Responsible AI

Susan Kennedy, Ph. D. | *Assistant Professor*
Department of Philosophy | *Santa Clara University*
Web: susan-kennedy.com

Santa Clara University

> “Machine intelligence is the last invention that humanity will ever need to make”

Nick Bostrom
Philosopher, University of Oxford

Source: *“What happens when our computers get smarter than we are?" TED Talk, March 2015.*

17 goals on the United Nations' 2030 Agenda for Sustainable Development:

- Ending poverty and world hunger
- Improving health and education
- Reducing inequality and injustice
- Clean water and sanitation
- … etc.

Promising Applications of **TinyML**

Industry

Environment

Humans

Microsoft's disastrous Tay experiment shows the hidden dangers of AI

Amazon scraps secret AI recruiting tool that showed bias against women

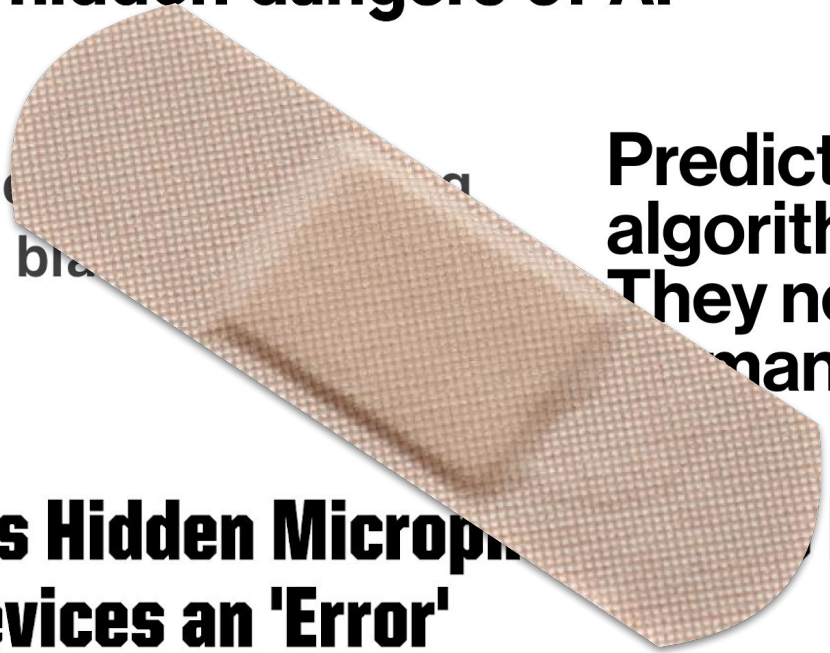
Predictive policing algorithms are racist. They need to be dismantled.

Google Calls Hidden Microphone in Its Nest Home Security Devices an 'Error'

Microsoft's disastrous Tay experiment shows the hidden dangers of AI

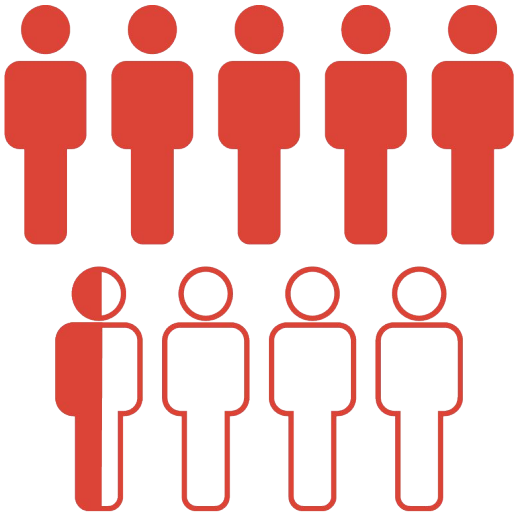
Amazon scraps s… …g tool that showed bi… women

Predictive policing algorithms are racist. They need to be …mantled.

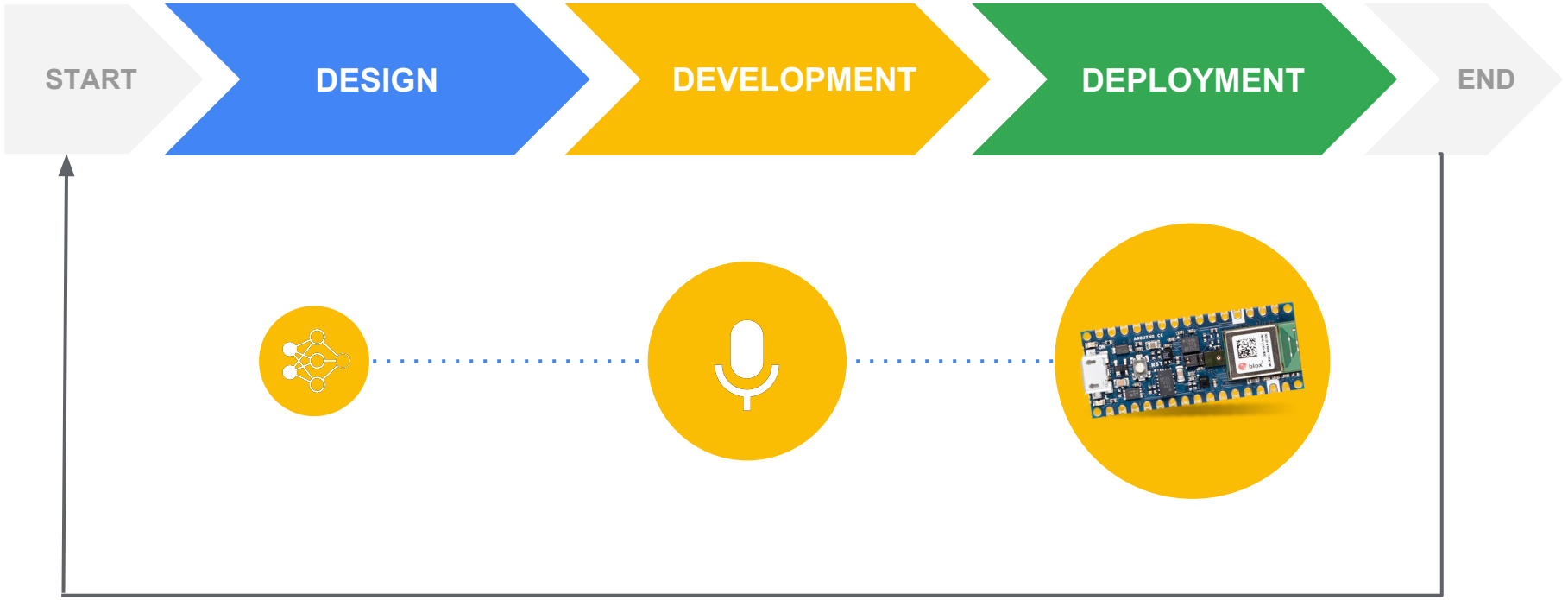


Google Calls Hidden Microph… Nest Home Security Devices an 'Error'

Pew Research shows that **65% of Americans** believe that **companies “often fail to anticipate how their products and services will impact society”**



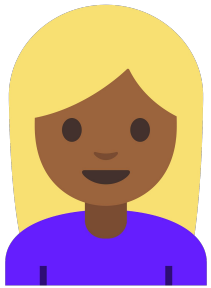
Embedding Ethics Throughout the Workflow



Responsible AI: Design

Stakeholder Analysis

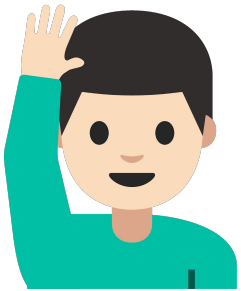
Direct



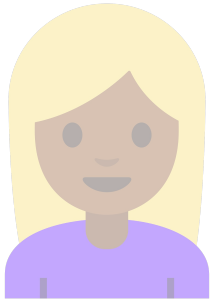
aka the “User(s)”

Stakeholder Analysis

Indirect

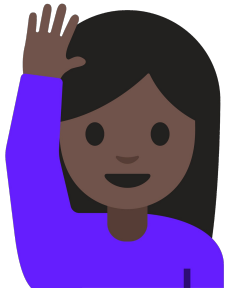


Direct



aka the “User(s)”

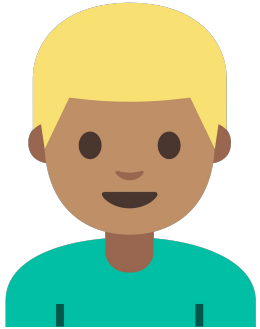
Indirect



What do the stakeholders **value**?

Direct (Doctor)

- Accurate diagnosis
- Training/skill set
- Ease of use
- Research advances



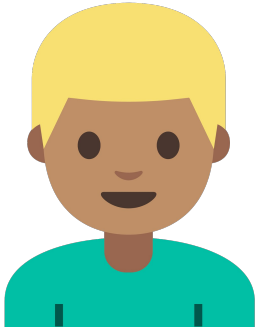
Indirect (Patient)

- Personal care
- Being informed / autonomy
- Trust
- Privacy

Do **value tensions** arise?

Direct (Doctor)

- **Accurate diagnosis**
- Training/skill set
- Ease of use
- Research advances



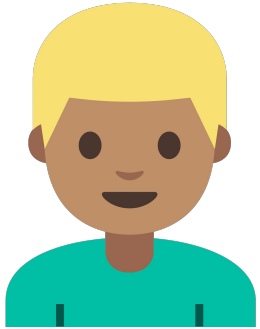
Indirect (Patient)

- Personal care
- **Being informed / autonomy**
- **Trust**
- Privacy

Do **value tensions** arise?

Direct (Doctor)

- Accuracy
- Training/skill set
- Ease of use
- **Research advances**



Indirect (Patient)

- Personal care
- Being informed / autonomy
- Trust
- **Privacy**

Which type of error is most harmful?

	Actual Disease = Yes	Actual Disease = No
Predicted Disease = Yes	True Positive	False Positive ***Type 1 Error***
Predicted Disease = No	False Negative ***Type 2 Error***	True Negative

Responsible AI: Development

The “**garbage in, garbage out**” problem

Bias: Defining the **Target Variable**

Using **biometric** sensors for a health wearable device, how should you define ***“healthy”***?

- **Heart rate**
- **Blood pressure**
- **Number of steps**

Bias: **Labeling** the Data

Labels applied to the training data must serve as **ground truth**

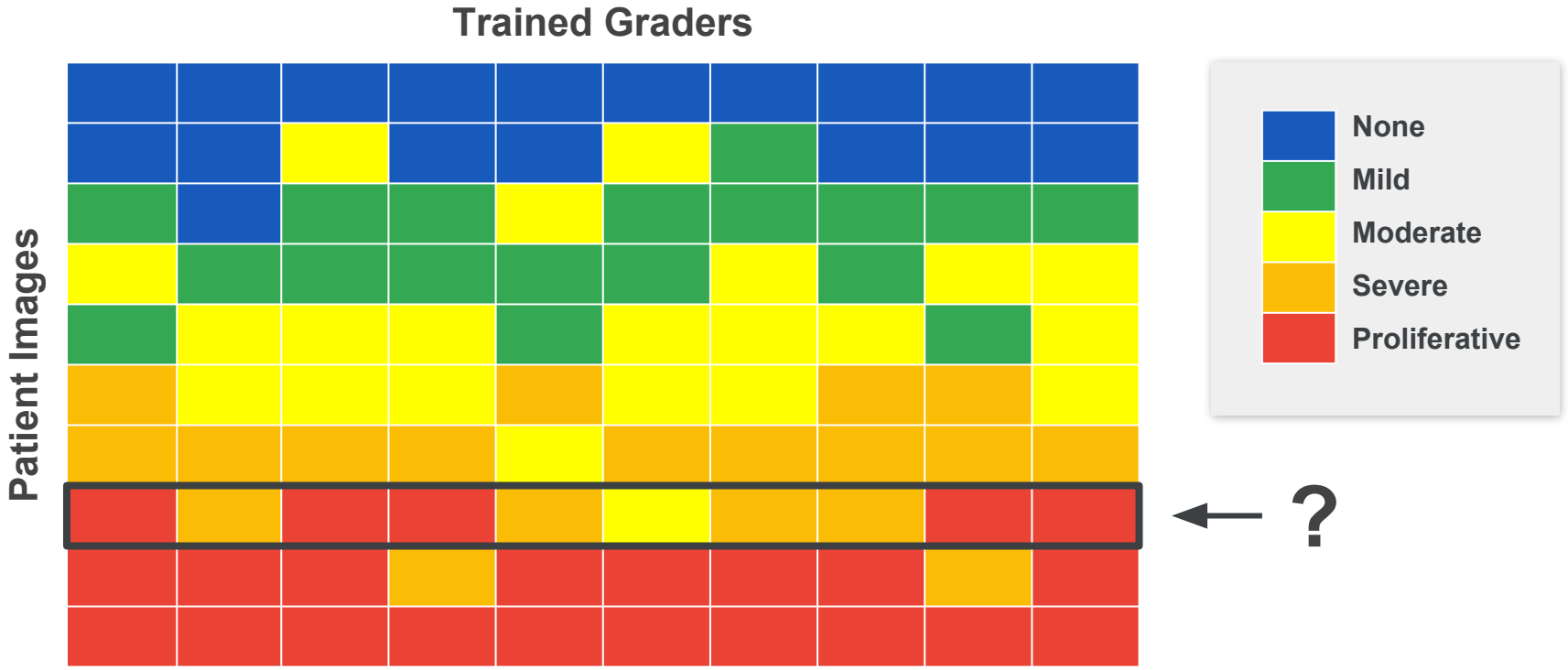
Horse

Human

Human

ERROR

Bias: **Labeling** the Data



Bias: **Prejudice** Reflected in Data

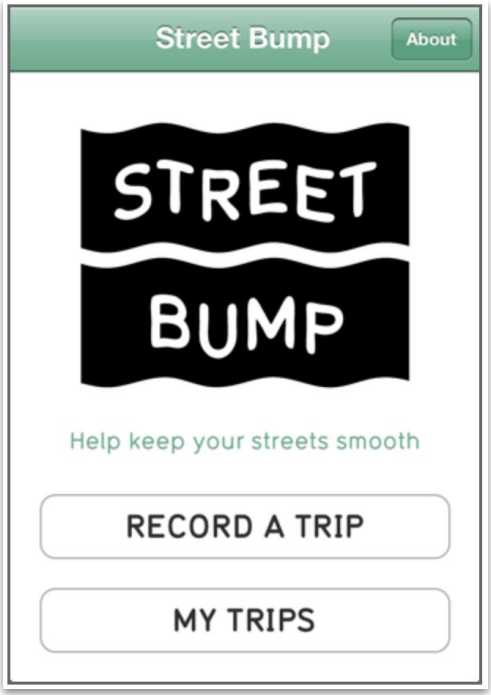


Dataset: *65%* of people cooking are *women*

Algorithm predicts: *85%* of people cooking are *women*

Bias: Measurement **Distortion**

Bias: Sampling the Data



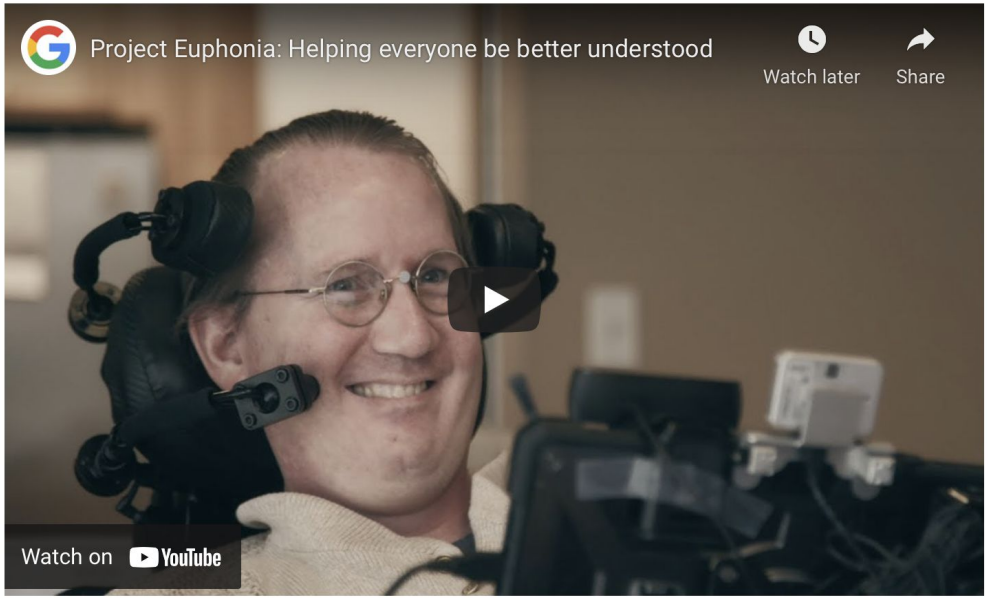
> “...we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big data sets?”

Kate Crawford
Principal Researcher at Microsoft and Professor at NYU Tandon School of Engineering

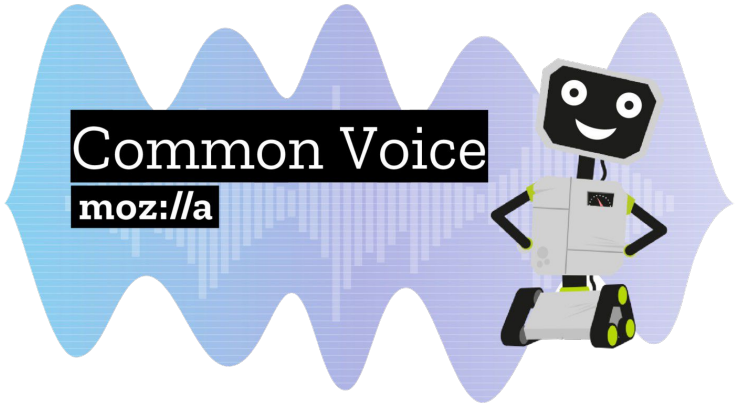
Source: *“The Hidden Biases in Big Data,” Harvard Business Review, April 2013.*

Project Euphonia

Google Research Initiative to **collect** data and **refine** speech recognition algorithms to work better for individuals with speech impairments



Open Datasets and **Crowdsourcing**



Accent

23% United States English, **8%** England English, **5%** India and South Asia, **4%** Australian English, **3%** Canadian English, **2%** Scottish English, **1%** Irish English, **1%** Southern African, **1%** New Zealand English

Age

23% 19–29, **14%** 30–39, **10%** 40–49, **6%** < 19, **4%** 50–59, **4%** 60–69, **1%** 70–79

Source: https://commonvoice.mozilla.org/en/datasets

Industry Solutions: **Datasheets for Datasets**

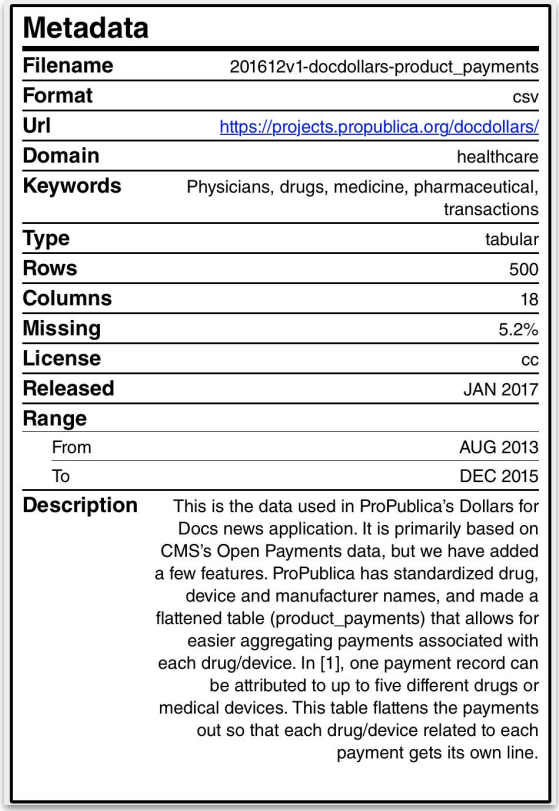
Questions for dataset creators to reflect on during the key stages of the dataset lifecycle:

- ***Motivation***
- ***Composition***
- ***Collection Process***
- ***Preprocessing/ labeling***
- ***Uses***
- ***Distribution***
- ***Maintenance***

paper authored by

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

Industry Solutions: **Data Nutrition Labels**



Metadata	
Filename	201612v1-docdollars-product_payments
Format	csv
Url	https://projects.propublica.org/docdollars/
Domain	healthcare
Keywords	Physicians, drugs, medicine, pharmaceutical, transactions
Type	tabular
Rows	500
Columns	18
Missing	5.2%
License	cc
Released	JAN 2017
Range	
From	AUG 2013
To	DEC 2015
Description	This is the data used in ProPublica's Dollars for Docs news application. It is primarily based on CMS's Open Payments data, but we have added a few features. ProPublica has standardized drug, device and manufacturer names, and made a flattened table (product_payments) that allows for easier aggregating payments associated with each drug/device. In [1], one payment record can be attributed to up to five different drugs or medical devices. This table flattens the payments out so that each drug/device related to each payment gets its own line.

DATA NUTRITION PROJECT

A standard label that highlights the **"key ingredients"** of a dataset:

- ***Provenance***
- ***Metadata***
- ***Missing units***
- ***Variables***

Unfairness in ML

Model exhibits **discriminatory biases**, perpetuates **inequality** or performs less well for historically **disadvantaged groups**

- ***All ML discriminates*** (it just means to recognize a distinction, differentiate)
- Fairness is concerned with **wrongful** discrimination


Age
Disability
Medical History
Marital Status
Protected Classes
Sex
Race
Religion
National Origin

Group Unawareness

Sensitive attributes are **not** included as features of the data (e.g. race, gender)

Pro: Avoids disparate treatment

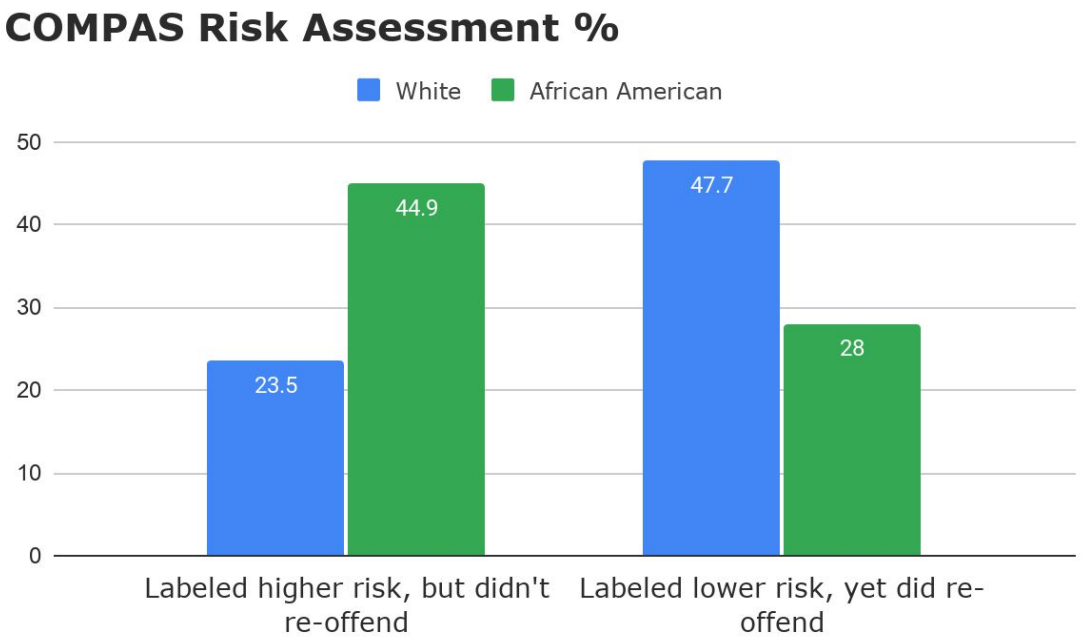
Con: Possibility of highly correlated features that are proxies of the sensitive attribute

Equal Accuracy

	Actually Healthy = Yes	Actually Healthy = No
Predicted Healthy = Yes	***True Positive***	False Positive
Predicted Healthy = No	False Negative	***True Negative***

The percentage of correct classifications should be the same for all individuals

Problem with Equal Accuracy



Northpointe's COMPAS Recidivism Prediction Tool

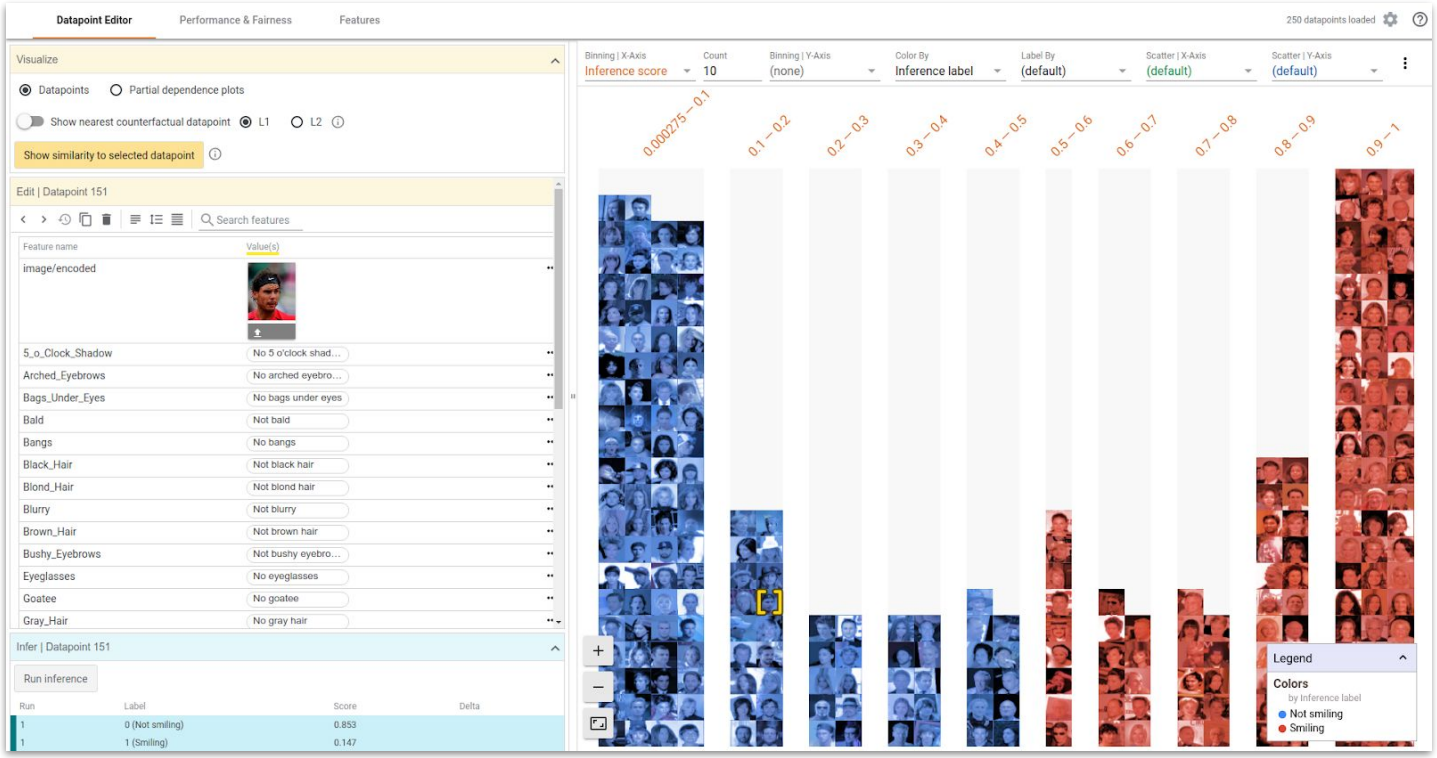
Industry Solutions: **Bias Testing Toolkits**

IBM Research Trusted AI

AI Fairness 360

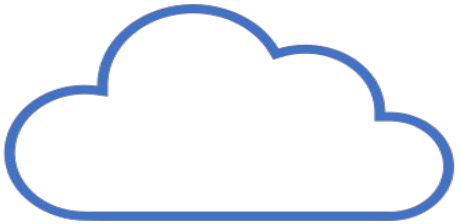
This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Google's **What-If Tool**



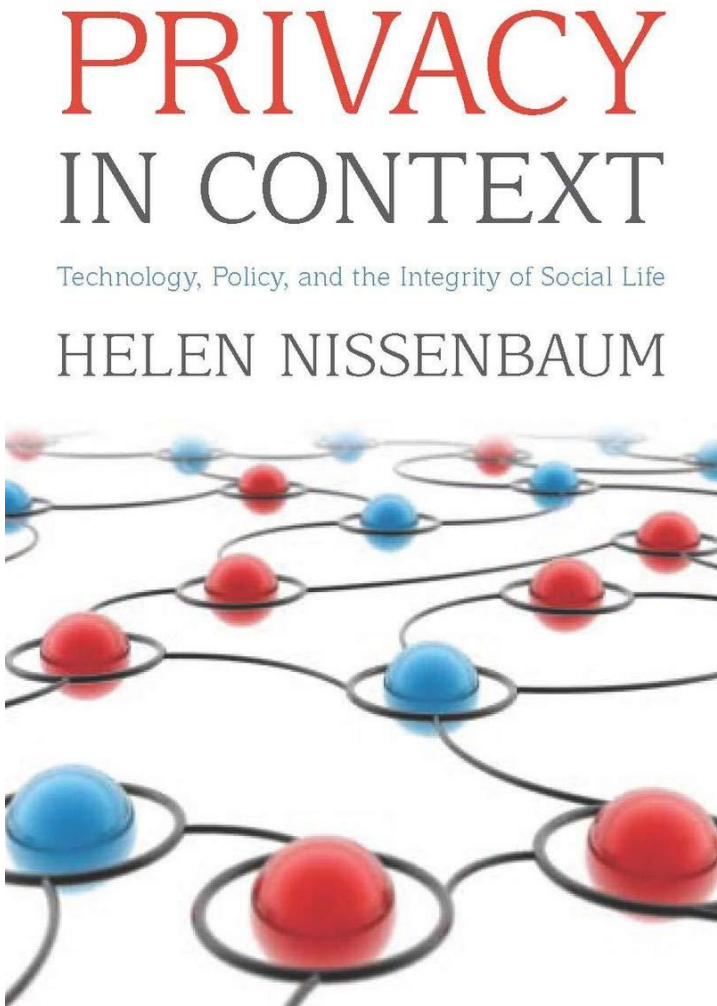
Responsible AI: Deployment

Privacy preserving?



Privacy in Context

- Context shapes our expectations of privacy
- Privacy is a right to the appropriate flow of personal information (contextual integrity)
- Privacy can either be preserved or violated by the introduction of new technologies



Context-Relative Informational Norms

Context: What is the prevailing context?

Actors: Who are the subjects, senders and recipients of information?

Attributes: What is the type or nature of information?

Transmission principle: What are the constraints on the flow of information?

If the new practice results in any changes to these features, the practice is **flagged** as violating privacy

Context: Establish the prevailing context

Actors: Establish key actors

Attributes: Ascertain what attributes are affected

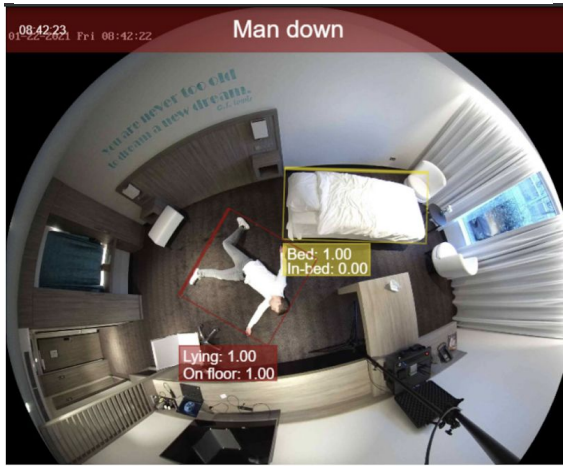
Transmission principle: Establish changes in transmission principles

Second Chances

Practices that are flagged as violating privacy may still be desirable all things considered

- Does the new practice provide better support for contextual values?
- Does it promote autonomy?
- Does it improve power relations?
- Does it create a fair distribution of costs and benefits

Kepler Night Nurse™



Context: Caregiving in health care facilities

Actors: Patient, Caregivers, **Annotators at Kepler**

Attributes: Video and Images of users

Transmission principle: Caregiver's mandate, confidential

How can **privacy be preserved**?

- **Minimize**
  - Avoid collecting unnecessary data, and dispose or delete data periodically
- **Protect**
  - Use encryption techniques to protect data
- **Informed consent**
  - Be transparent with users about how their data is being collected and used
- **Map the flow of information**
  - Context, the type of information, and who has access, etc.

Responsible AI: Post-Deployment

Sustainability *of* TinyML

	Microprocessor	*vs*	Microcontroller
Platform		>	
Power	30W–100W	**~1000X**	150µW–23.5mW

Environmental Impact

Operational (Recurring)

- Product use
- Operational energy consumption
- e.g., training, inference

Capital (one-time)

- Supply chain for raw materials
- Chip manufacturing
- e.g., hardware production, transport, end-of-life processing

> Development that meets the needs of the **present** without compromising the ability of future generations to meet their ***own*** needs

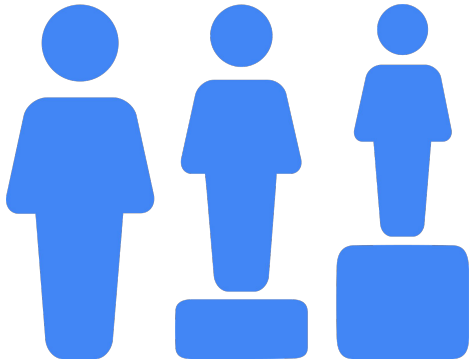
World Commission on Environment and Development
Brundtland Report 1987

Source: World Commission on Environment and Development. (1987). Report of the World Commission on Environment and Development: *Our Common Future*. UN Documents: Gathering a Body of Global Agreements.

Equitable Resource Distribution

Equity

Fair distribution of burdens, benefits, resources, etc.

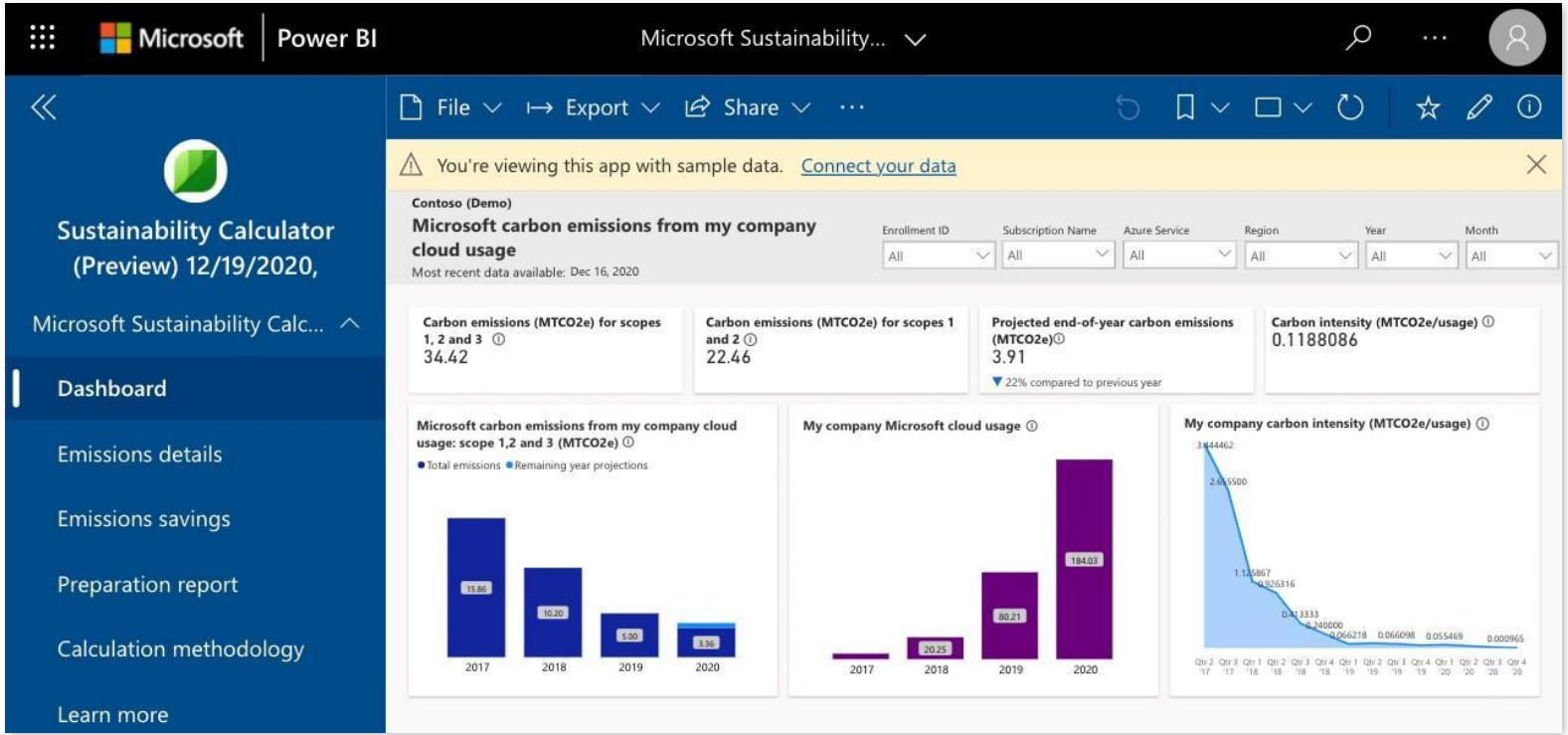


- ***Intra***generational justice
  Within a generation
- ***Inter***generational justice
  Between generations

Sustainability **Pledges**

Google	Carbon neutral since 2007, carbon free by 2030
amazon	100% renewable energy by 2025, carbon neutral by 2040
Microsoft	Carbon negative by 2030, remove historical carbon emissions by 2050

Sustainability **Calculator**



AI is a science *and* an art form

There is no substitute for critical thinking!

Embedded Ethics