Responsible AI in Practice

Dr. Nashlie Sephus

Principal Tech Evangelist, Amazon AI
AWS

Responsible AI: Outline

What is Responsible AI, and what is it not?

What is leadership accountable for?

What are the differences between ML services, ML use cases, and ML applications?

What are the kinds of problems you need to address?

When and how do you incorporate feedback from stakeholders?

Traditional software solutions

1. We spec with human language
2. Customers do not expect to test
3. New releases perform the same or better on all inputs

Machine learning solutions

1. We spec with datasets
2. Customers should test
3. New releases perform the same or better overall

Scaling fairness, explainability, and privacy across the AWS machine learning stack

AI SERVICES

HEALTHCARE AI

NEW
Amazon HealthLake

Amazon Transcribe for Medical

Amazon Comprehend for Medical

INDUSTRIAL AI

NEW
AWS Panorama + Appliance

NEW
Amazon Monitron

NEW
Amazon Lookout for Equipment

NEW
Amazon Lookout for Vision

ANOMALY DETECTION

NEW
Amazon Lookout for Metrics

CODE AND DEVOPS

NEW
Amazon DevOps Guru

Amazon CodeGuru

VISION

Amazon Rekognition

SPEECH

Amazon Polly

Amazon Transcribe
+Medical

TEXT

Amazon Comprehend
+Medical

Amazon Translate

Amazon Textract

SEARCH

Amazon Kendra

CHATBOTS

Amazon Lex

PERSONALIZATION

Amazon Personalize

FORECASTING

Amazon Forecast

FRAUD

Amazon Fraud Detector

CONTACT CENTERS

Contact Lens
Voice ID
For Amazon Connect

ML SERVICES

Amazon SageMaker

Label data

SAGEMAKER STUDIO IDE

NEW
Aggregate & prepare data

NEW
Store & share features

Auto ML

Spark/R

NEW
Detect bias

Visualize in notebooks

Pick algorithm

Train models

Tune parameters

NEW
Debug & profile

Deploy in production

Manage & monitor

NEW
CI/CD

Human review

NEW: SageMaker JumpStart

NEW: Model management for edge devices

FRAMEWORKS & INFRASTRUCTURE

TensorFlow

MXNet

PyTorch

GLUON

Keras

scikit learn

HOROVOD

DeepGraphLibrary

AWS Deep Learning AMIs & containers

GPUs & CPUs

Amazon Elastic Inference

AWS Trainium

AWS Inferentia

FPGA

AI/ML on AWS: Innovation, choice, and flexibility

200+

new capabilities for machine learning and artificial intelligence in 2021

100,000+

customers have used machine learning on AWS

Customers from across industries & geographies

“Facial recognition AI can’t identify trans and non-binary people”

—Quartz, October 2019

“How photos of your kids are powering surveillance technology

Millions of Flickr images were sucked into a database called MegaFace. Now some of those faces may have the ability to sue.”

—The New York Times, October 2019

“San Francisco Bans Facial Recognition Technology”

—The New York Times, May 2019

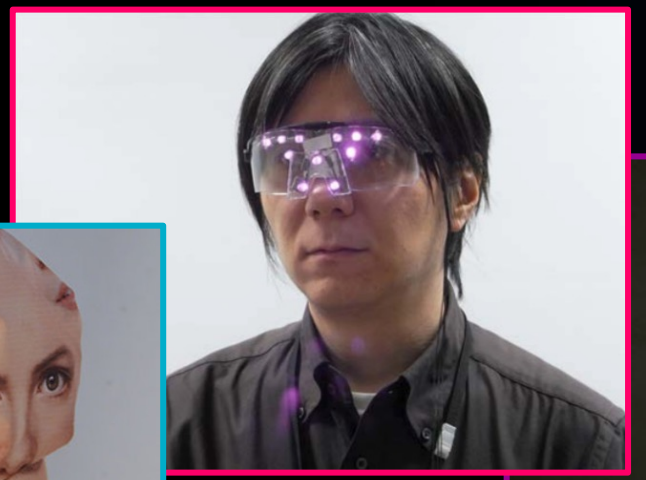
Adversarial examples

“This jewelry is a brilliant shield against face-recognition intrusions”

—Fast Company, July 2019

“These clothes use outlandish designs to trick facial recognition software into thinking you’re not a human”

—Insider, January 2020



The reach of AI and ML is growing

Global spending on AI-based technologies will reach **$204 billion by 2025**

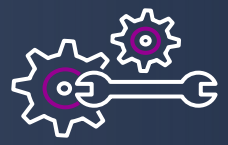
57% of executives said that AI would **transform their organization** in the next three years

AI plays a role in nearly **every part of an organization** from chatbots to personalized recommendations

With the growth of AI comes the recognition that we must all use it **responsibly**

Our commitment

Our commitment to develop AI and machine learning in a responsible way is integral to our approach



Transform responsible AI from theory to practice

Nurture and educate a more diverse generation of leaders in ML

Integrate responsible AI into the end to end ML lifecycle

Advance the science behind responsible AI

“**Responsible AI**
[…] AI that is innovative, trustworthy and that respects human rights and democratic values. ”

OECD

https://oecd.ai/en/ai-principles

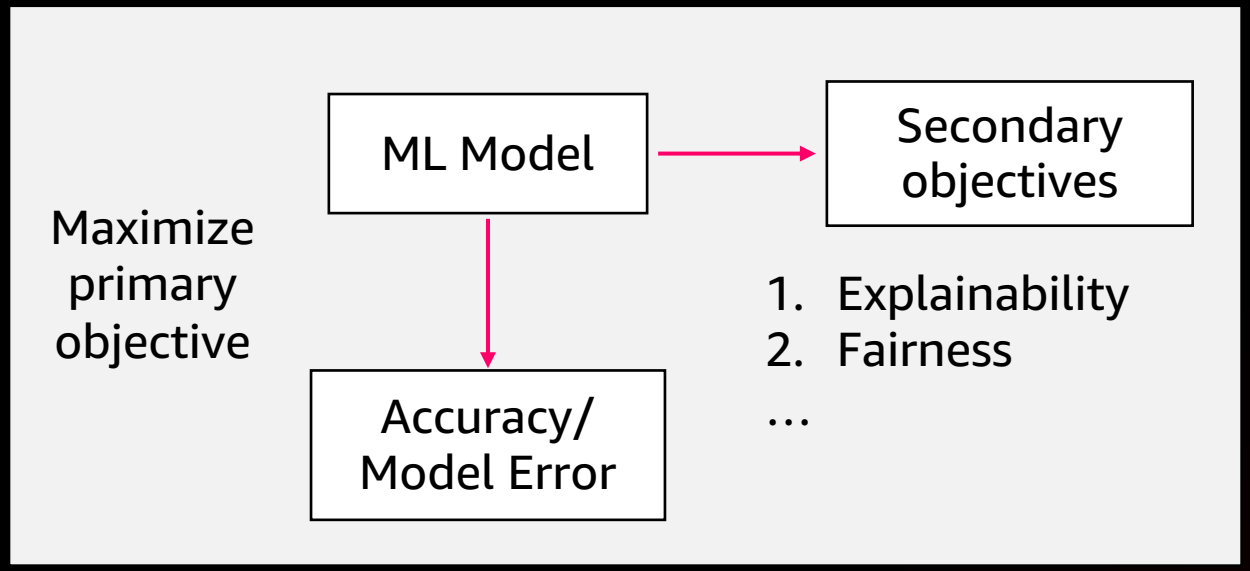
→ **Dimensions of responsible AI**

Dimensions of responsible AI

Dimension	Example Metric
Privacy & Security	Is data used in accordance with privacy & legal considerations, and protected from theft and exposure?
Fairness	Are there harmful disparities in system performance across subpopulations?
Explainability	Does the system offer a clear rationale for its decisions?
Robustness	How hard is it to confuse or fool the system, e.g. with “adversarial” examples?
Transparency	Are users enabled to make informed choices about their use of the system?
Governance	How do you enforce and ensure these responsible AI practices are being carried out amongst all stakeholders?

Tradeoffs in responsible AI

1. Dimensions of responsible AI can be at odds with primary objective to have best possible model performance.

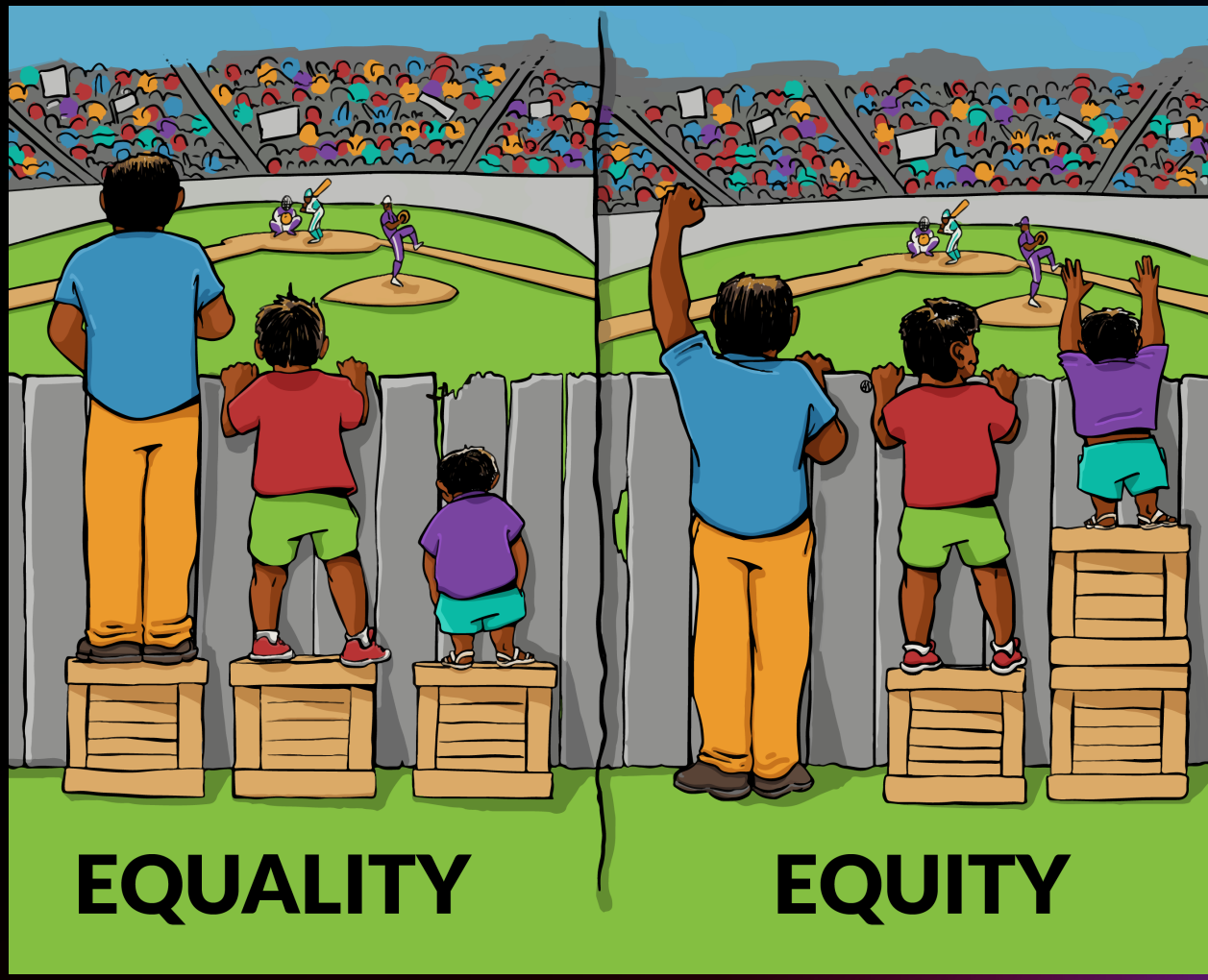


2. It is challenging to maximize all dimensions of responsible AI simultaneously.
   - E.g.: Preserving privacy makes data more coarse → potentially degrades ability to explain model behavior.
   - E.g.: Making model results and components more transparent → potentially security & privacy risk.

Fairness criteria & metrics

To operationalize theoretical constructs like fairness, we need mathematical constructs:

- Equality of opportunity
- Demographic Parity
- Fairness through Unawareness
- …

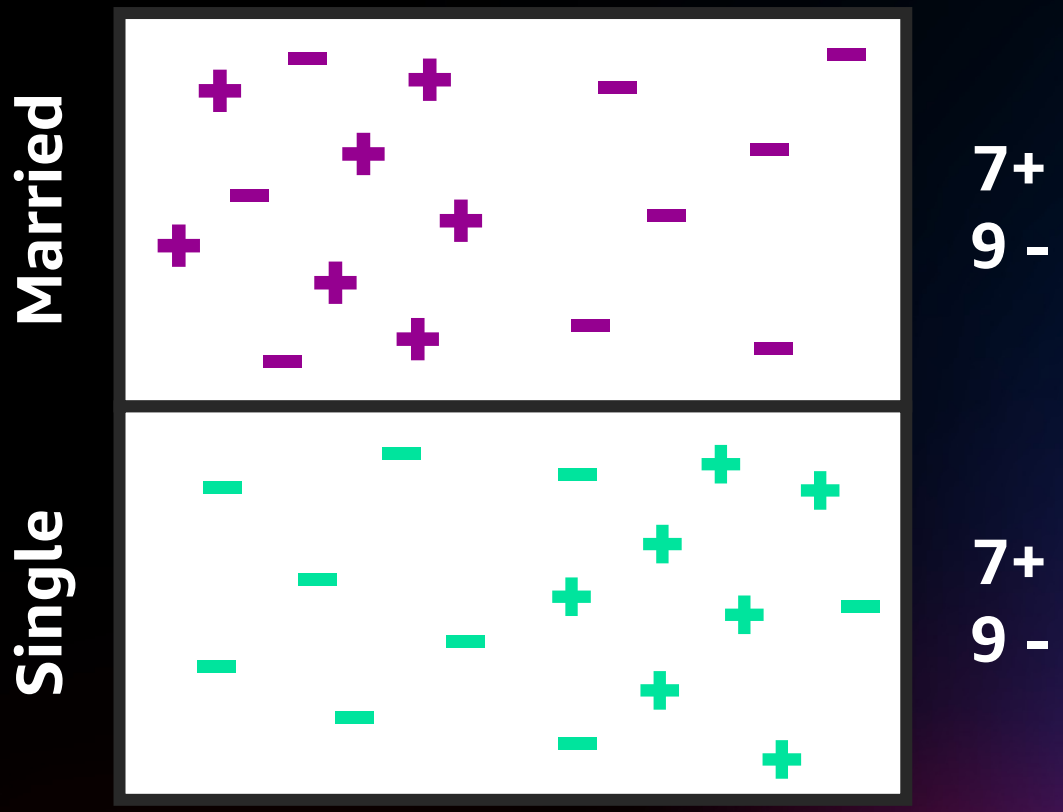


Source: https://interactioninstitute.org/illustrating-equality-vs-equity/

Intersectional fairness

- Individuals can exhibit multiple demographic attributes.
- If multiple attributes intersect, new sub-groups emerge.

Cannot build models for separate groups as against law (in USA).

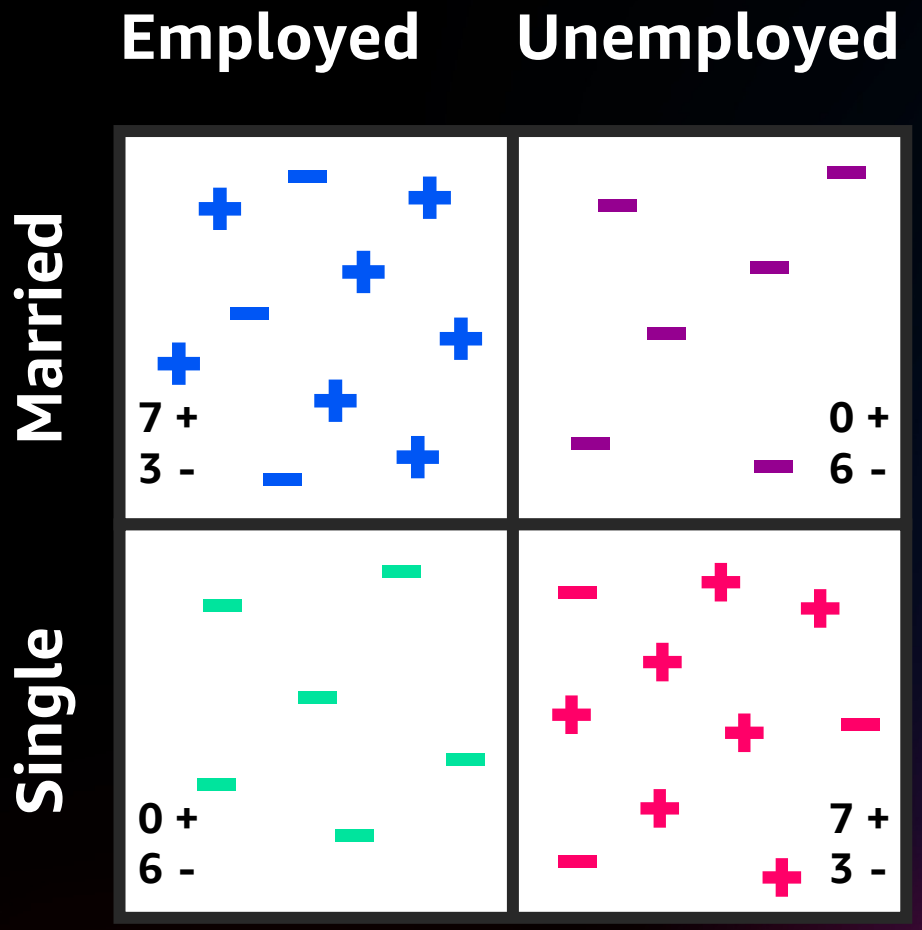


Intersectional fairness example

Attributes (e.g. married vs. single individuals) can intersect with other sensitive attributes (e.g. employed vs. unemployed)

→ new sub-groups emerge (e.g. married & unemployed, …)

Sub-groups may experience additional disparity (see positive/negative outcomes per sub-group).



Why is responsible AI complex? How can we succeed?

Success is **use case** specific	→	Simplify defining and measuring success
Technically hard (e.g., root cause analysis of bias)	→	Invent and simplify for all AI DevOps steps
Rapid changes in AI landscape (users, technologies, companies, societies)	→	Improve continuously over the long term

Simplify defining and measuring success

AMAZON SAGEMAKER GROUND TRUTH PLUS: TURNKEY SOLUTION FOR DATA LABELING

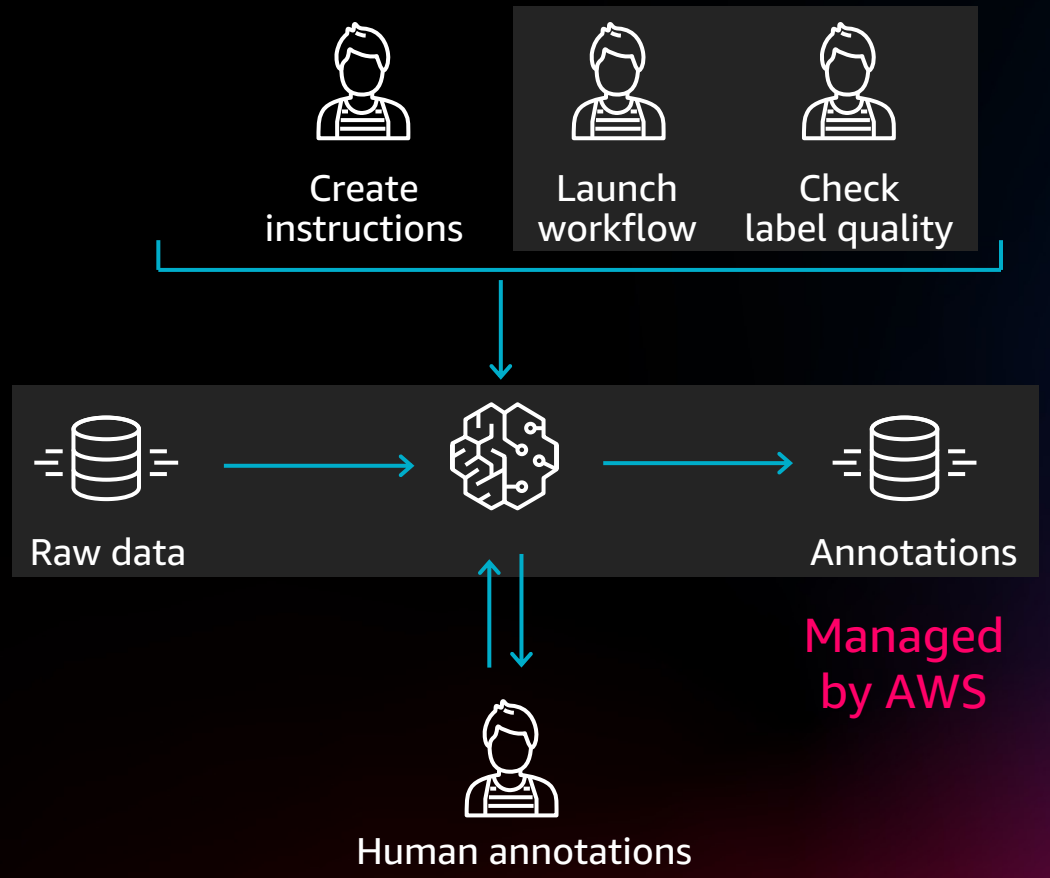
Quality guarantees

+

Throughput guarantees

+

Accessible to
data operations managers &
technical program managers



Amazon SageMaker Clarify

Detect bias in ML models and understand model predictions

Identify imbalances in data

Detect bias during data preparation

Check your trained model for bias

Evaluate the degree to which various types of bias are present in your model

Explain overall model behavior

Understand the relative importance of each feature to your model's behavior

Explain individual predictions

Understand the relative importance of each feature for individual inferences

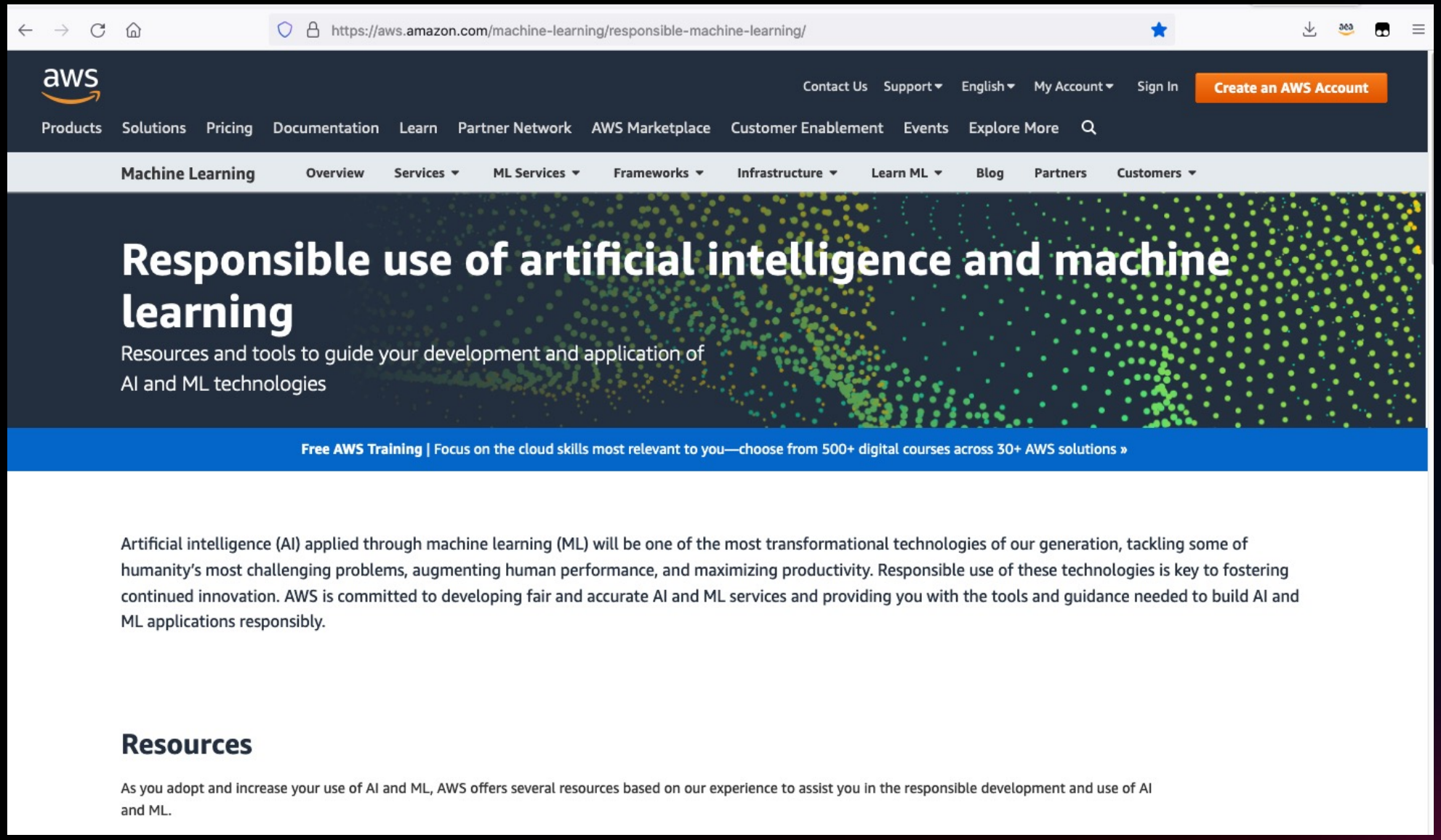
Detect drift in bias and model behavior over time

Provide alerts and detect drift over time due to changing real-world conditions

Generate automated reports

Produce reports on bias and explanations to support internal presentations

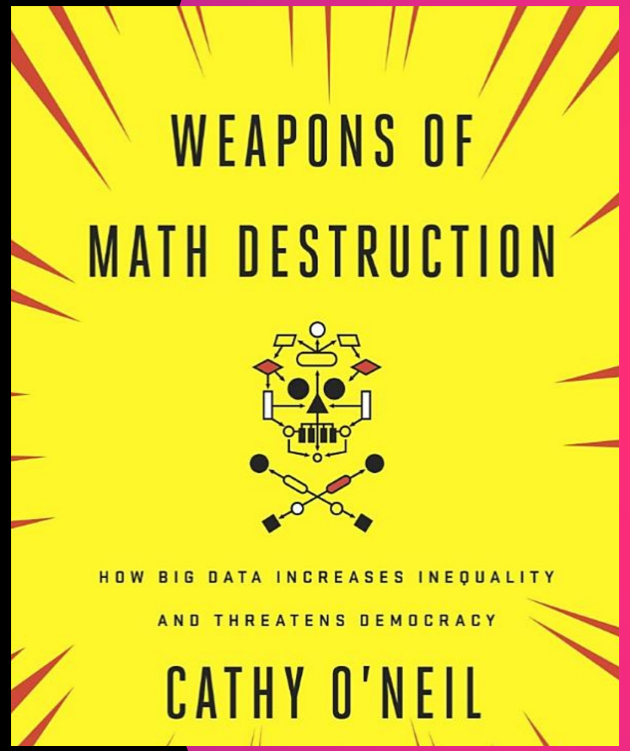
aws

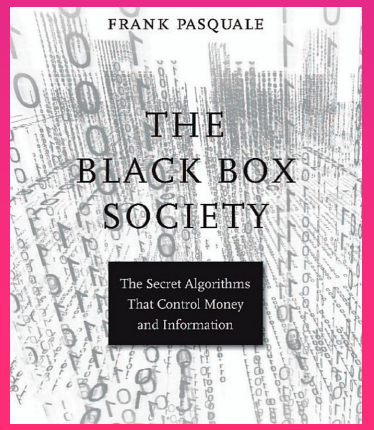


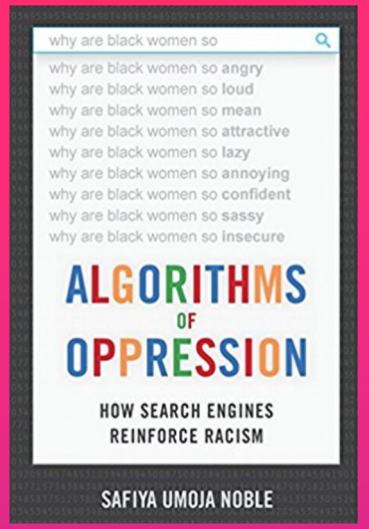
https://aws.amazon.com/machine-learning/responsible-machine-learning/

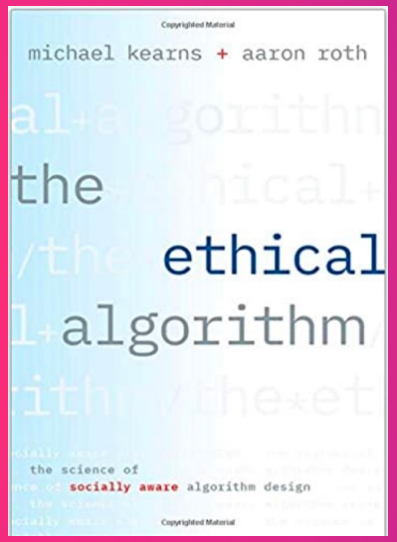
Algorithmic bias

- Ethical challenges posed by AI systems
- Inherent biases present in society
- Reflected in training data
- AI/ML models prone to amplifying such biases



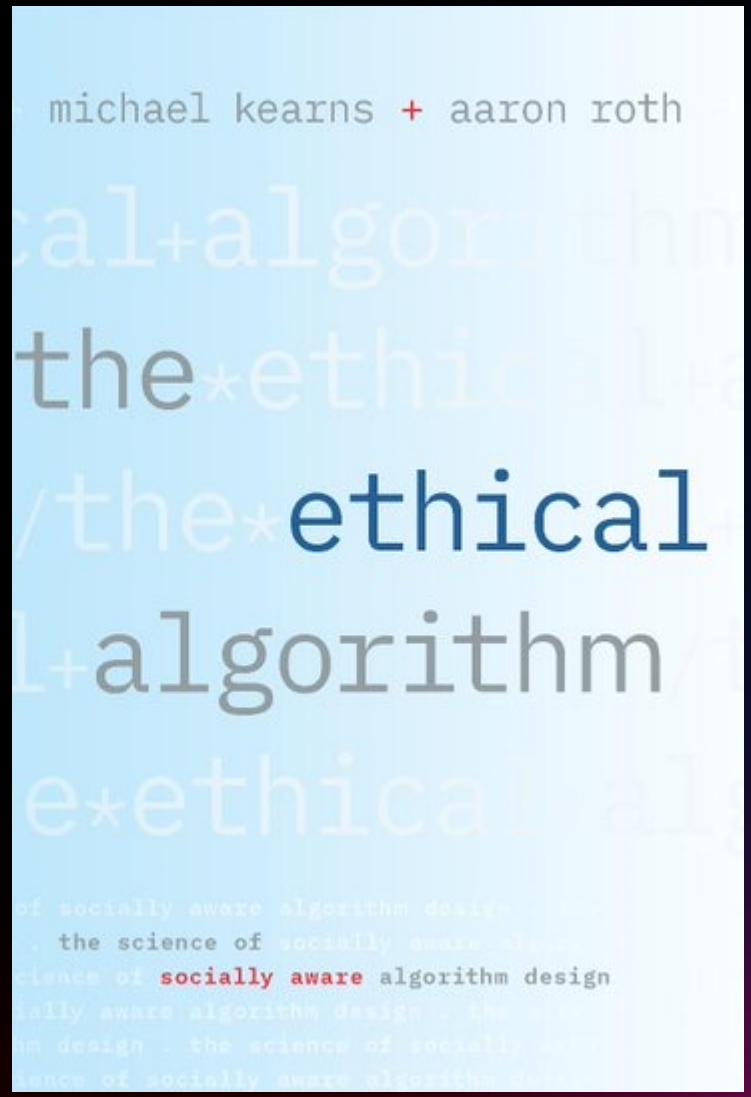




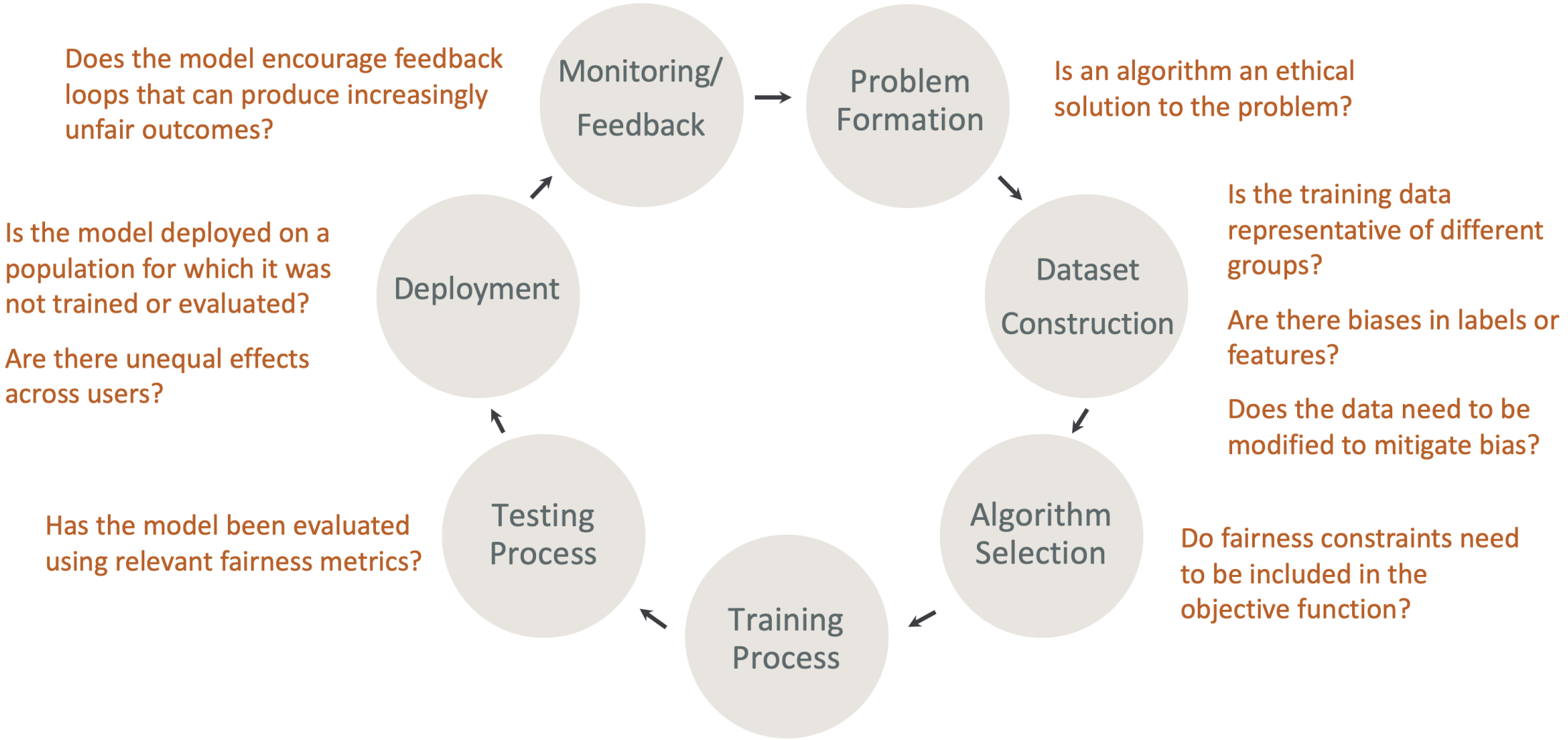


Role of government

- Use cases vary in risk – music recommendations vs. driving
- Governments should act to ensure appropriate use of AI
- To help develop risk-based regulations, we work with
  - Standards bodies (ISO)
  - The OECD working groups
  - Jurisdictions (when asked)



Responsible AI by design in ML lifecycle



Model Cards, AI Service Card Cards, and Datasheets

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.

Motivation: Datasets

How are datasets collected?

Where did the data come from?

When it comes to humans, were they aware?

Had the individuals given consent?

How are dataset owners being held accountable for consequences that may arise?

How do we create greater transparency about data?

AI Service Cards – Key Considerations

Dimension	Example Metric
Privacy & Security	Is data used in accordance with privacy & legal considerations, and protected from theft and exposure?
Fairness	Are there harmful disparities in system performance across subpopulations?
Explainability	Does the system offer a clear rationale for its decisions?
Robustness	How hard is it to confuse or fool the system, e.g. with “adversarial” examples?
Transparency	Are users enabled to make informed choices about their use of the system?
Governance	How do you enforce and ensure these responsible AI practices are being carried out amongst all stakeholders?

Bias

How will you measure and mitigate differences in the service's performance or downstream impact on subpopulations of the general public worldwide (defined by differing combinations of personal attributes such as – but not limited to – skin tone, face or body shape, ethnicity, race, age, gender, sexual orientation, language dialect, physical ability, health status, region, and religious, political, national, or other group affiliation)?

How will you minimize use and adverse downstream impact of the service on subpopulations for which it is not designed or underperformant?

Efficient testing for bias

- Development teams are under multiple constraints
  - Time
  - Money
  - Human resources
  - Access to data
- How can we **efficiently** test for bias?
  - Prioritization
  - Strategic testing

Privacy

How will you improve your training data and performance, while maximizing the confidentiality of customer-submitted content?

Privacy in ML

- Privacy for highly sensitive data – model training and analytics using secure enclaves, homomorphic encryption, federated learning/on-device learning, or a hybrid
- Privacy-preserving model training, robust against adversarial membership inference attacks (dynamic settings + complex data/model pipelines)
- Privacy-preserving mechanisms for data marketplaces

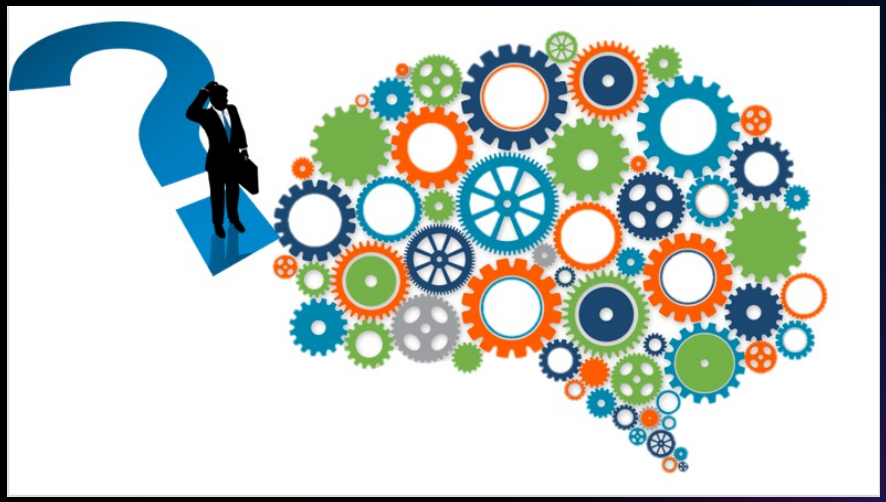
Explainability

Will the service, either alone or as a component of a larger customer system, be required by law (e.g., GDPR), customers, or users to explain any ML-based inferences it makes?

If so, who will consume the explanations, and how will the service track the quality of the explanations?

Explainability in ML

- Actionable explanations
- Balance between explanations & model security
- Robustness of explanations to failure modes (interaction between ML components)
- Application-specific challenges
  - Conversational AI systems – contextual explanations
  - Gradation of explanations
- Tools for explanations across AI lifecycle
  - Pre- and post-deployment for ML models
  - Model developer vs. end user–focused



Robustness

What risks could be posed to customers by maliciously crafted inputs, including deepfakes and sources constructed by generative adversarial networks (GANs), and how will you mitigate them?

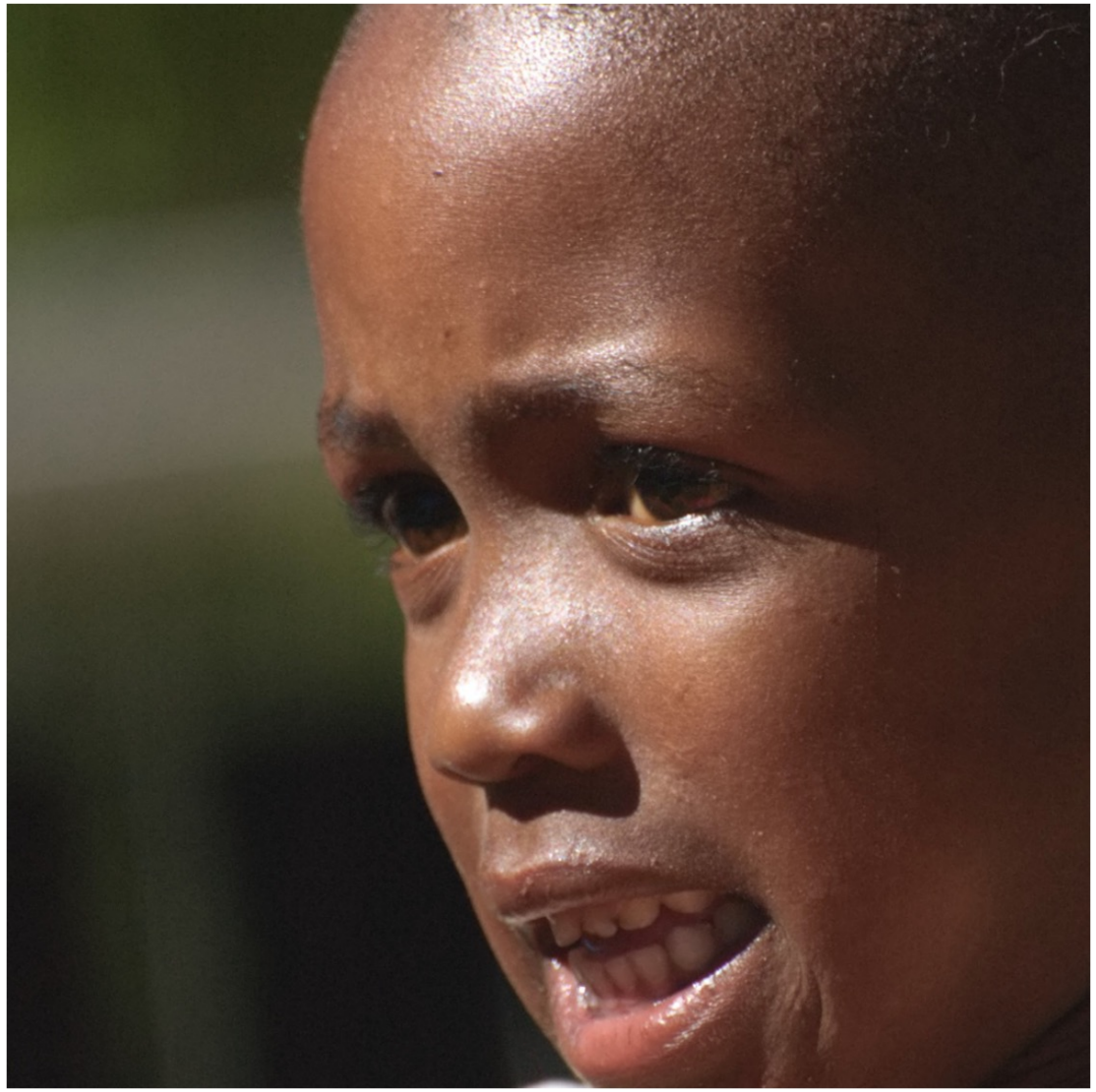
“Artificial Intelligence Is Coming for Our Faces: Trained for a week on a massive data set of portraits, a neural network spits out striking images of nonexistent people”

—Wired, June 2019

whichfaceisreal.com

whichfaceisreal.com

whichfaceisreal.com



whichfaceisreal.com

Transparency

What information about the service’s training, performance, and biases will your customers and the public want to know?

Governance

How do you enforce and ensure these responsible AI practices are being carried out amongst all stakeholders?

Beyond accuracy

- Performance and cost
- Fairness and bias
- Transparency and explainability
- Privacy
- Security
- Safety
- Robustness

Process best practices

Identify product goals

Get the right people in the room

Identify stakeholders

Policy

Select a fairness approach

Technology

Analyze and evaluate your system

Mitigate issues

Monitor continuously and determine escalation plans

Auditing and transparency

Lessons learned

- Testing for unseen areas amongst intersectionality is key
- Taking into account confidence scores/thresholds and error bars when measuring for biases is necessary
- Representation matters
- Transparency, reproducibility, and education can promote change
- Confidence in your product’s fairness requires fairness testing
- Fairness testing has a role throughout the product iteration lifecycle
- Contextual concerns should be used to prioritize fairness testing

Thank you!

Dr. Nashlie Sephus
Principal Tech Evangelist