Harvard

Yá'át'ééh 👋

CRESTLEX 3.0

CReating **E**ffective **ST**em
Learning **EX**periences

with **Navajo Tech**

> “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.*

SUSTAINABLE DEVELOPMENT GOALS

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
- Industry, innovation and infrastructure
- ... etc.

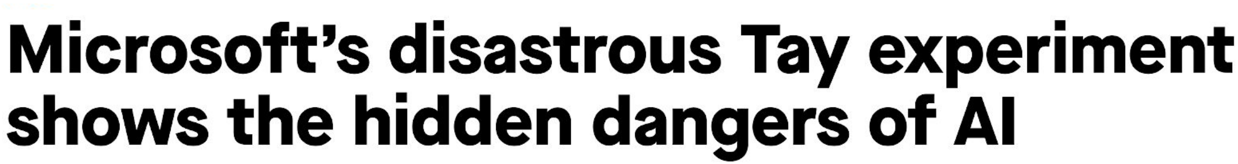
Promising Applications of **TinyML**

Industry

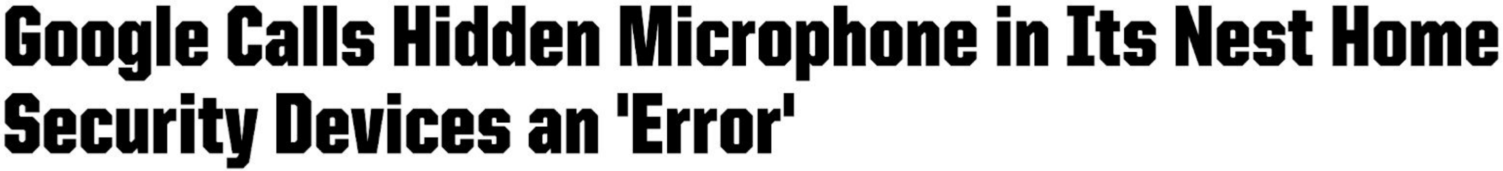
Environment

Humans

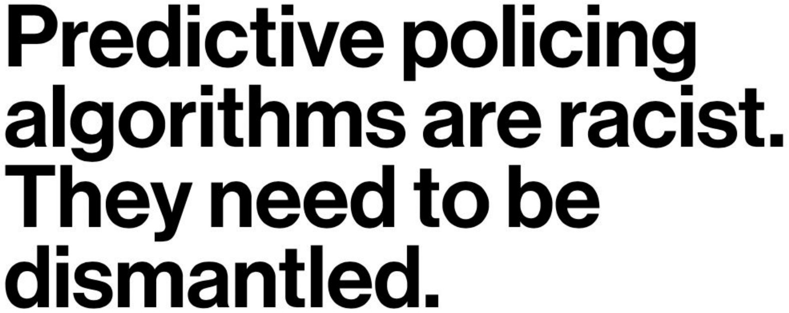
AI Failures



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

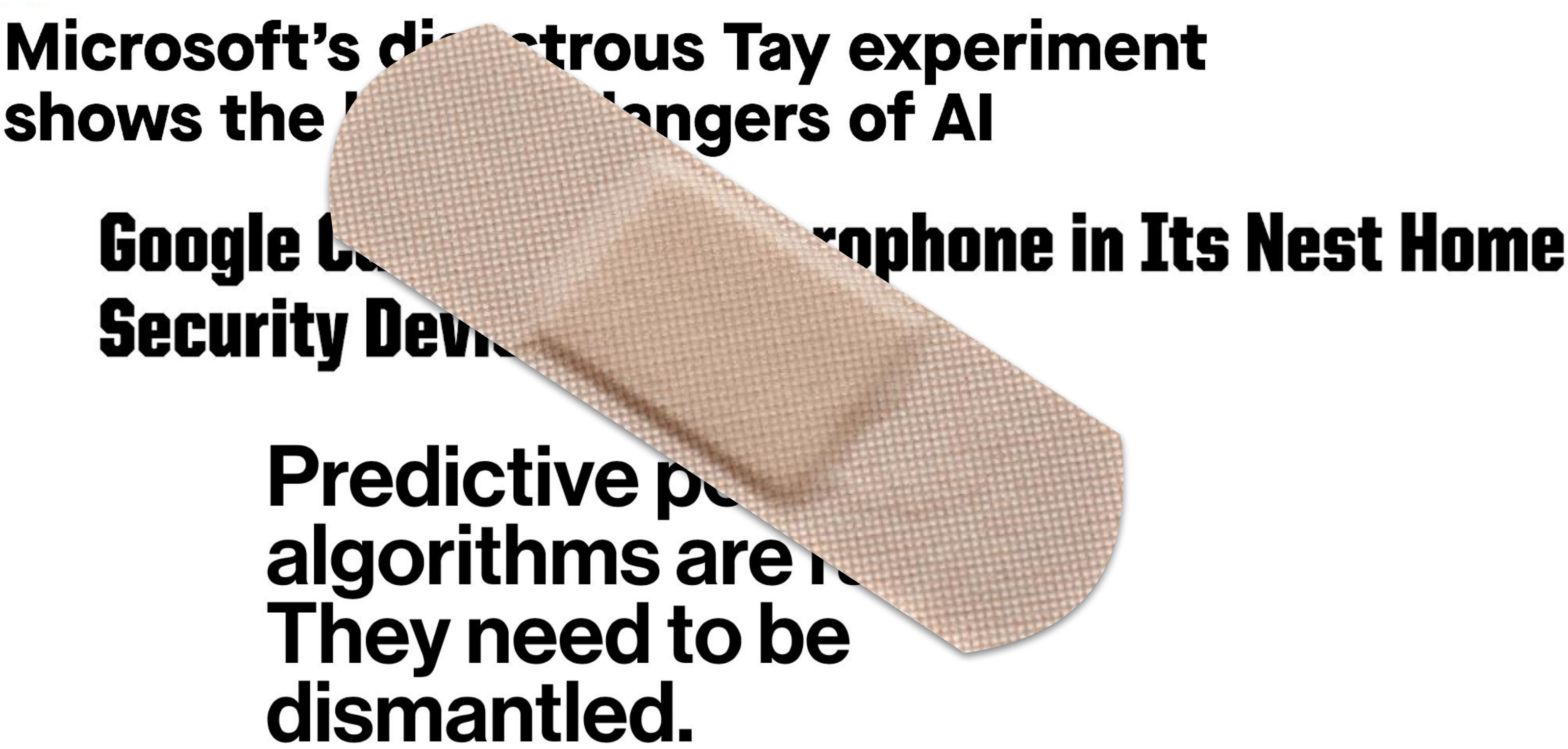


Google Calls Hidden Microphone in Its Nest Home Security Devices an 'Error'



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

AI Failures

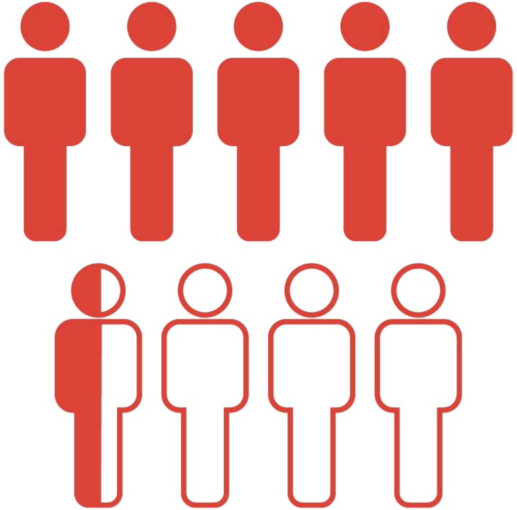


Microsoft's di trous Tay experiment
shows the angers of AI

Google C rophone in Its Nest Home
Security Devi

Predictive p
algorithms are r
They need to be
dismantled.

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



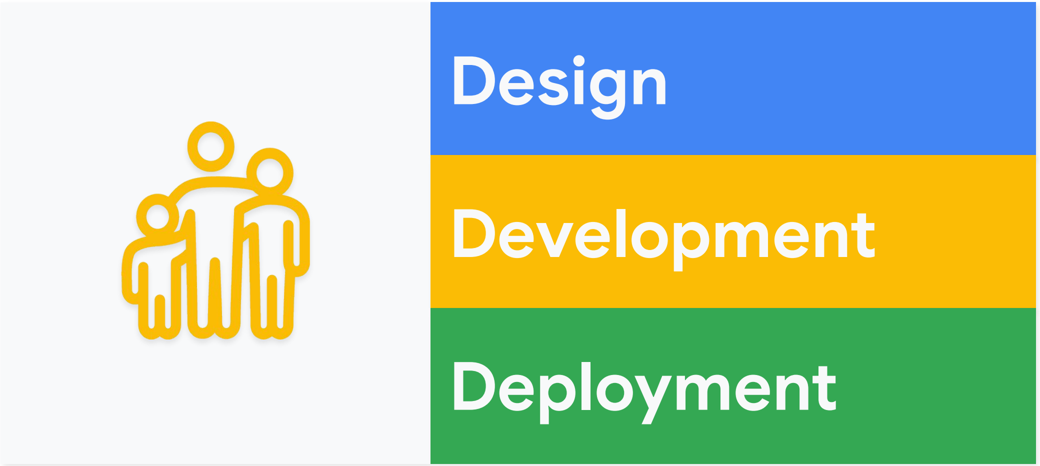
PARTNERSHIP ON AI

amazon DeepMind Google

IBM facebook Microsoft

- **2016:** Partnership on AI founded to ***benefit people and society***

Human Centered AI



Design

Development

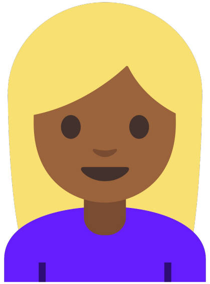
Deployment

Keeping **human values** in the loop throughout **all stages** of a product’s lifecycle

Responsible AI: Design

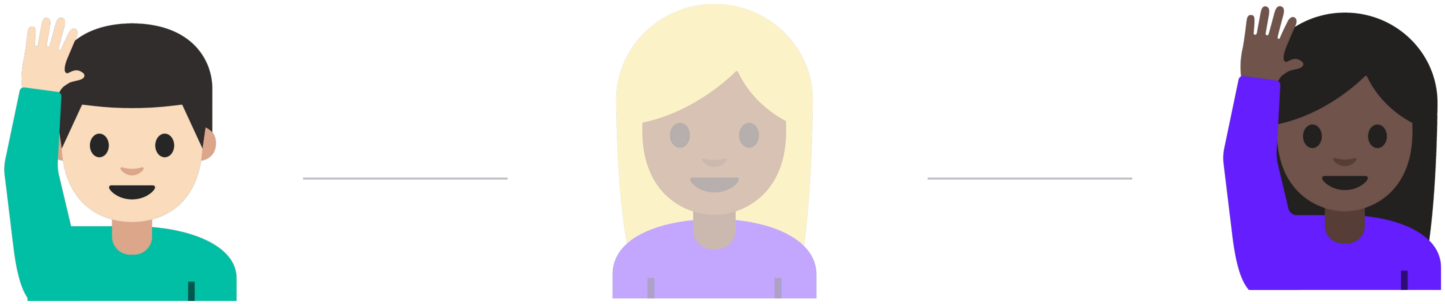
Stakeholder Analysis

Direct



aka the “User(s)”

Stakeholder Analysis

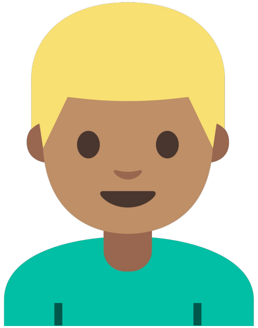


Indirect

What do the stakeholders **value**?

Direct (Doctor)

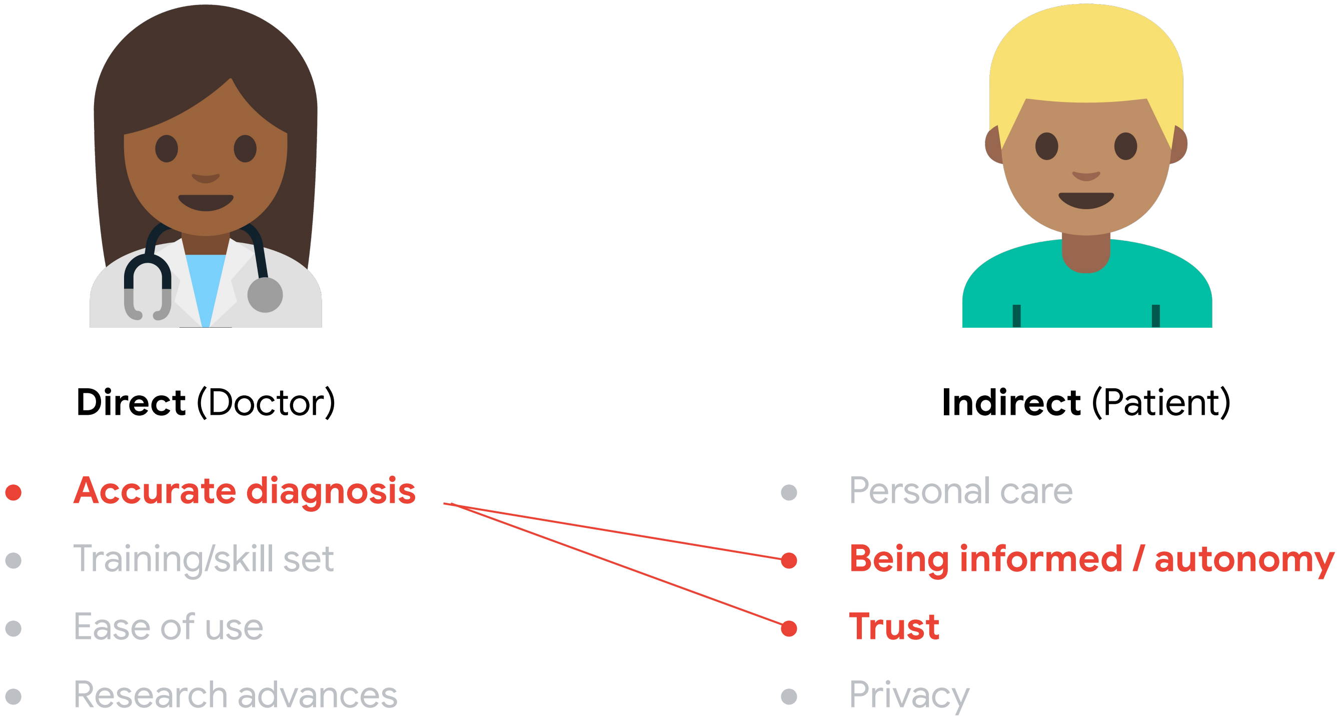
- Accurate diagnosis
- Training/skill set
- Ease of use
- Research advances



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

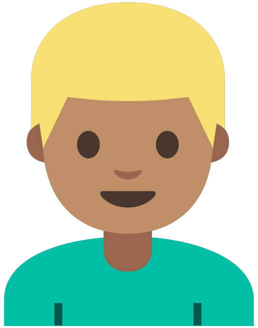
Indirect (Patient)

- 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**

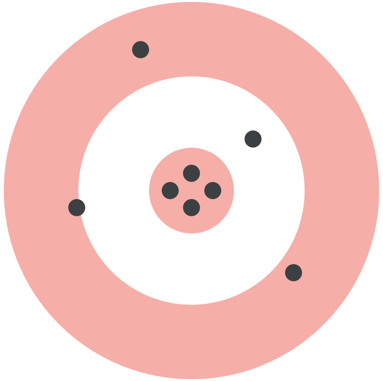
Which type of error is most harmful?

	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

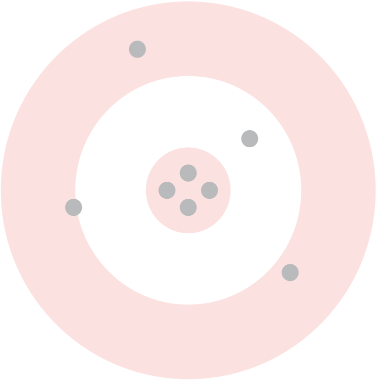
What is bias?

Not all errors
are attributed to bias

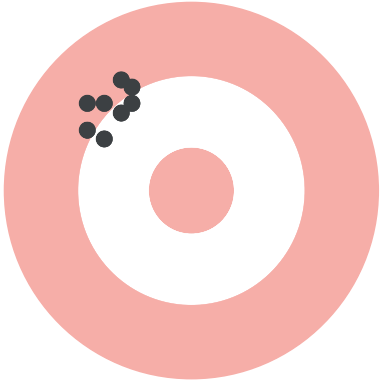


What is bias?

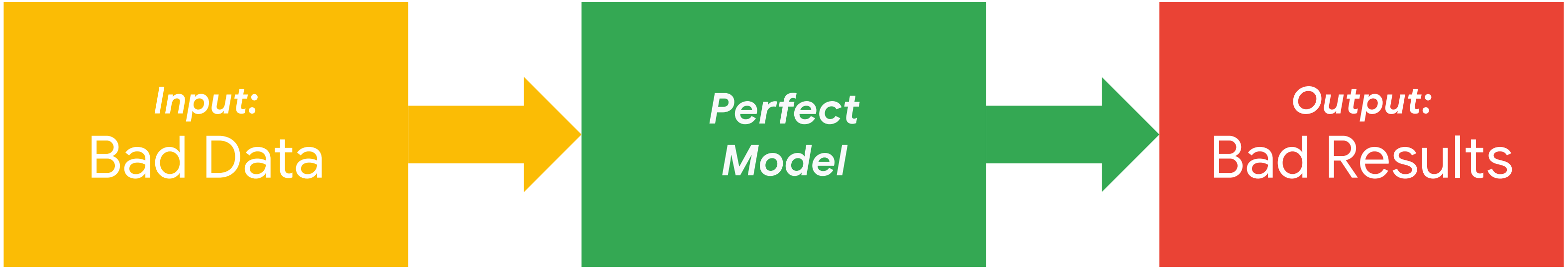
Not all errors are attributed to bias



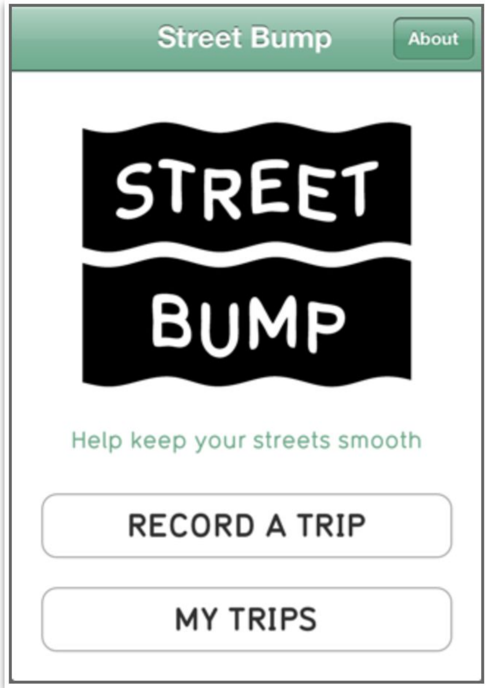
Bias is a deviation in a ***predictable*** (i.e., not random) direction



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



Bias: Sampling the Data



Street Bump

About

STREET BUMP

Help keep your streets smooth

RECORD A TRIP

MY TRIPS

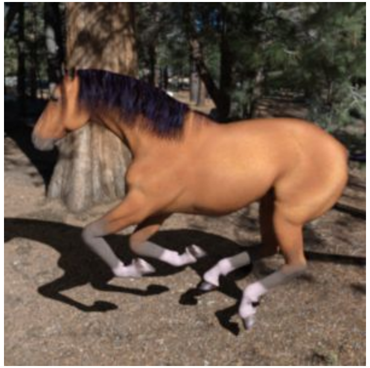
Bias: Defining the **Target Variable**

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

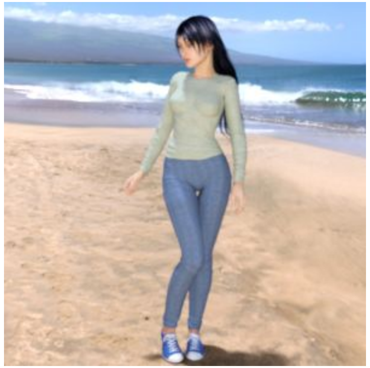
- **Heart rate**
- **Blood pressure**
- **Number of steps**

Bias: **Labeling** the Data

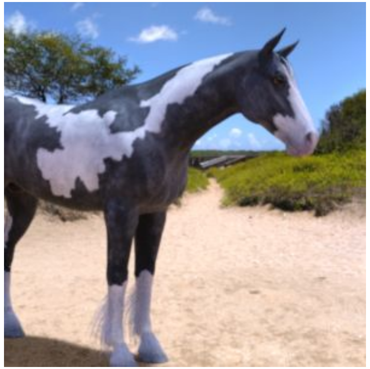
Labels applied to the training data must serve as **ground truth**



Horse



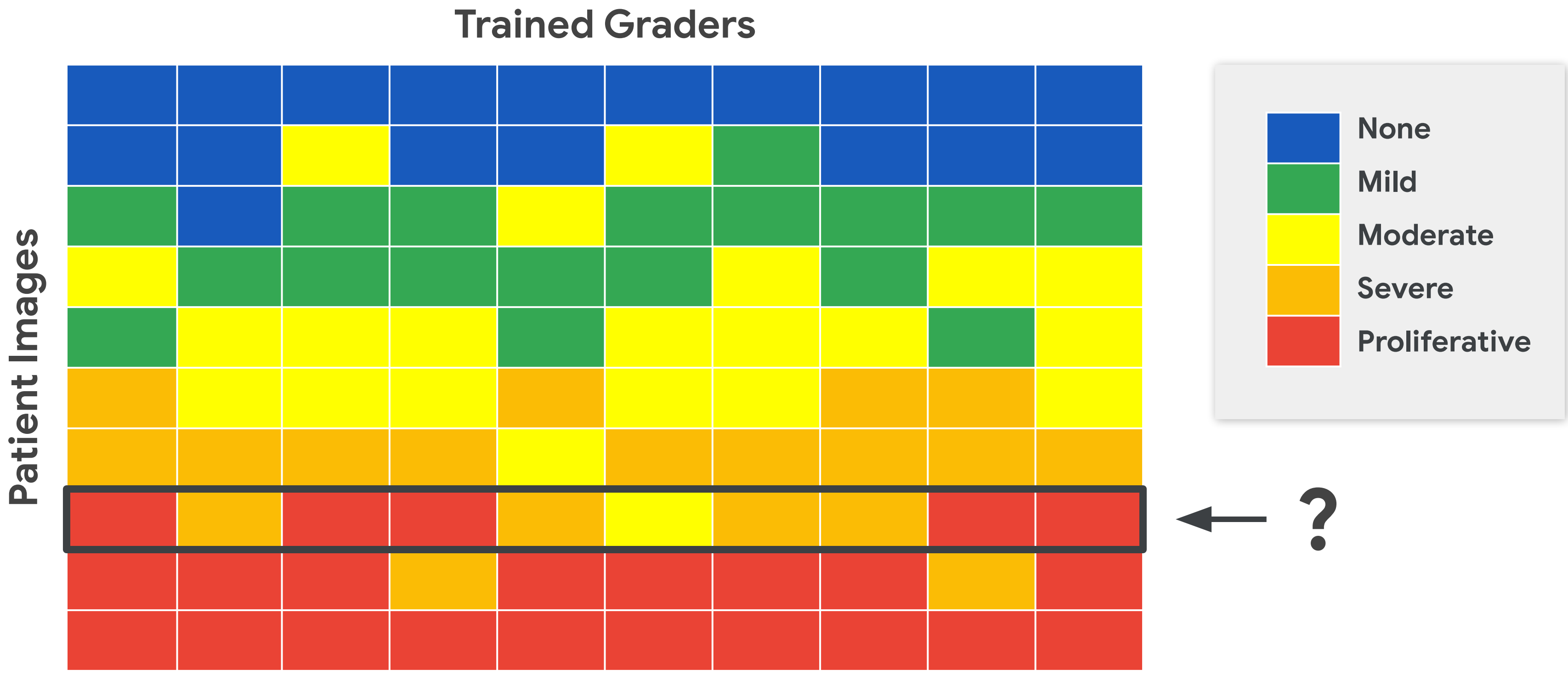
Human



Human

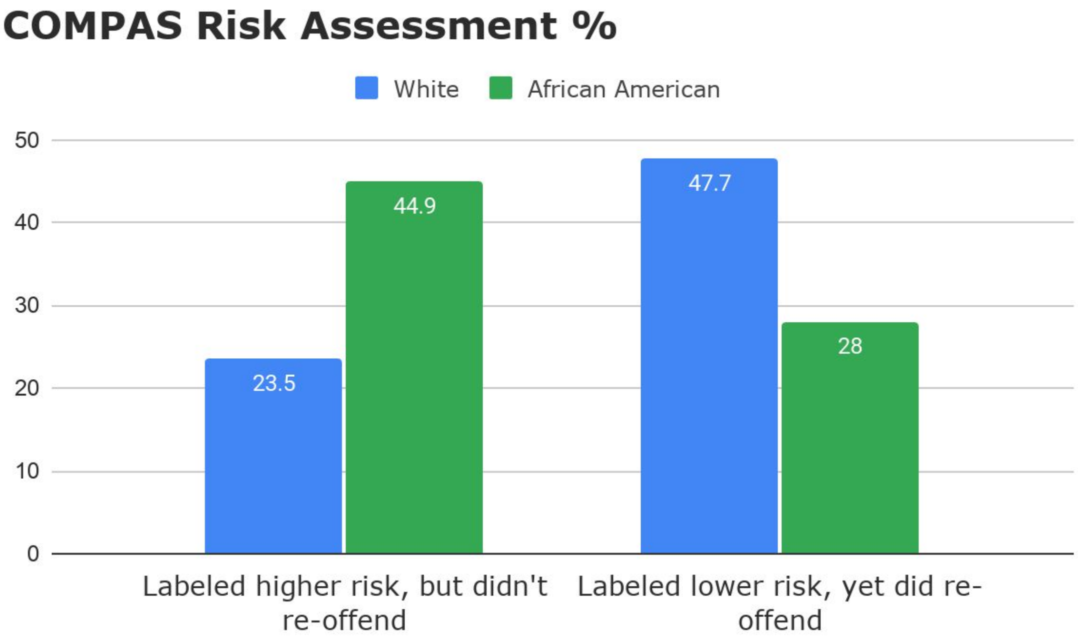
ERROR

Bias: **Labeling** the Data



🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
🟦	🟦	🟨	🟦	🟦	🟨	🟩	🟦	🟦	🟦
🟩	🟦	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩
🟨	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟨	🟨
🟩	🟨	🟨	🟨	🟩	🟨	🟨	🟨	🟩	🟨
🟧	🟨	🟨	🟨	🟧	🟨	🟨	🟧	🟧	🟨
🟧	🟧	🟧	🟧	🟨	🟧	🟧	🟧	🟧	🟧
🟥	🟧	🟥	🟥	🟧	🟨	🟧	🟧	🟥	🟥
🟥	🟥	🟥	🟧	🟥	🟥	🟥	🟥	🟧	🟥
🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥

Bias: **Prejudice** Reflected in Data



Northpointe's COMPAS Recidivism Prediction Tool

Bias: **Prejudice** Reflected in Data

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

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

Designer Solutions

- Carefully **research your users in advance**, be aware of potential outliers
- **Ensure your team** of data scientists and data labelers is ***diverse***
- Where possible, **combine inputs from multiple sources** to ensure data diversity
- **Create a gold standard** for data labeling
- Seek out **domain experts** to review your data

Industry Solutions: **Datasheets for Datasets**

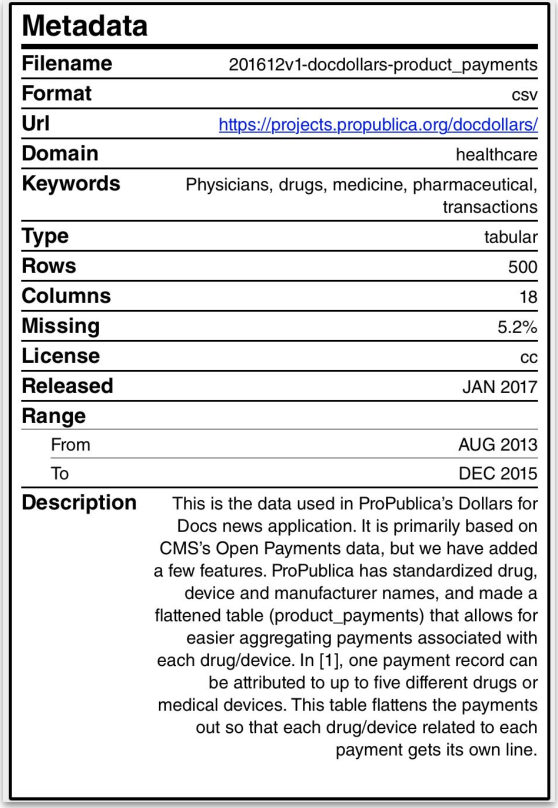
Questions for dataset creators to reflect on during the key stages of the dataset lifecycle:

- ***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***

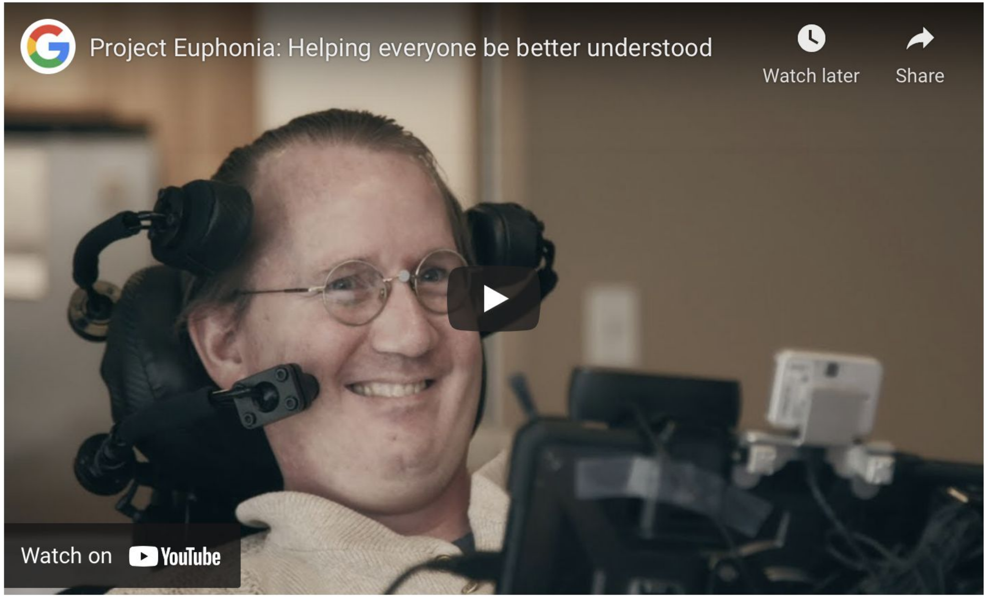
> “…we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big data sets?”

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

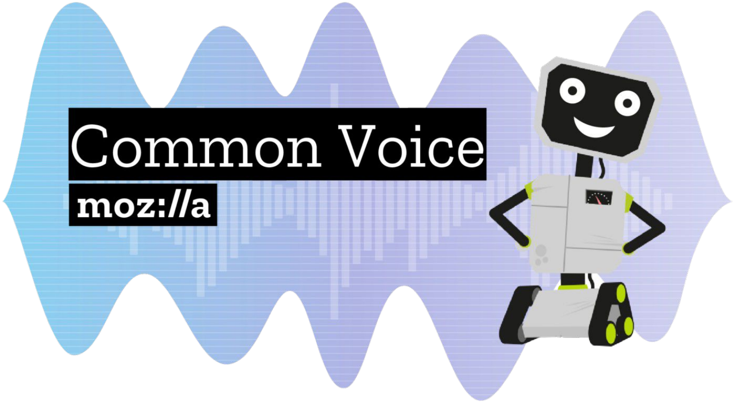
Source: *“The Hidden Biases in Big Data,” Harvard Business Review, April 2013.*

Project Euphonia

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


Project Euphonia: Helping everyone be better understood

Open Datasets and **Crowdsourcing**



Accent

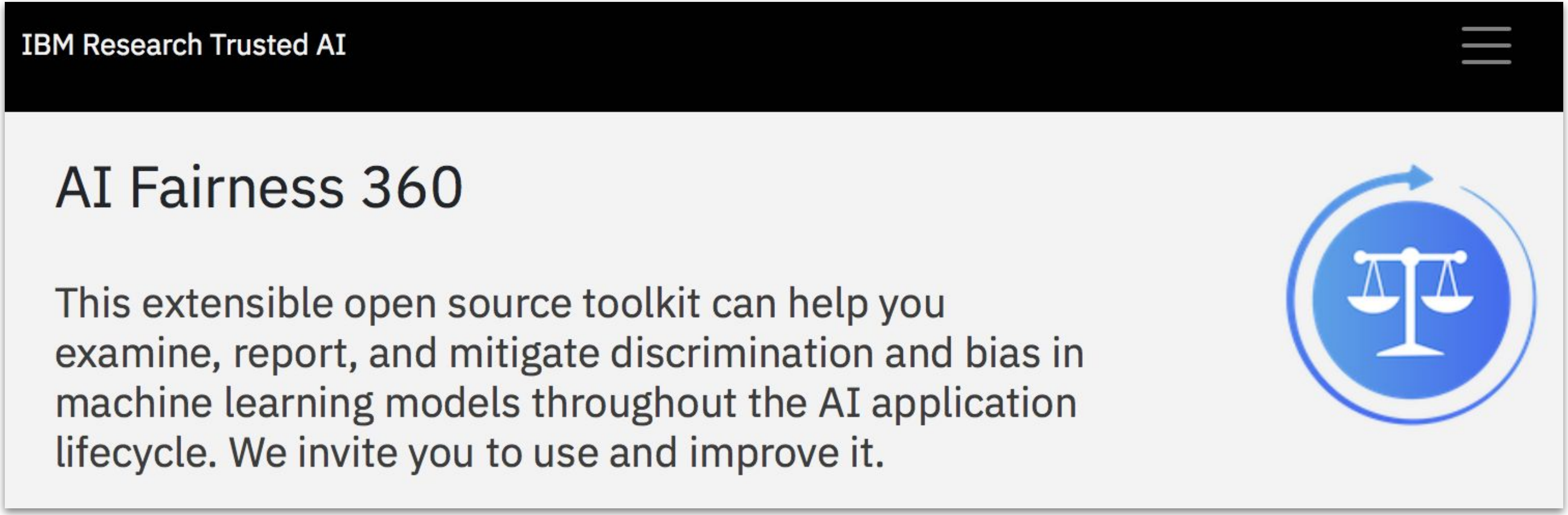
23% United States English, **8%** England English, **5%** India and South Asia, **4%** Australian English, **3%** Canadian English, **2%** Scottish English, **1%** Irish English, **1%** Southern African, **1%** New Zealand English

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: **Bias Testing Toolkits**



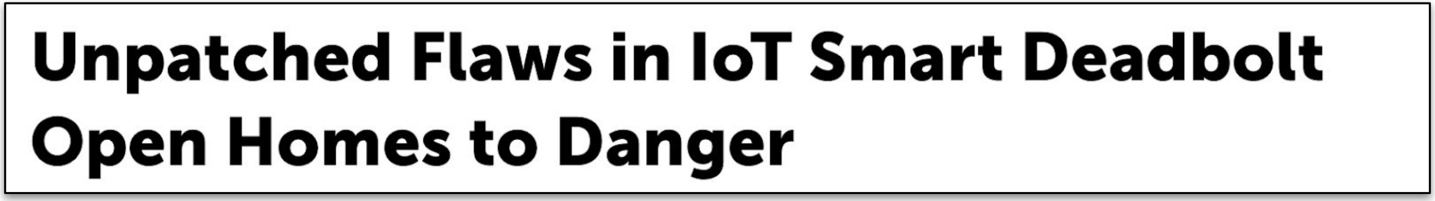
IBM Research Trusted AI

AI Fairness 360

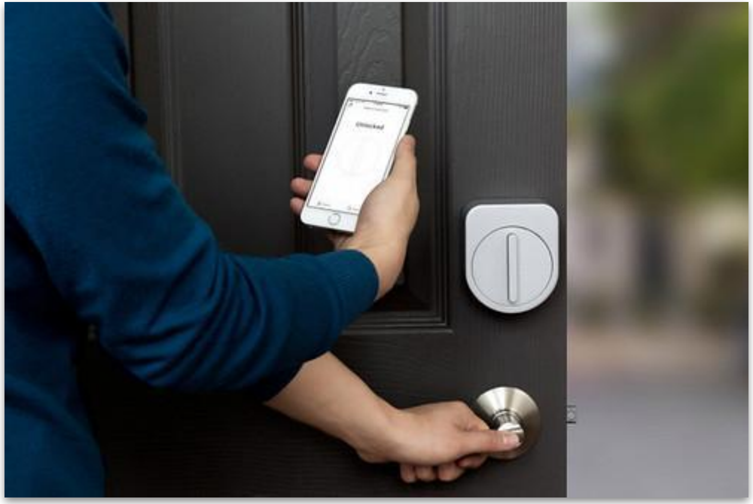
This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Responsible AI: Deployment

Attack: Exploiting **Vulnerabilities**



Unpatched Flaws in IoT Smart Deadbolt Open Homes to Danger



Adversarial Attacks: **TinyML**

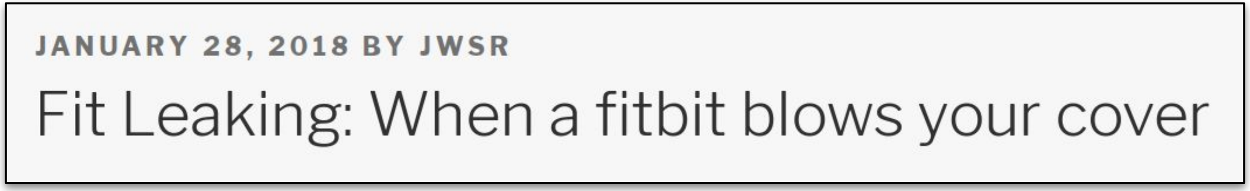
Fooling the machine

failure to trigger wake word

DolphinAttack

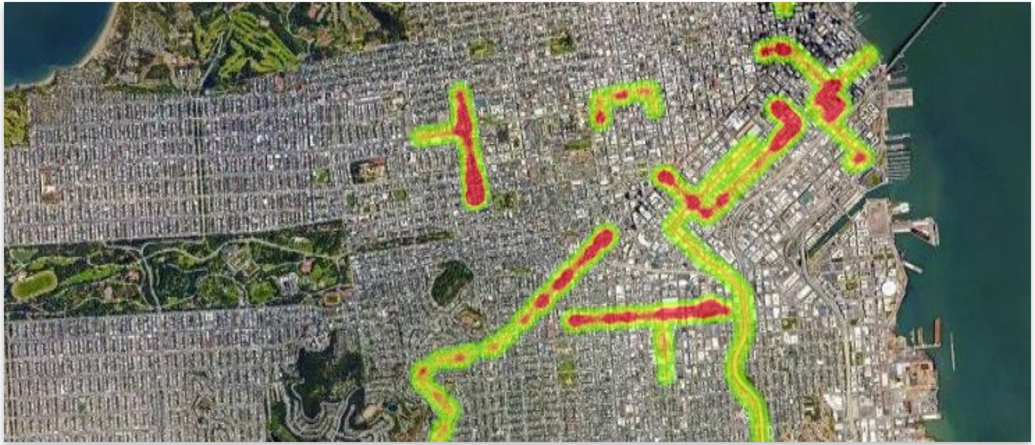
succeeds in triggering wake word

Data Leaks

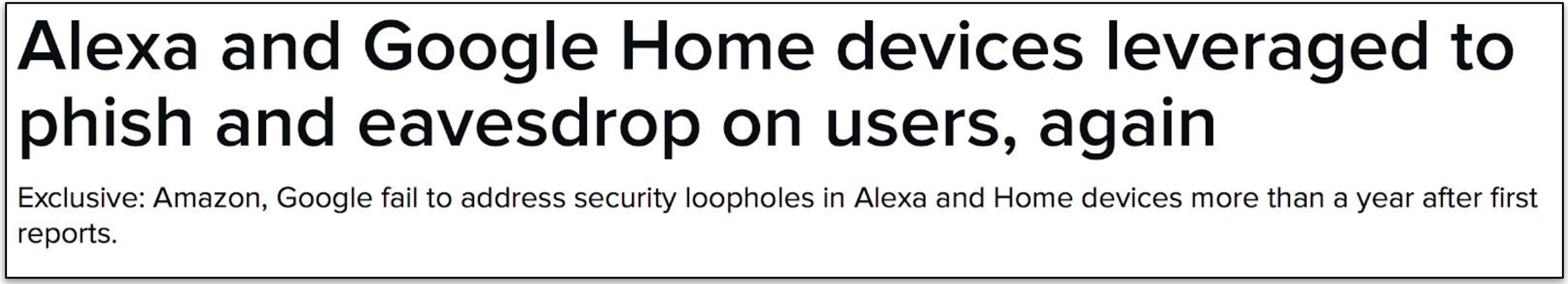


JANUARY 28, 2018 BY JWSR

Fit Leaking: When a fitbit blows your cover

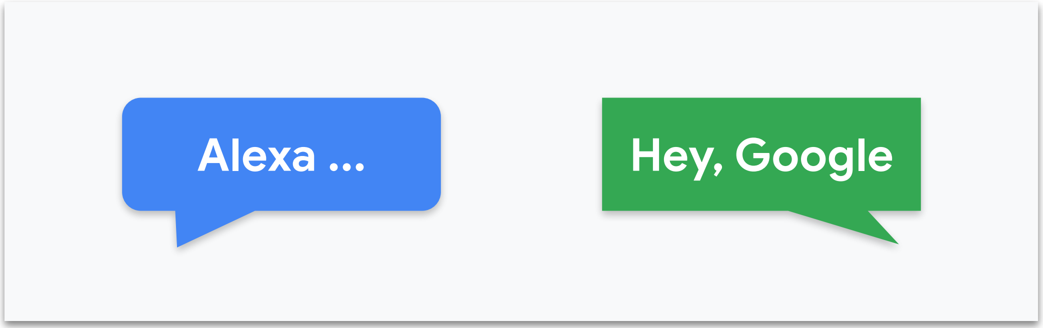


Data Breaches



Alexa and Google Home devices leveraged to phish and eavesdrop on users, again

Exclusive: Amazon, Google fail to address security loopholes in Alexa and Home devices more than a year after first reports.



Why is privacy **valuable**?

- Prevent **information-based harms**
  - Minimize opportunities for hackers to gain inappropriate access to data
- Prevent informational **injustice** and **discrimination**
  - Consider the context, the type of information, and who has access
- Preserve **autonomy** and **human dignity**
  - Obtain informed consent

How can **privacy be preserved**?

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

AI is a science *and* an art form

There is no substitute for critical thinking!