AI Ethics Webinar Series: Part II

June 4th, 2020

Amy Paul

USAID Center for Digital Development

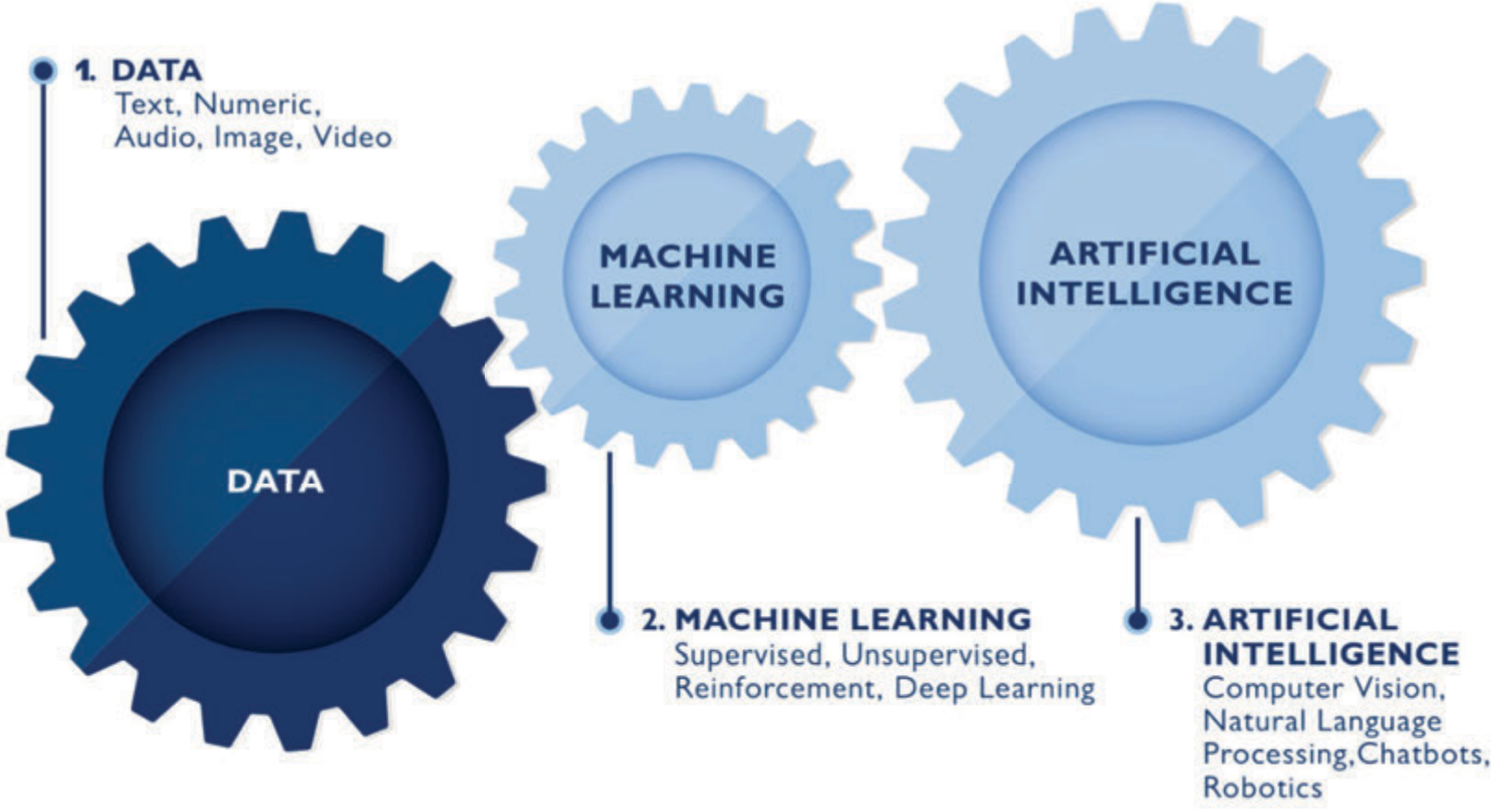
Objectives

- Introduce relevant AI/ML applications in international development

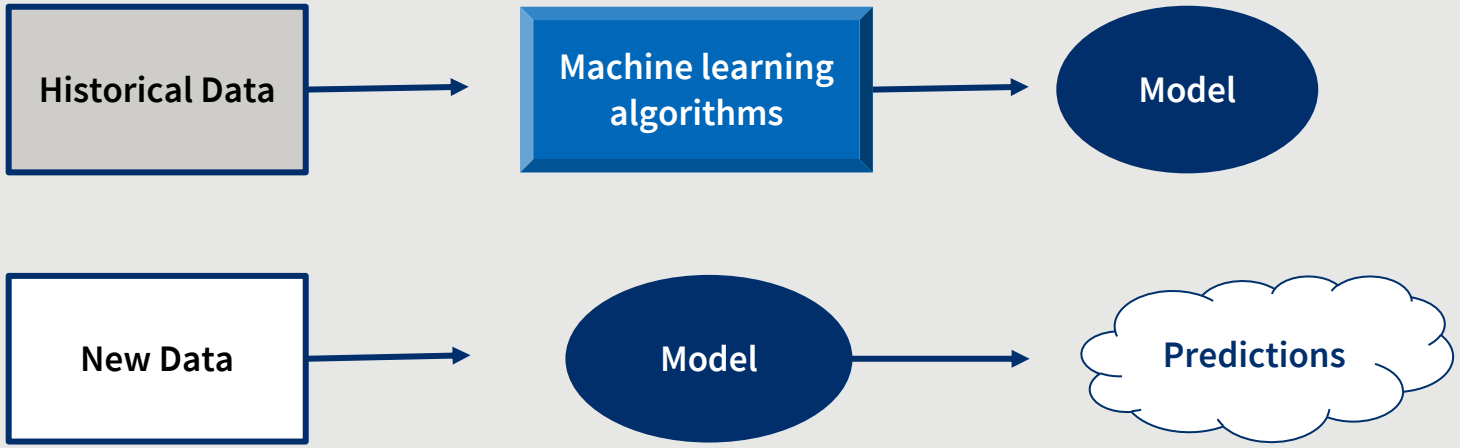
- Highlight key ethics considerations for responsible use of AI/ML technology in international development

- Identify risks that may arise in the development and use of AI/ML technology with respect to ethics considerations

- Discuss approaches to safeguarding against risks


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

Predictive Modeling with Machine Learning



Key AI capabilities

Computer vision processes images or video in order to identify objects or interpret scenes or events.

Natural language processing (NLP) analyzes or synthesizes text of human languages such as English, Spanish, or Arabic.

MIT, Harvard: Identifying Infection in surgical scars

Key AI capabilities (...continued)

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 and Interpretation creates new text, images, video from understanding of key patterns in training text, images, video.

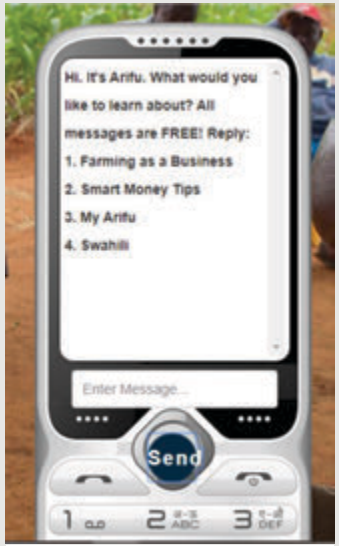
Rainforest Connection: Detecting Illegal Deforestation via monitoring chainsaw sounds

Practical Examples: Chatbots

Users request information from a system, often using written or spoken queries.

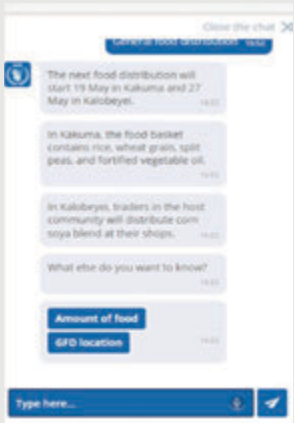
Key capabilities: Natural language processing, speech recognition, conversational interfaces (chatbots)

Application areas: Health, Agriculture, Financial inclusion, humanitarian aid



Arifu chatbot and digital learning platform for farmers

Plan International's TESSA chatbot



World Food Programme's Foodbot

Practical Examples: Text and Image Classification in Humanitarian response

Uses satellite imagery and other geographic information (e.g. geotagged user data) rapidly process information for situational awareness, disaster response planning, resource allocation, etc.

Key capabilities: Computer vision, geospatial analysis, NLP

Application areas: Humanitarian assistance, disaster response, law enforcement, policy planning

CMU Software Engineering Institute: Disaster damage assessment with computer vision

Ushahidi crisis mapping platform in Haiti

Practical Examples: Personalized predictions/ratings

Uses diverse types of data (behavioral, demographic, economic, etc.) to predict specific outcomes or behaviors.

Key capabilities: Advanced Analytics, NLP

Application areas: Employment, credit scoring, law enforcement, health

https://twitter.com/talamobile/status/671525771546988544

Practical Examples: Vision & Audio diagnostics

Analyzes images or sounds (often captured via smartphone) to diagnose disease.

Key capabilities: Computer vision, speech or audio recognition, NLP

Application areas: Health, Agriculture

MIT, Harvard: Identifying Infection in surgical scars

Ubenwa.ai uses machine learning to analyze baby cries in order to identify perinatal asphyxia at early stage

Makerere University AI Lab tests smartphone app for diagnosing malaria

Values for Machine Learning and AI-
Key considerations relevant to Fairness

- Equity
- Representativeness
- Explainability
- Auditability
- Transparency
- Suitability & Added Value

USAID
FROM THE AMERICAN PEOPLE

Equity

- Does an ML model disproportionately benefit or harm some individuals or groups more than others?

Equity

- Does an ML model disproportionately benefit or harm some individuals or groups more than others?

What does equity mean with respect to machine learning?

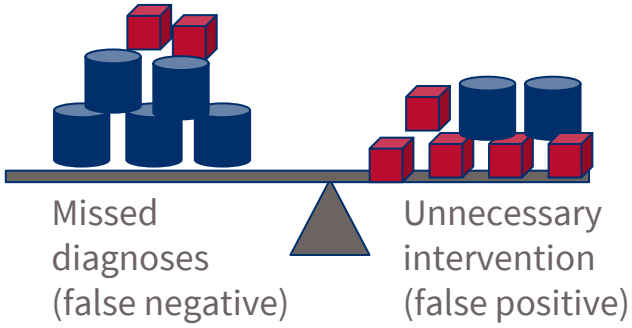
(See this article by Abdi Latif Dahir for more info on African languages in voice recognition tools)

ML models can perform significantly better for one group than another, creating an uneven opportunity to utilize ML technology *(e.g. language and image processing tools)*

Equity

Does an ML model disproportionately benefit or harm some individuals or groups more than others?

What does equity mean with respect to machine learning?



ML models can fails equally often across groups, but produce systematic differences in the type of error each group experiences

(e.g. diagnostics, scoring/eligibility applications)

Equity

Does an ML model disproportionately benefit or harm some individuals or groups more than others?

What does equity mean with respect to machine learning?

ML models can be technically accurate, yet reinforce existing inequities and social bias

(e.g., credit scoring, hiring, recommender apps)

Representativeness

Is the data used to train the ML models representative of the people who will be affected by the model's application?

Image credit: Belle Demont

Representativeness

Is the data used to train the ML models representative of the people who will be affected by the model's application?

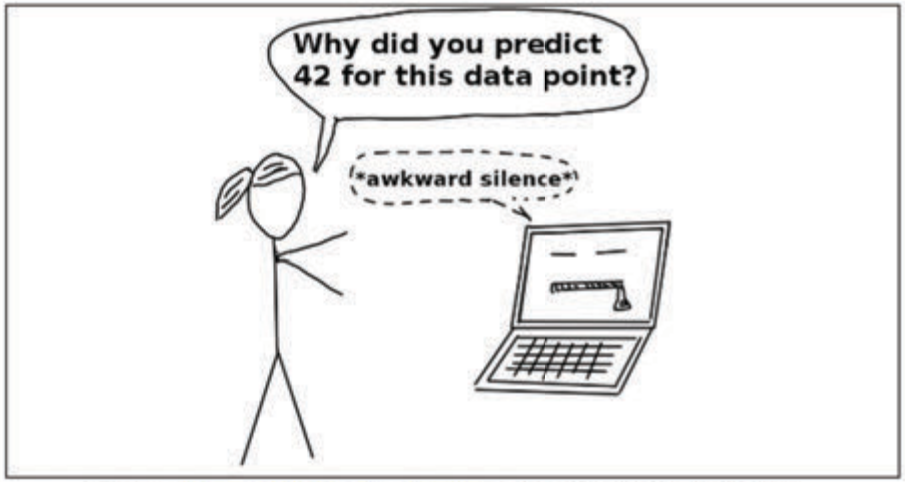
Why does representativeness matter in machine learning?

Image credit: https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

If data aren't representative of the real-world context in which model is used, ML models can produce misleading results that contribute to inequitable outcomes

Explainability

Can individual predictions or decisions be explained in human-friendly terms?

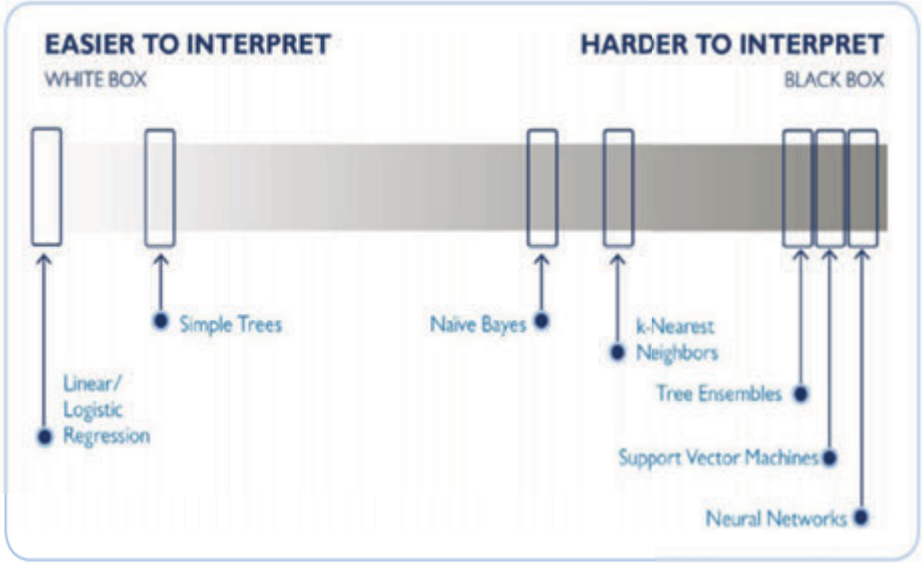


Source: Interpretable Machine Learning, a book by Christopher Molnar

Why is explainability important?

Explainability

Can individual predictions or decisions be explained in human-friendly terms?



The choice of algorithm affects both model accuracy and our understanding of how predictions are made. If we can't determine how a model is using input data, it is harder to identify when they produce unfair outcomes.

Auditability

Can the model's decision-making processes and recommendations be queried by external actors?

Auditability

Auditability: Can the model's decision-making processes and recommendations be queried by external actors?

Why is auditability important?

Model Cards for Model Reporting FAT* '19, January 29–31, 2019, Atlanta, GA, USA

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images, augmentative applications, such as providing details for people who are blind, or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect, smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset: Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [40]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the 5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 – 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [1].
- Given gender classes are binary (male not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analysis

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Opening up the model's decision-making process for question and inspection increases likelihood of identifying potential harms and biases ahead of time

Accountability

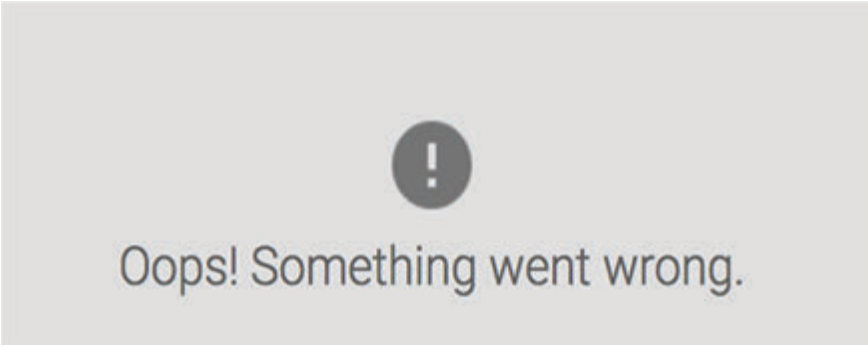
Are there mechanisms in place to ensure that someone will be responsible for responding to feedback and redressing harms, if necessary?

Illustration credit: Frits Ahlefeldt, http://hikingartist.com.

Accountability

Are there mechanisms in place to ensure that someone will be responsible for responding to feedback and redressing harms, if necessary?

What does accountability mean with respect to machine learning?



Without strong commitments to monitor outcomes, work collaboratively. and willingness to learn from failures, unintentional harms of machine-learning based tools may go unaddressed

Other Considerations: Relevance & Added Value

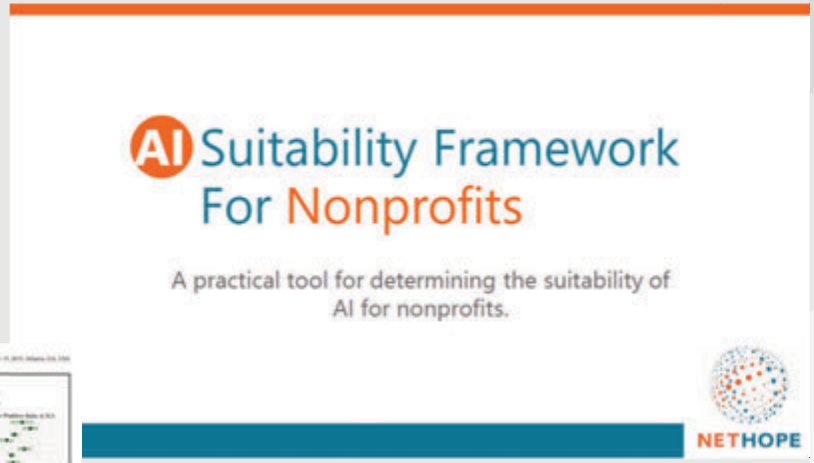
- Is the use of ML in your context solving a **relevant** problem?
- Is the application of ML technology adding **value** (e.g. informing more accurate, timely, actionable results?)

Safeguards for Fairness in Machine Learning

USAID
FROM THE AMERICAN PEOPLE

Awareness Raising and Practical Tools and Resources

- Webinar and trainings
- Toolkits
- Tools that make it easier for non-coders to engage with and test machine learning based applications



Technical Approaches to Fairness in Machine Learning

- Addressing fairness consideration in the technical decisions of model development
  - data selection
  - choice of algorithm
  - model performance metrics

- Technical approaches to bias checks, greater interpretability

Capacity Strengthening and Diversification of Workforce

- Including people of different cultural, educational, economic, social, ethnic backgrounds to understand and address issues related to fairness

- Strengthening capacity for local innovation and technology development

Data Science Africa

Since 2013

Black in AI

Interest

Strengthening Digital Ecosystems

- Strengthening inclusive and secure data systems

- Supporting local innovation and diversity in start-up ecosystems

- Shaping policy frameworks for fair and accountable use of digital technology

END

Next Up!

June 18th, 2020
11am ET

USAID
FROM THE AMERICAN PEOPLE