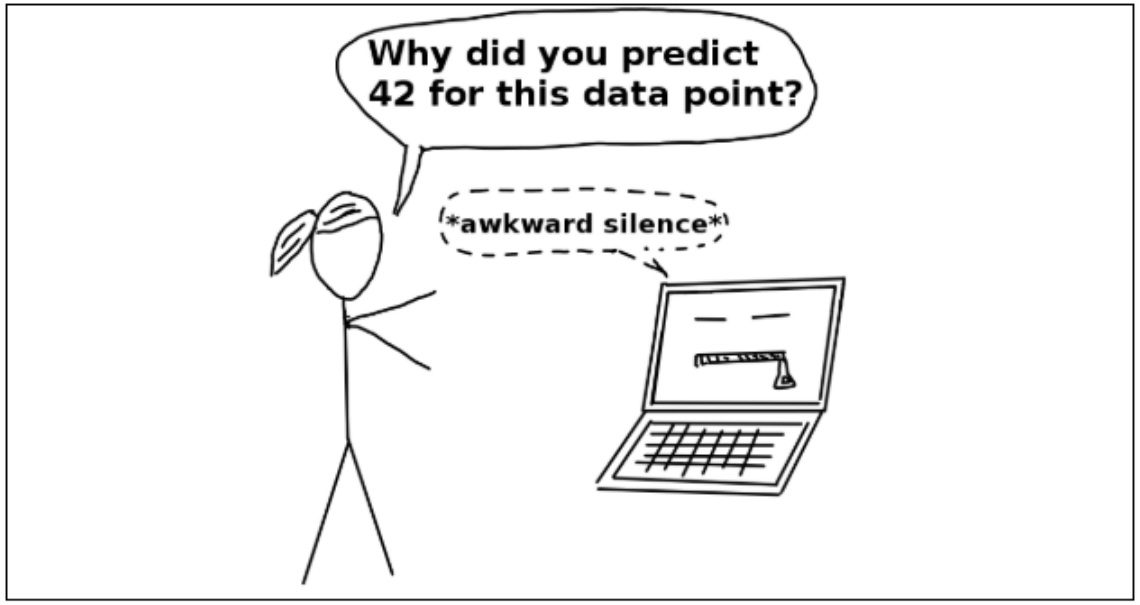
How well do we understand how ML models are working? Would we recognize bias or inequities when (or before) they occur?

- **Explainability** - to what extent can the predictions made by ML model be understood in non-technical terms? Can we interpret the relationships underlying the model's predictions?

- **Auditability** - to what extent can outside actors query AI/ML models (eg, to check for bias)?

How confident can we be that model results are not based on underlying biases in the data?

To what extent could we figure out what would need to change to get a different result?



Source: Interpretable Machine Learning, a book by Christopher Molnar

Key considerations relevant to Fairness and AI

What happens when things go wrong?

- **Accountability**
  - What mechanisms are in place to identify when mistakes are made?
  - To what extent will feedback be sought from those affected by the predictions the model makes?
  - What can be done to redress possible harms that result from mistakes?

What are some things we can we do to mitigate concerns?

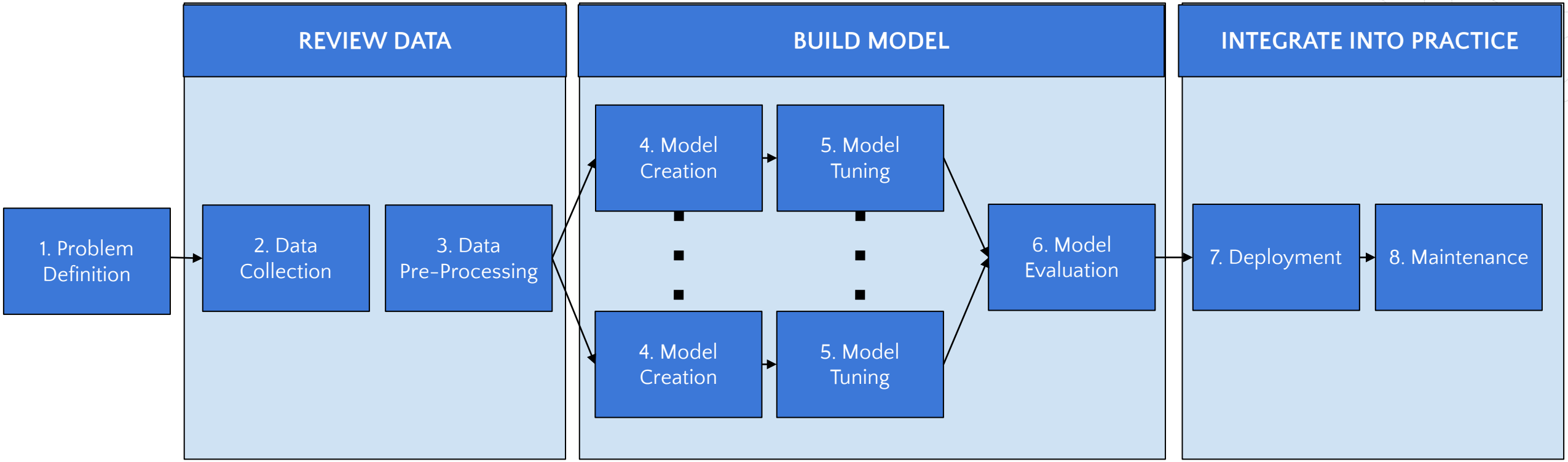
Project Level

- Ask the right questions
- Define which attributes you don't want to bias model predictions
- Identifying sources of bias (historical biases, individual biases, biases in data)
- Exploring technical approaches to testing for bias and implementing fairness

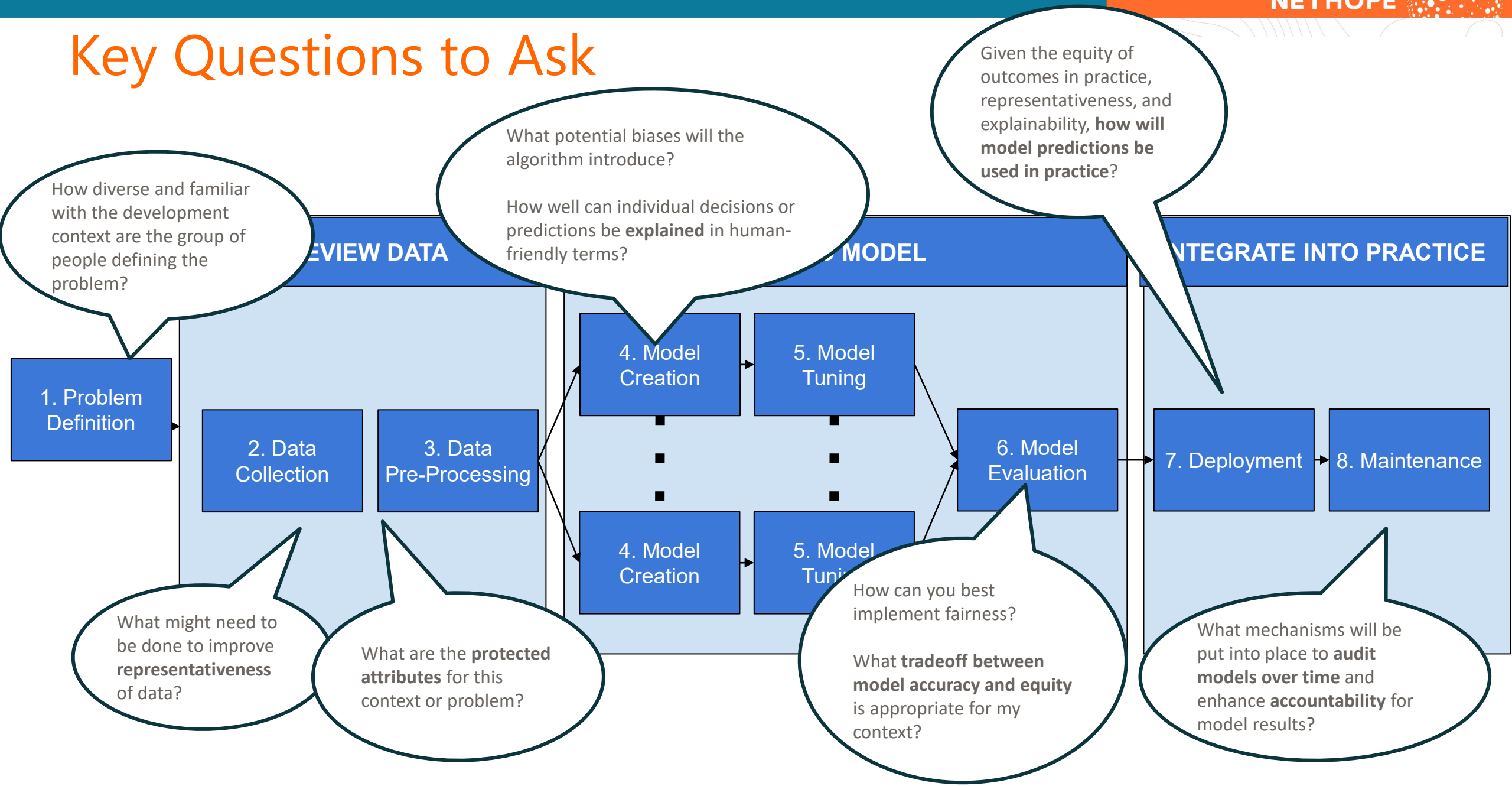
System Level

- Strengthening representativeness of data available for training ML models
- Support for auditing model outcomes, including consideration for open data and open algorithms
- Strengthening digital ecosystems and enabling environment for AI/ML
- Diversifying the workforce and organizations working in AI/ML

Machine Learning (ML) Project Overview



Key Questions to Ask



AI Suitability Framework for Nonprofits: https://nethope.app.box.com/file/643893297448

Protected Attributes

Traits that should not be used as a basis for decision-making in machine learning projects. Sometimes they are legally mandated. Your organization and data science team will need to define which traits to treat as protected in your context.

Typically, protected attributes include:

- race
- age
- gender
- sexual orientation
- religion
- socio-economic status

Other considerations

- Is the use of ML in your context solving a **relevant** problem?
- Is the application of ML technology adding **value** (eg informing more accurate, timely, actionable results?)
- Does your organization have sufficient **capacity** to implement the solution?
- Are there **other concerns (besides fairness**) you have about the proposed use of ML/AI?

NETHOPE

PLAN INTERNATIONAL

YOUTH EMPLOYMENT SOLUTIONS

Case Study: TESSA Chatbot

What can go wrong? How do you address the issues?

PLAN INTERNATIONAL

YOUTH EMPLOYMENT SOLUTIONS

TESSA

Plan International's **T**raining, **E**mployment and **S**upport **S**ervices **A**ssistant

- **Problem:** Marginalized youth in Asia are unable to effectively communicate their skills and link to suitable economic opportunities
- **Current approach:** Community Development Facilitators support youth to 'formalize' their skills, then link them to opportunities
- **Limitations:** Quality not quantity
- **Solution:** AI-powered chatbot on Facebook Messenger - TESSA
- **Benefits:** Approachable, accessible, consistent quality

What can go wrong?

- TESSA **reinforces** the status quo in terms of **gender inequalities** in the labor market and **gender-stereotypical** skills, jobs, and careers through its engagement with and recommendations for the users

PLAN INTERNATIONAL

YOUTH EMPLOYMENT SOLUTIONS

Why?

The beast of bias

- Team
- Data
- Problem framing
- Implementation
- Model / Tools

→ Narrowed down options and opportunities for girls

PLAN INTERNATIONAL

YOUTH EMPLOYMENT SOLUTIONS

How to address these issues?

- **Team:** Diversify/restructure team to address power dynamics
- **Data:** Develop ML/AI for recognizing patterns in data, analyze, evaluate and flag.
- **Problem framing:** Principles and process to increase intentional inclusion
- **Model/Tools:** Assess for bias
- **Implementation:** Implementing an agile inclusion methodology.

PLAN INTERNATIONAL

YOUTH EMPLOYMENT SOLUTIONS

Unless we intentionally include, we will unintentionally exclude

Q&A

INSERT:

Photo of Speaker 1

Speaker 1
Name, Title, Organization, Email

INSERT:

Photo of Speaker 2

Speaker 2
Name, Title, Organization, Email

INSERT:

Photo of Speaker 3

Speaker 3
Name, Title, Organization, Email

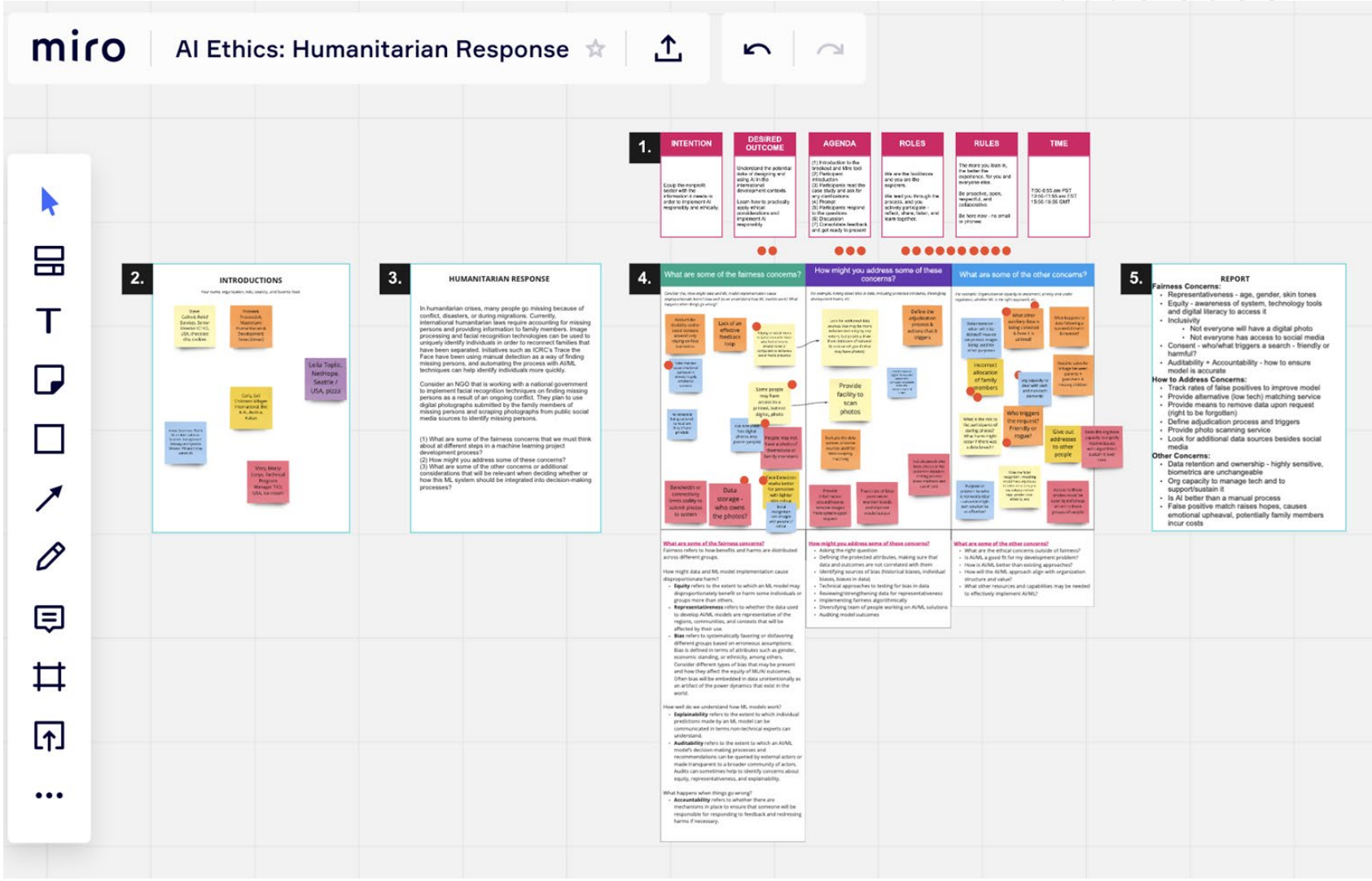
Short break

NETHOPE

Breakouts

Breakouts: Introduction

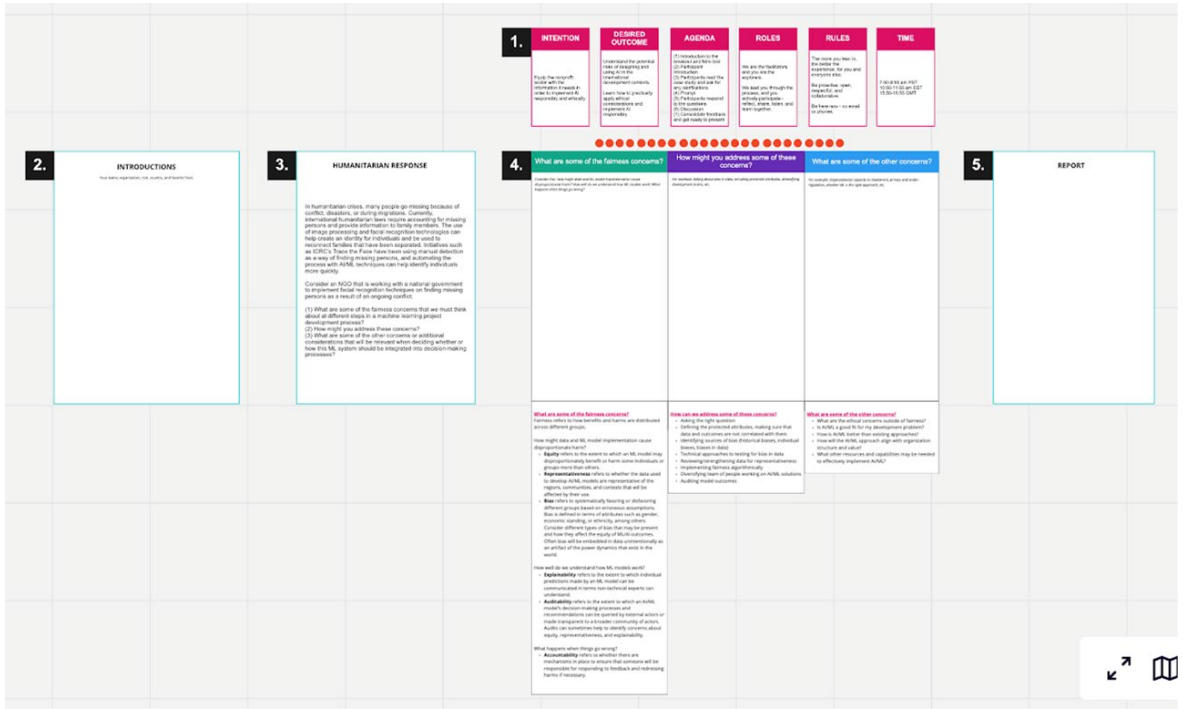
- 5 breakouts centered around use cases from humanitarian response, health, education, agriculture, workforce
- Hands-on, collaborative work using Miro and Zoom, followed by a report-out from each group

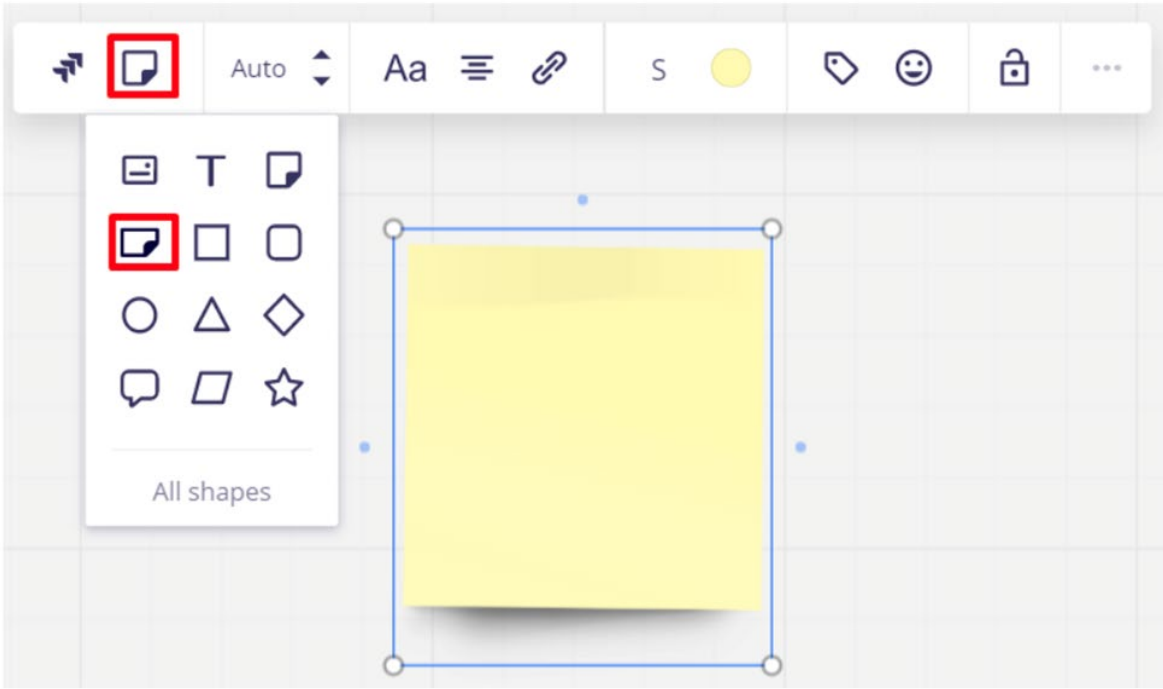


Breakouts: Facilitators

PHOTO	PHOTO	PHOTO	PHOTO	PHOTO	PHOTO
Name Organization	**Name** Organization	**Name** Organization	**Name** Organization	**Name** Organization	**Name** Organization
Humanitarian Response	**Health**	**Education**	**Agriculture**	**Workforce**	**Main room**

Breakouts: Miro demo





Welcome back, hope you enjoyed the breakout!

Please get ready to share.

Report-out and Discussion

Each group will have 2 minutes to share their key insights, using their Miro board as a backdrop.

- Fairness concerns
- How to address fairness-related concerns
- Other concerns or additional considerations

Few minutes at the end to share observations, feedback, additional insights on the use cases.

Order: workforce, health, education, agriculture, humanitarian response

Breakouts: Use Cases

Workforce

Machine learning is used to screen resumes from job applicants and determine which ones should be offered interviews.

Health

Machine learning is used to better predict which people living with HIV/AIDS are most at risk of being "lost to follow up" in the first 12 months of their treatment.

Education

Machine learning is used to evaluate the quality of student writing among high school students across India, with the goal to make tailored recommendations of support needed to improve writing performance.

Agriculture

Machine learning is used to improve farmer income by helping them determine where and when to purchase inputs and sell crops as well as how to connect with the appropriate markets.

Humanitarian Response

A facial recognition system designed to help identify and find missing persons (missing due to conflict, disaster or migration) with the goal to reconnect them with their families.

Next steps:

- Continue learning about ethical, responsible development and use of AI.

  See resources on the next slide.
- If you are a NetHope Member, join NetHope's AI Working Group:

  http://bit.ly/ET_WorkingGroup

Resources

- Key AI Ethics concepts
- Exploring Fairness in Machine Learning for International Development by MIT D-Lab, with the support from USAID
- AI Ethics: 5 Considerations for Nonprofits
- NetHope's AI Ethics webinars:
    - Part I (recording, slides)
    - Part II (recording, slides)
    - Part III (recording, slides)
- NetHope AI Suitability Toolkit for Nonprofits
- USAID: Reflecting the Past, Shaping the Future: Making AI Work for International Development
- AI Primer (recording, slides)

Thank you

for participating in the **AI Ethics for Nonprofits** Workshop!

NETHOPE

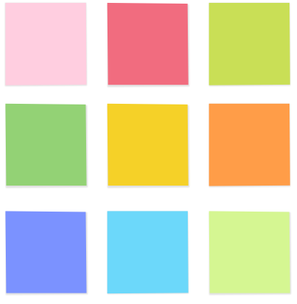
1.

INTENTION	DESIRED OUTCOME	AGENDA	ROLES	RULES	TIME
Equip the nonprofit sector with the information it needs in order to implement AI responsibly and ethically.	Understand the potential risks of designing and using AI in the international development contexts. Learn how to practically apply ethical considerations and implement AI responsibly.	(1) Introduction to the breakout (2) Participant introduction (3) Participants read the case study and ask for any clarifications (4) Participants respond to the questions (5) Discussion (6) Consolidate feedback and get ready to present	We are the facilitators and you are the explorers. We lead you through the process, and you actively participate - reflect, share, listen, and learn together.	The more you lean in, the better the experience, for you and everyone else. Be proactive, open, respectful, and collaborative. Be here now - no email or phones.	75min

2.

INTRODUCTIONS

Your name, organization, role, country, and favorite food.



3.

HUMANITARIAN RESPONSE

In humanitarian crises, many people go missing because of conflict, disasters, or during migrations. Currently, international humanitarian laws require accounting for missing persons and providing information to family members. Image processing and facial recognition technologies can be used to uniquely identify individuals in order to reconnect families that have been separated. Initiatives such as ICRC's Trace the Face have been using manual detection as a way of finding missing persons, and automating the process with AI/ML techniques can help identify individuals more quickly.

Consider an NGO that is working with a national government to implement facial recognition techniques on finding missing persons as a result of an ongoing conflict. They plan to use digital photographs submitted by the family members of missing persons and scraping photographs from public social media sources to identify missing persons.

(1) What are some of the fairness concerns that we must think about at different steps in a machine learning project development process?
(2) How might you address some of these concerns?
(3) What are some of the other concerns or additional considerations that will be relevant when deciding whether or how this ML system should be integrated into decision-making processes?

4.

What are some of the fairness concerns?	How might you address some of these concerns?	What are some of the other concerns?
Consider this: How might data and ML model implementation cause disproportionate harm? How well do we understand how ML models work? What happens when things go wrong?	*For example: Asking about bias in data, including protected attributes, diversifying development teams, etc.*	*For example: Organizational capacity to implement, privacy and under-regulation, whether ML is the right approach, etc.*

What are some of the fairness concerns?
Fairness refers to how benefits and harms are distributed across different groups.

How might data and ML model implementation cause disproportionate harm?
- **Equity** refers to the extent to which an ML model may disproportionately benefit or harm some individuals or groups more than others.
- **Representativeness** refers to whether the data used to develop AI/ML models are representative of the regions, communities, and contexts that will be affected by their use.
- **Bias** refers to systematically favoring or disfavoring different groups based on erroneous assumptions. Bias is defined in terms of attributes such as gender, economic standing, or ethnicity, among others. Consider different types of bias that may be present and how they affect the equity of ML/AI outcomes. Often bias will be embedded in data unintentionally as an artifact of the power dynamics that exist in the world.

How well do we understand how ML models work?
- **Explainability** refers to the extent to which individual predictions made by an ML model can be communicated in terms non-technical experts can understand.
- **Auditability** refers to the extent to which an AI/ML model's decision-making processes and recommendations can be queried by external actors or made transparent to a broader community of actors. Audits can sometimes help to identify concerns about equity, representativeness, and explainability.

What happens when things go wrong?
- **Accountability** refers to whether there are mechanisms in place to ensure that someone will be responsible for responding to feedback and redressing harms if necessary.

How might you address some of these concerns?
- Asking the right question
- Defining the protected attributes, making sure that data and outcomes are not correlated with them
- Identifying sources of bias (historical biases, individual biases, biases in data)
- Technical approaches to testing for bias in data
- Reviewing/strengthening data for representativeness
- Implementing fairness algorithmically
- Diversifying team of people working on AI/ML solutions
- Auditing model outcomes

What are some of the other concerns?
- What are the ethical concerns outside of fairness?
- Is AI/ML a good fit for my development problem?
- How is AI/ML better than existing approaches?
- How will the AI/ML approach align with organization structure and value?
- What other resources and capabilities may be needed to effectively implement AI/ML?

5.

REPORT

NETHOPE