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?

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.