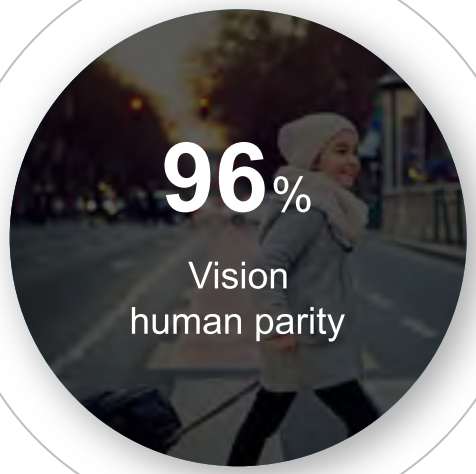
Microsoft AI

Developing AI Responsibly

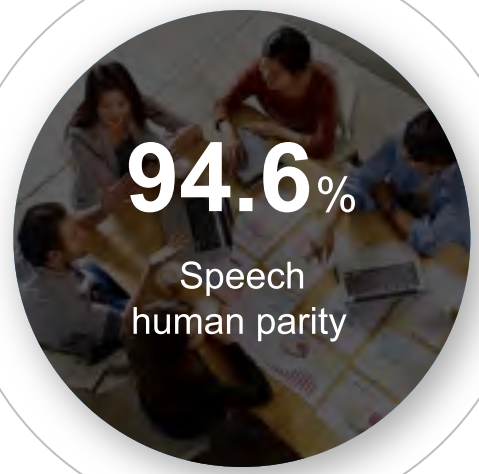
“AI isn’t just another piece of technology. It could be one of the world’s most fundamental pieces of technology the human race has ever created.”

-Satya Nadella

Reaching human parity

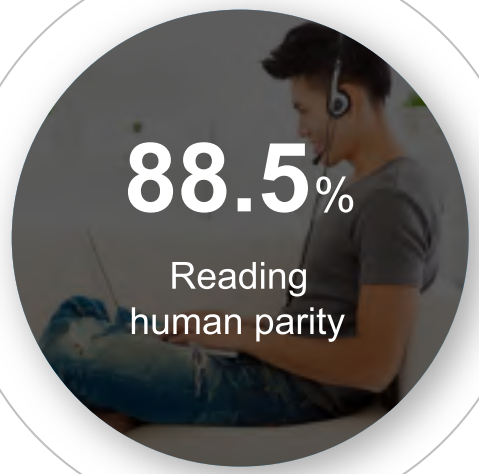


2016



2017

2018



2018

AI will have a considerable impact on business and society as a whole

But this impact raises a host of complex and challenging questions

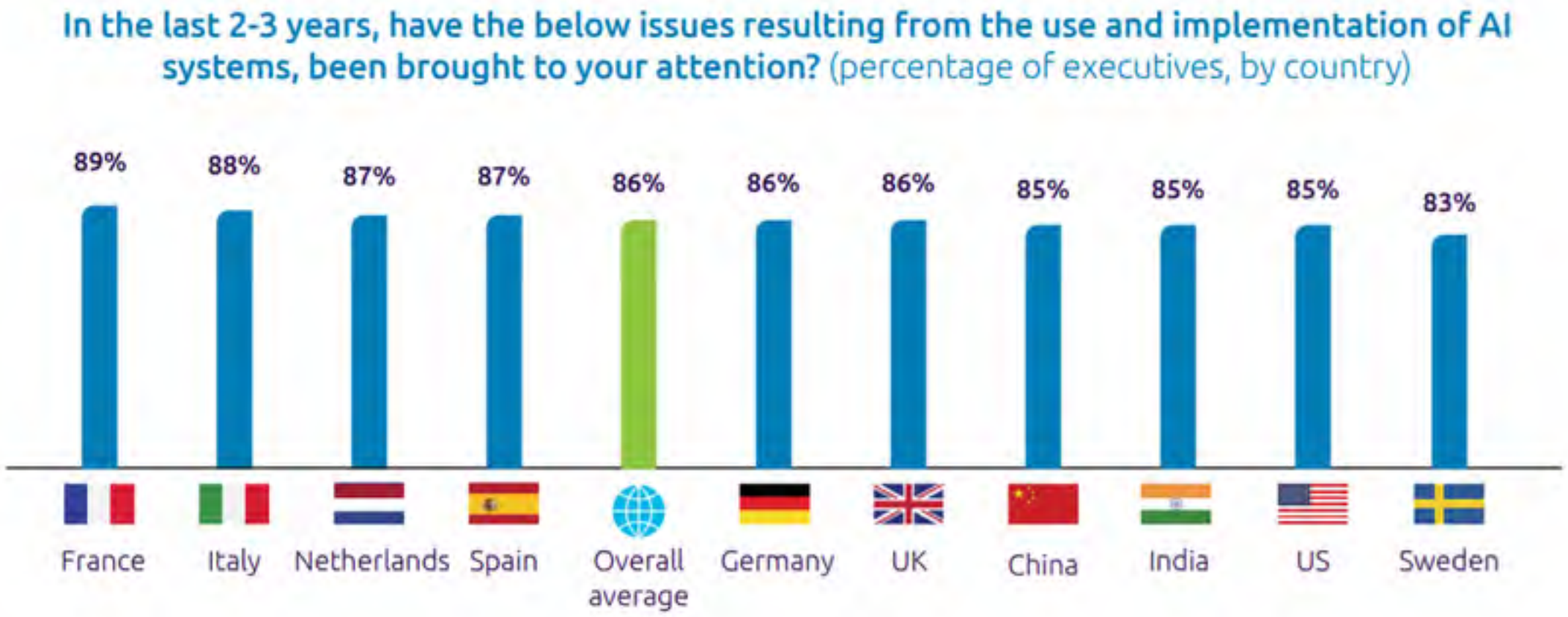
How do we design, build, and use AI systems that create a positive impact on people and society?

How do we best ensure that AI is guarantees key properties such as fairness, privacy and safety?

How can we keep pace with the rate of innovation?

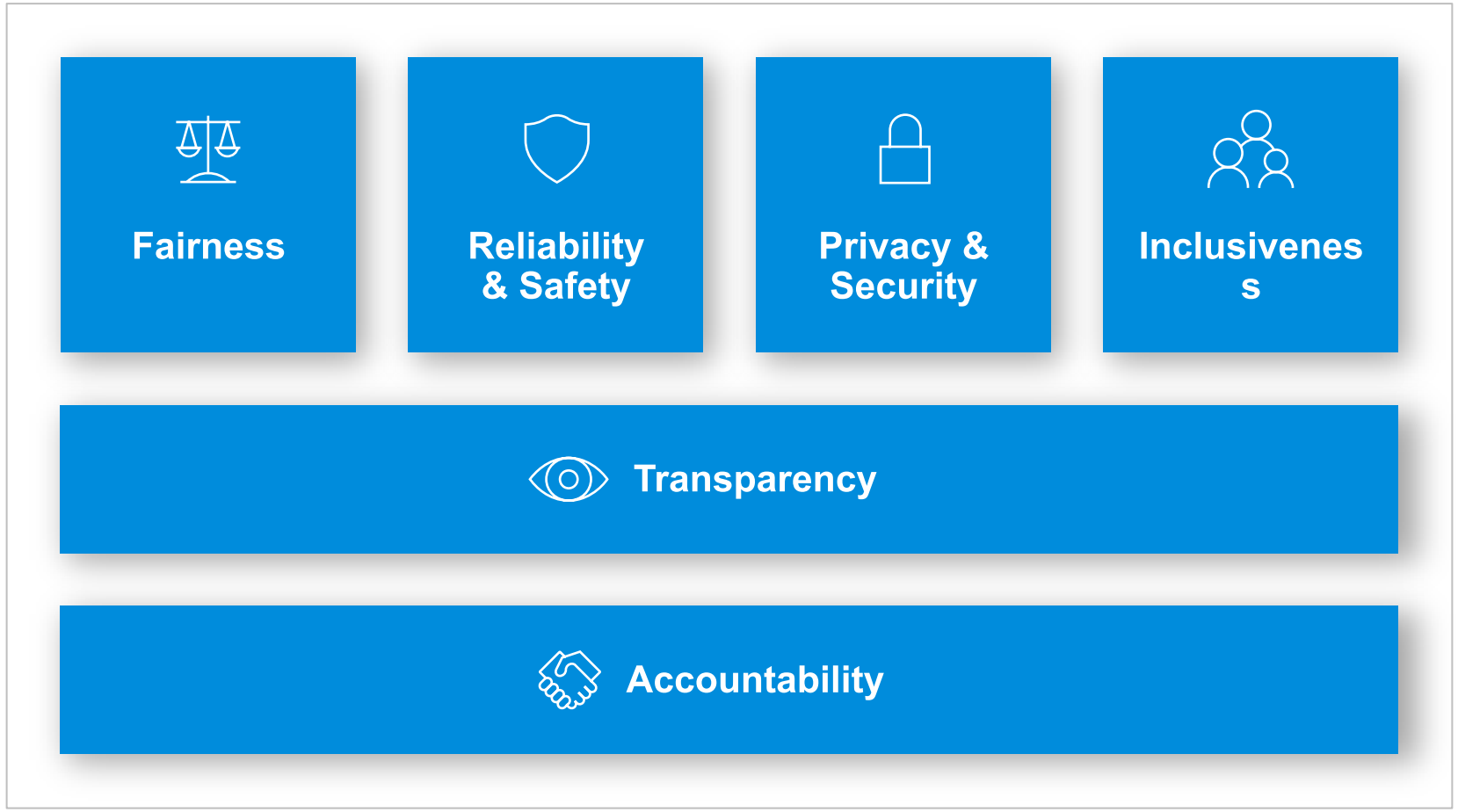
How do we support the diversity of applications?

Nearly nine in ten organizations across countries have encountered ethical issues resulting from the use of AI



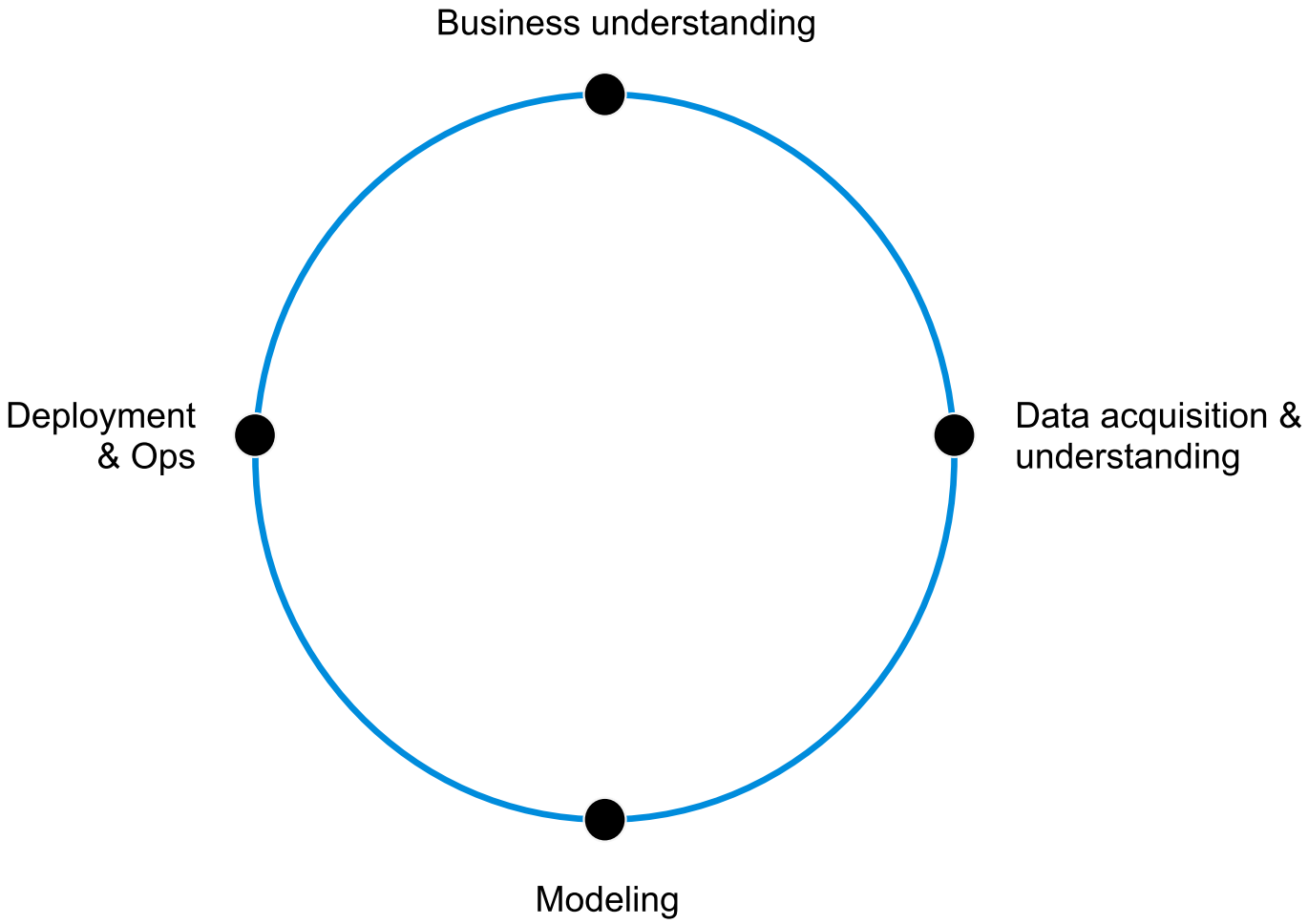
7.5.19 Report from Capgemini

Principles for AI



Putting our ethical principles into action

Responsible AI is a complex and broad topic



State of the Art in Industry

People

Bring in domain expertise and diversity

Process

Focus on processes, best practices, and reviews

Measure

Emphasize analysis and testing

Beyond Accuracy

- Accuracy
- Performance
- Cost
- Fairness
- Inclusiveness
- Privacy
- Security
- Safety
- Reliability

Machine Learning on Azure

Domain Specific Pretrained Models
To reduce time to market


Vision


Speech


Language


Search

Familiar Data Science tools
To simplify model development


PyCharm


Jupyter


Visual Studio Code


Command line

Popular Frameworks
To build machine learning and deep learning solutions


Pytorch


TensorFlow


Scikit-Learn


Onnx

Productive Services
To empower data science and development teams


Azure Databricks


Azure Machine Learning


Machine Learning VMs

Powerful Infrastructure
To accelerate deep learning


CPU

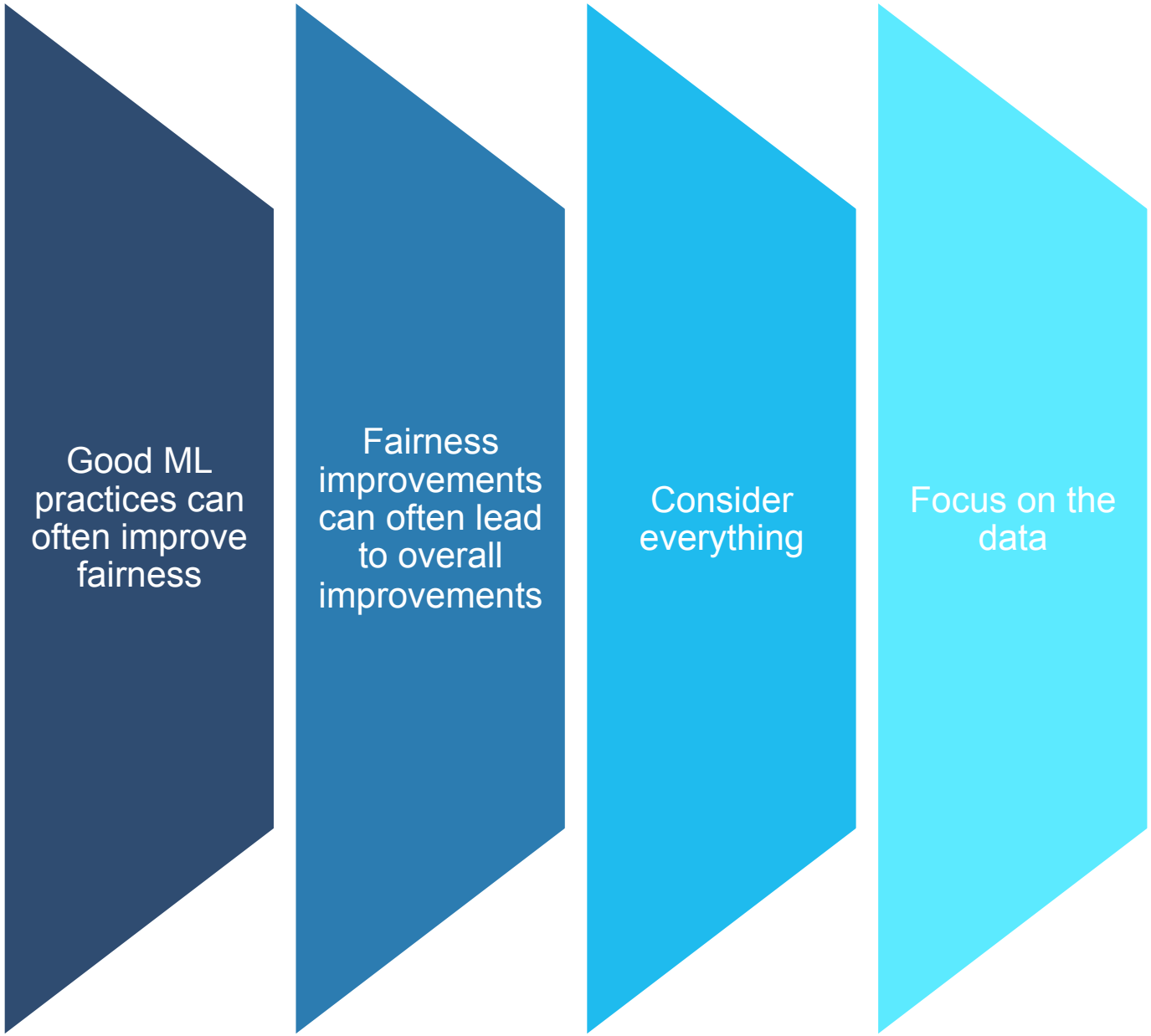

GPU


FPGA

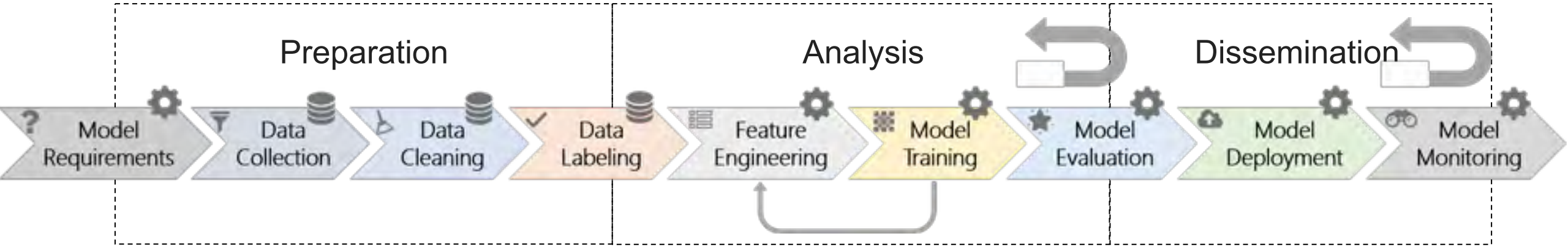
Fairness

AI systems should **treat everyone fairly** and avoid affecting similarly situated groups of people in

Fairness in Practice



Measurement Fairness Tooling



Data Integrity	Label Quality	Model Integrity	Model Effect	Monitoring
• Representiveness • Quality • Intersectional properties	• Labelers • Coverage	• Representive Power of Features • Correlations • Interpretability	• Measure different fairness metrics • Calibration • Demographic Parody • False Positive Rate • False Negative Rate	• Track key metrics

FAIRLEARN

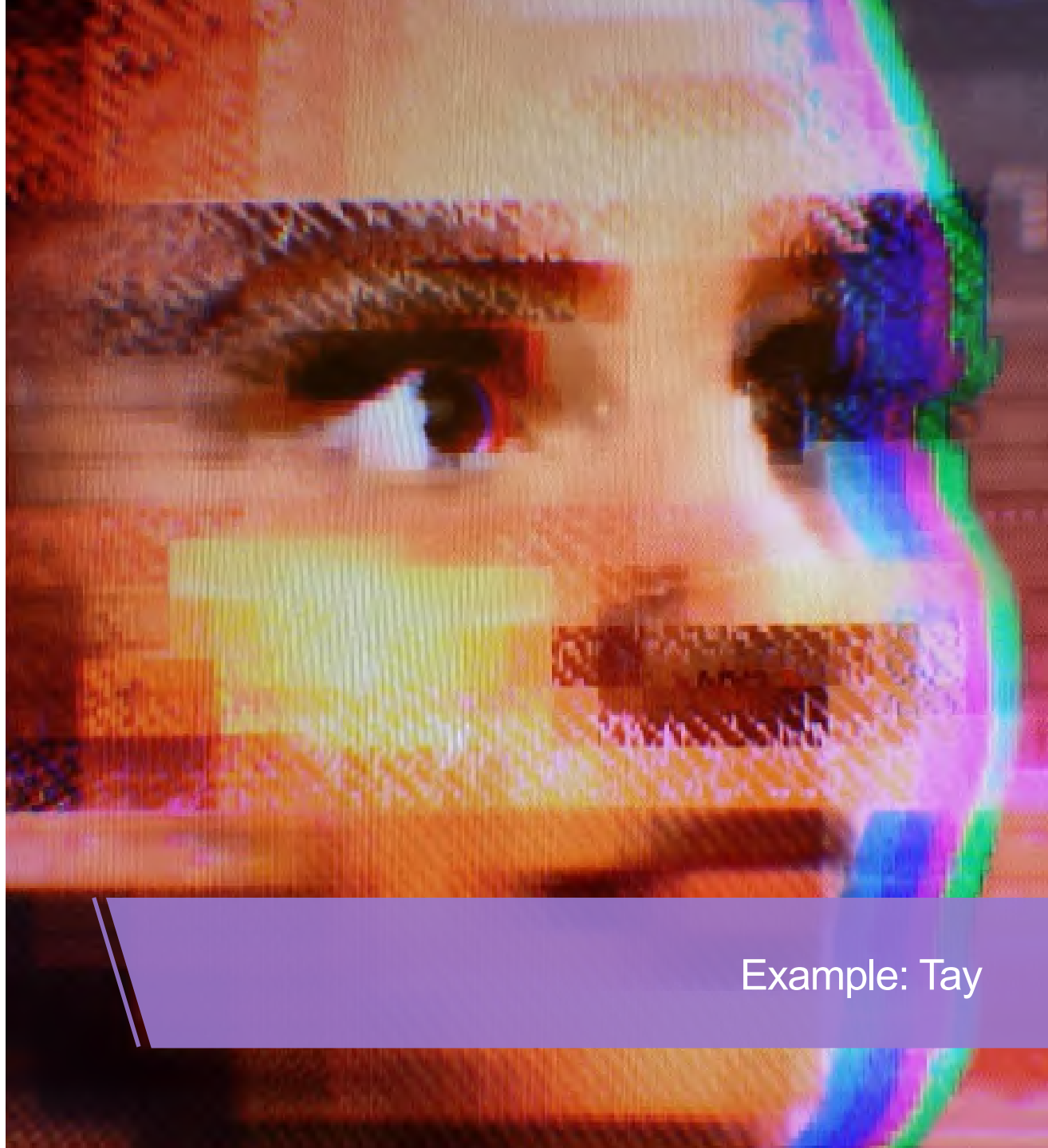
github.com/microsoft/fairlearn

Reliability and Safety

AI systems should **operate reliably, safely, and consistently** under normal circumstances and in unexpected conditions

Privacy and Security

Like other technologies, AI systems should be able to **protect private information and resist attacks**



Promising Technologies

- Secure execution environments
- Homomorphic encryption
- Multi-party computation
- Differential privacy

Differential Privacy

Injecting Noise with Formal Privacy Rules to avoid disclosing data

Advantages:

- Privacy guarantees are closed under composition
- Privacy guarantees are robust to post-processing
- Privacy guarantees are future-proof
- Privacy guarantees are provable and tunable
- Privacy guarantees are public and explainable
- Protects against database reconstruction attacks

Disadvantages:

- Every use of the private data must be tallied in the privacy-loss budget

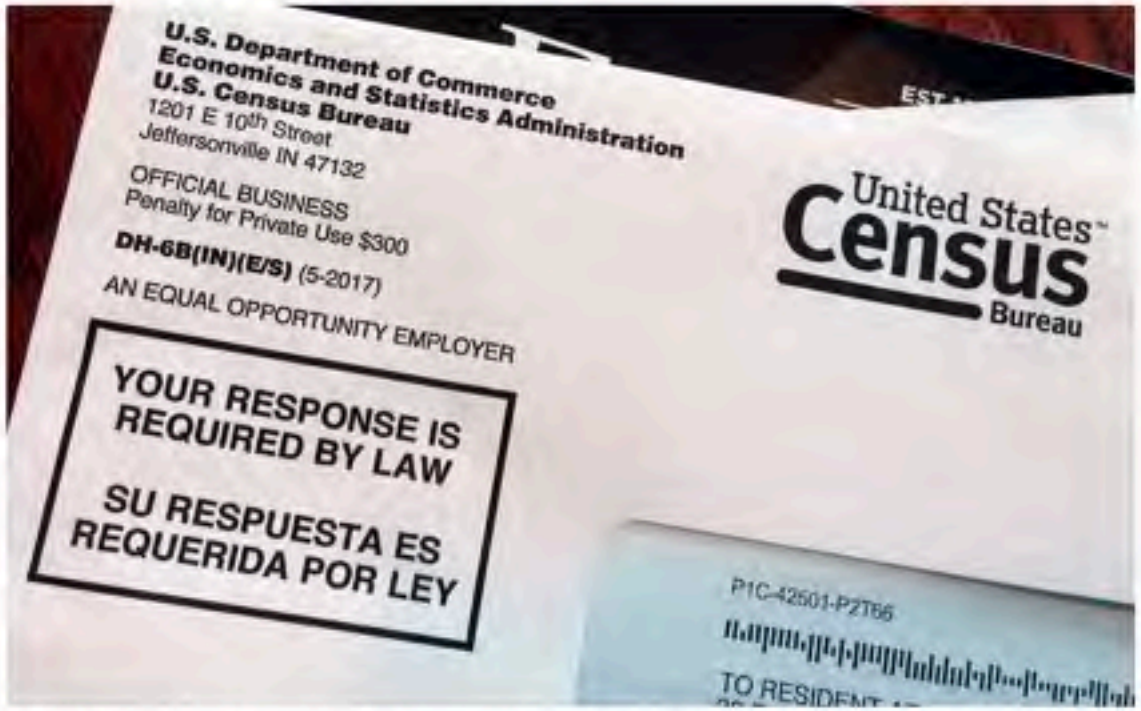
[John M. Abowd]

The New York Times

TheUpshot

To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

Guaranteeing people's confidentiality has become more of a challenge, but some scholars worry that the new system will impede research.



A 2018 census test letter mailed to a resident in Providence, R.I. The nation's test run of the 2020 Census is in Rhode Island. Michelle R. Smith/Associated Press

Inclusiveness

AI systems should **empower everyone and engage people**

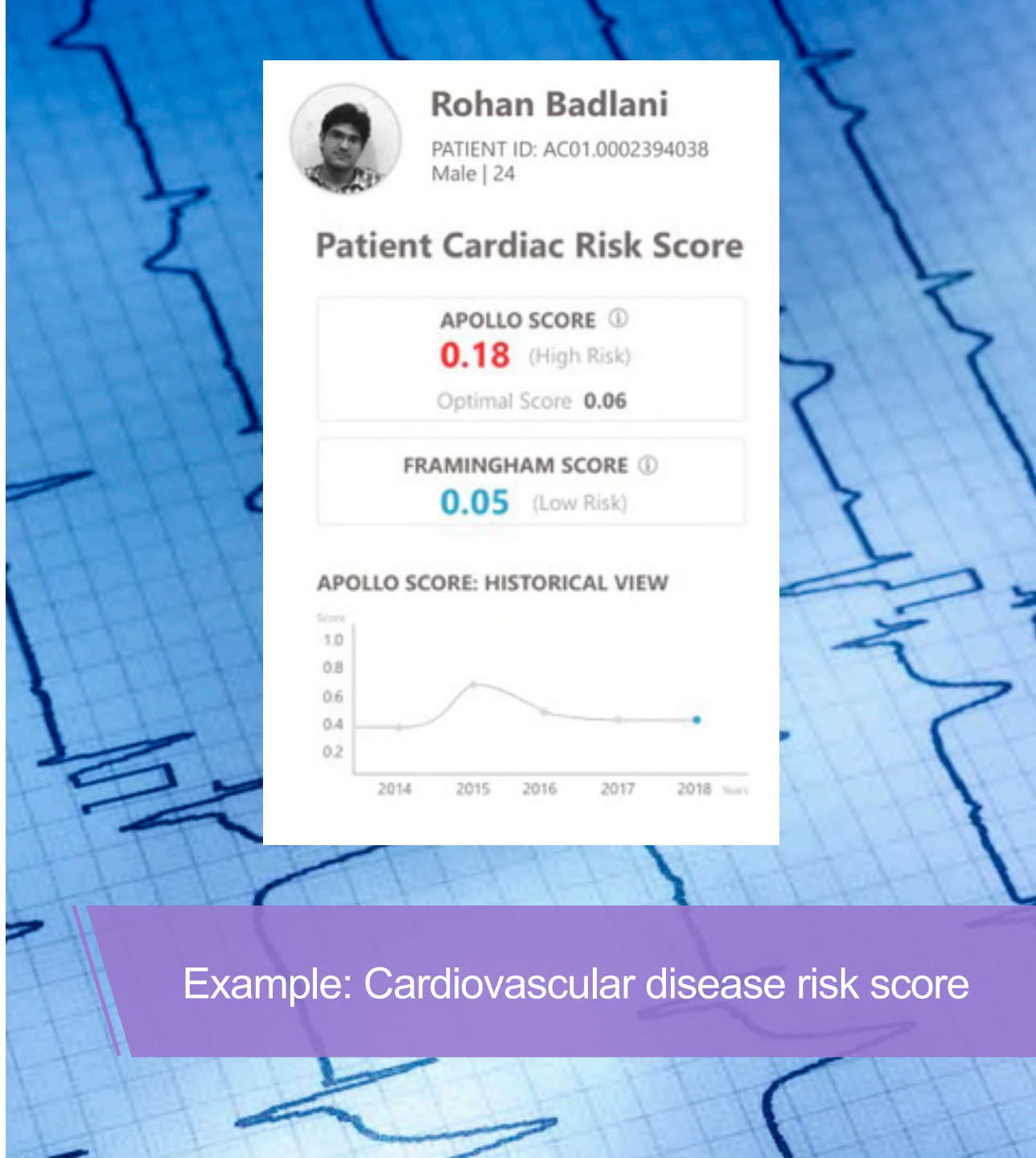
Inclusive Design

Inclusive Design is a methodology, born out of digital environments, that enables and draws on the full range of human diversity. Most importantly, this means including and learning from people with a range of perspectives.

https://www.microsoft.com/design/inclusive/

Transparency

People should be able to **understand how AI systems make decisions,** especially when those decisions impact people's lives



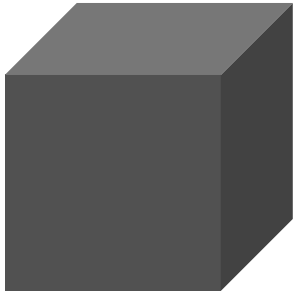
Example: Cardiovascular disease risk score

Interpretability

Tools to **understand how the system is working**

Glassbox Models

Explainable Boosting
Linear Models
Decision Tree
Rule Systems
...

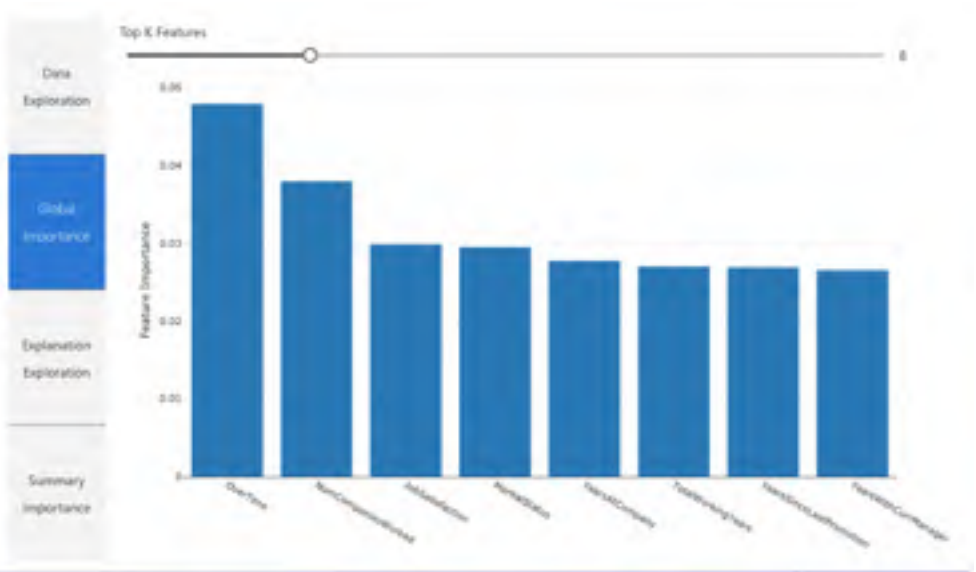


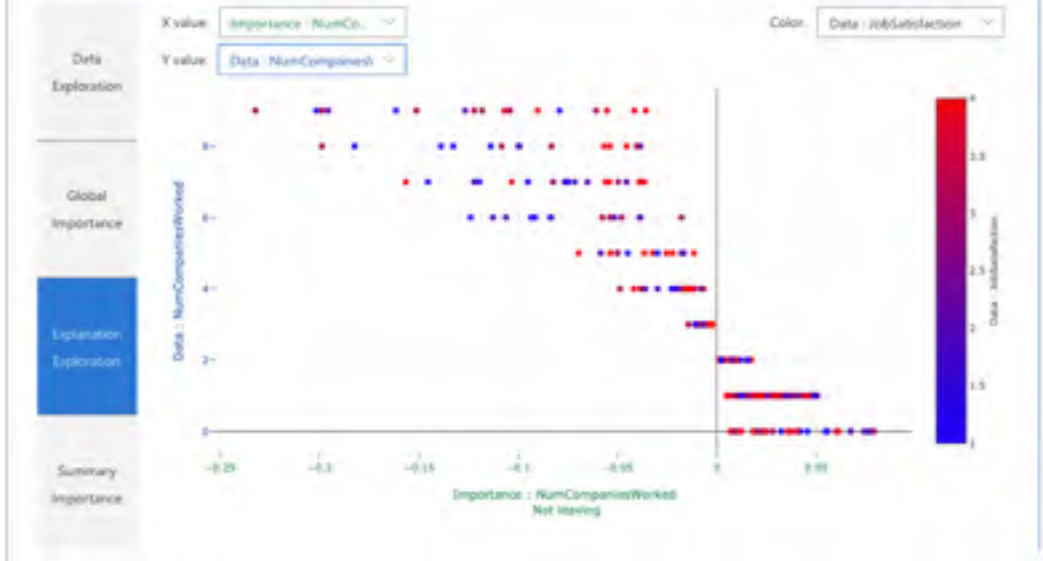
Blackbox Explainers

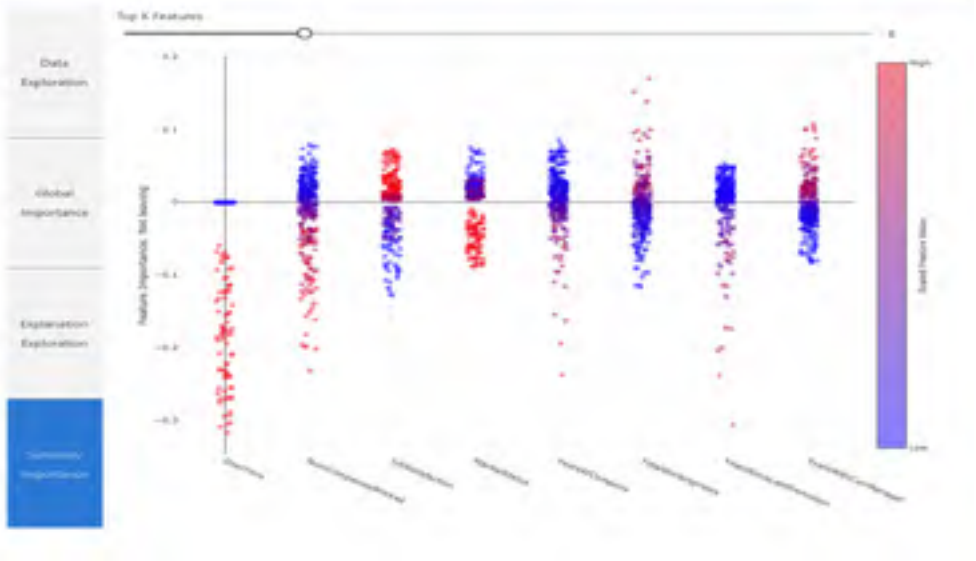
LIME
SHAP
Partial Dependence
Sensitivity Analysis
...

INTERPRETML

github.com/microsoft/interpret

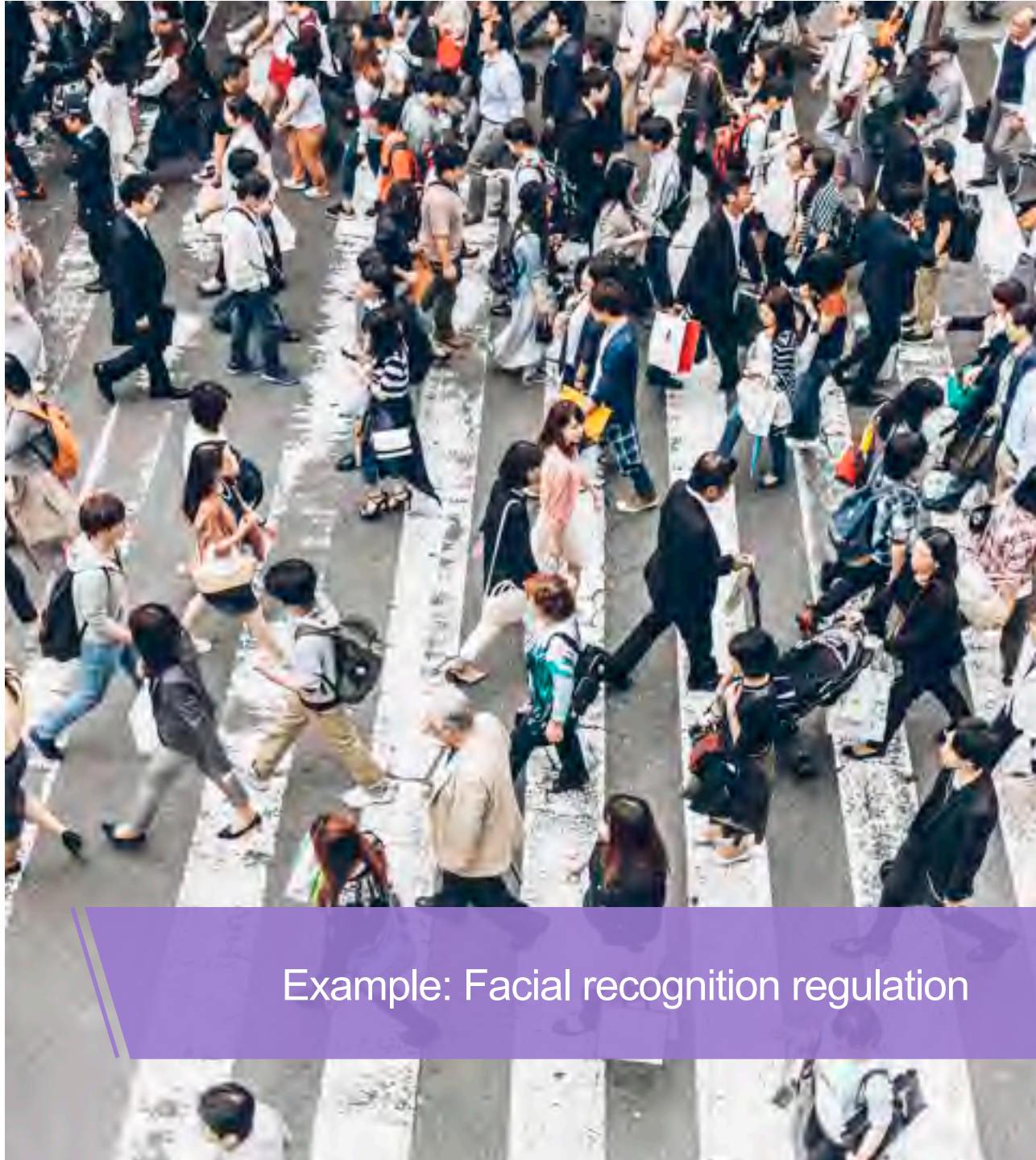






Accountability

People must **maintain responsibility for and meaningful control** over AI systems



Responsible Meta-Data

arXiv:1803.09010v4 [cs.DB] 14 Apr 2019

Datasheets for Datasets

Timnit Gebru*[1], Jamie Morgenstern[2], Briana Vecchione[3], Jennifer Wortman Vaughan[4], Hanna Wallach[4], Hal Daumé III[4,5], and Kate Crawford[4,6]

[1]Google
[2]Georgia Institute of Technology
[3]Cornell University
[4]Microsoft Research
[5]University of Maryland
[6]AI Now Institute

April 16, 2019

Abstract

The machine learning community currently has no standardized process for documenting datasets. To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using datasets, and the characteristics of these datasets will fundamentally influence a model's behavior. A model is unlikely to perform well in the wild if its deployment context doesn't match its training or evaluation datasets, or if these datasets reflect unwanted biases. Mismatches like this can have especially severe consequences when machine learning is used in high-stakes domains such as criminal justice [2, 20, 44], hiring [29], critical infrastructure [10, 35], or finance [28]. And even in other domains, mismatches may lead to loss of revenue or public relations setbacks.

Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in datasets. Much like a faulty capacitor in a circuit,

arXiv:1810.03993v2 [cs.LG] 14 Jan 2019

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

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [15]) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for two supervised models: One trained to detect smiling faces in images, and one trained to detect toxic comments in text. We propose model cards as a step towards the responsible democratization of machine learning and related artificial intelligence technology, increasing transparency into how well artificial intelligence technology works. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed evaluation numbers and other relevant documentation.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Social and professional topics** → *User characteristics*; • **Software and its engineering** → **Use cases**; *Documentation*; *Software evaluation*; • **Human-centered computing** → *Walkthrough evaluations*;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

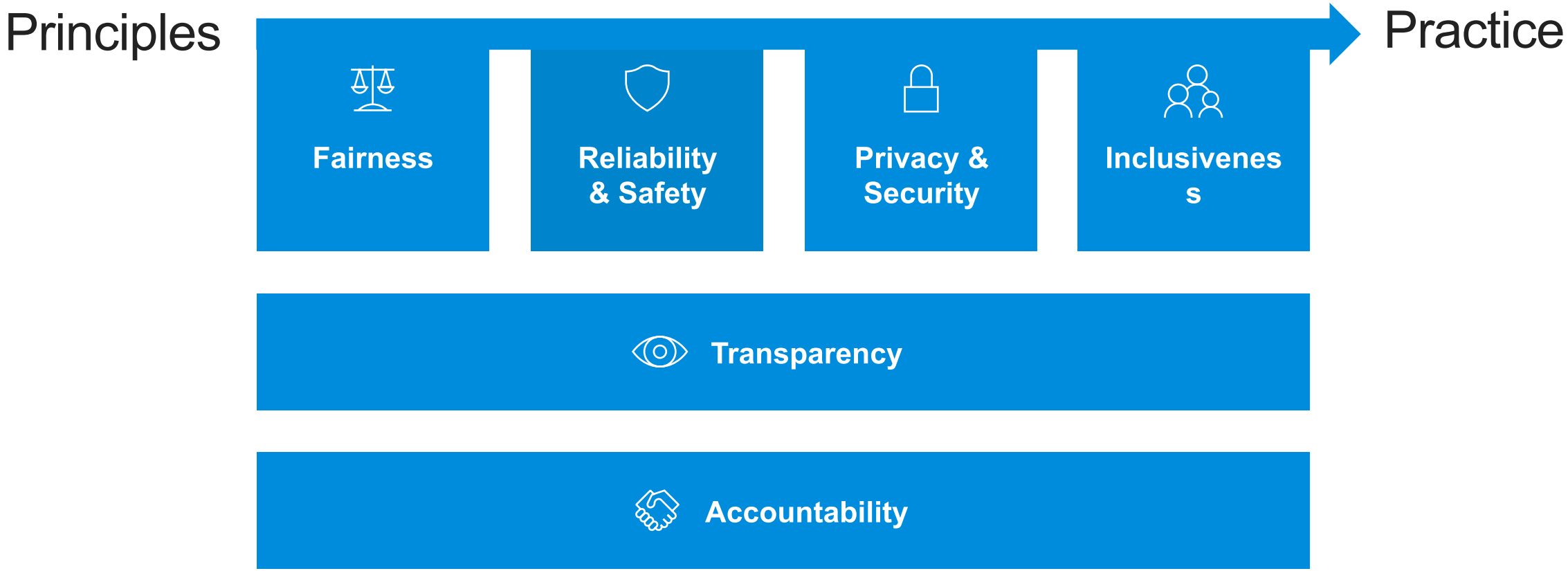
1 INTRODUCTION

Currently, there are no standardized documentation procedures to communicate the performance characteristics of trained machine learning (ML) and artificial intelligence (AI) models. This lack of documentation is especially problematic when models are used in applications that have serious impacts on people's lives, such as in health care [13, 42, 44], employment [1, 13, 29], education [23, 45], and law enforcement [2, 7, 20, 34].

Researchers have discovered systematic biases in commercial machine learning models used for face detection and ranking [4, 9, 40], attribute detection [5], criminal justice [10], toxic comment detection [11], and other applications. However, these systematic errors were only exposed after models were put into use, and negatively affected users reported their experiences. For example, after MIT Media Lab graduate student Joy Buolamwini found that commercial face recognition systems failed to detect her face [4], she collaborated with other researchers to demonstrate the disproportionate errors of computer vision systems on historically marginalized groups in the United States, such as darker-skinned women [5, 41]. In spite of the potential negative effects of such reported biases, documentation accompanying trained machine learning models (if supplied) provide very little information regarding model performance characteristics, intended use cases, potential pitfalls, or other information to help users evaluate the suitability of these systems to their context. This highlights the need to have detailed documentation accompanying trained machine learning models, including metrics that capture bias, fairness and inclusion considerations.

As a step towards this goal, we propose that released machine learning models be accompanied by short (one to two page) records we call model cards. Model cards (for model reporting) are complements to "Datasheets for Datasets" [21] and similar recently proposed documentation procedures [3, 28] that report details of

Progression of Responsible AI



How do we proceed?

Focus on awareness and education

Build tools for measurement and mitigation

Develop application and industry specific-guidance

How will you impact AI?