AI Risk Management Framework

From National Institute of Standards and Technology (NIST), U.S. Department of Commerce

Version 1.0 released on Jan, 2023

https://www.nist.gov/itl/ai-risk-management-framework

Content

- Part I Motivation
  - Trustworthy and Responsible AI
  - Sources of AI Risks
  - Who should be involved?
  - Understanding Risk, Impact, and Harms
  - AI Risk and trustworthy
- Part 2 Framework Core

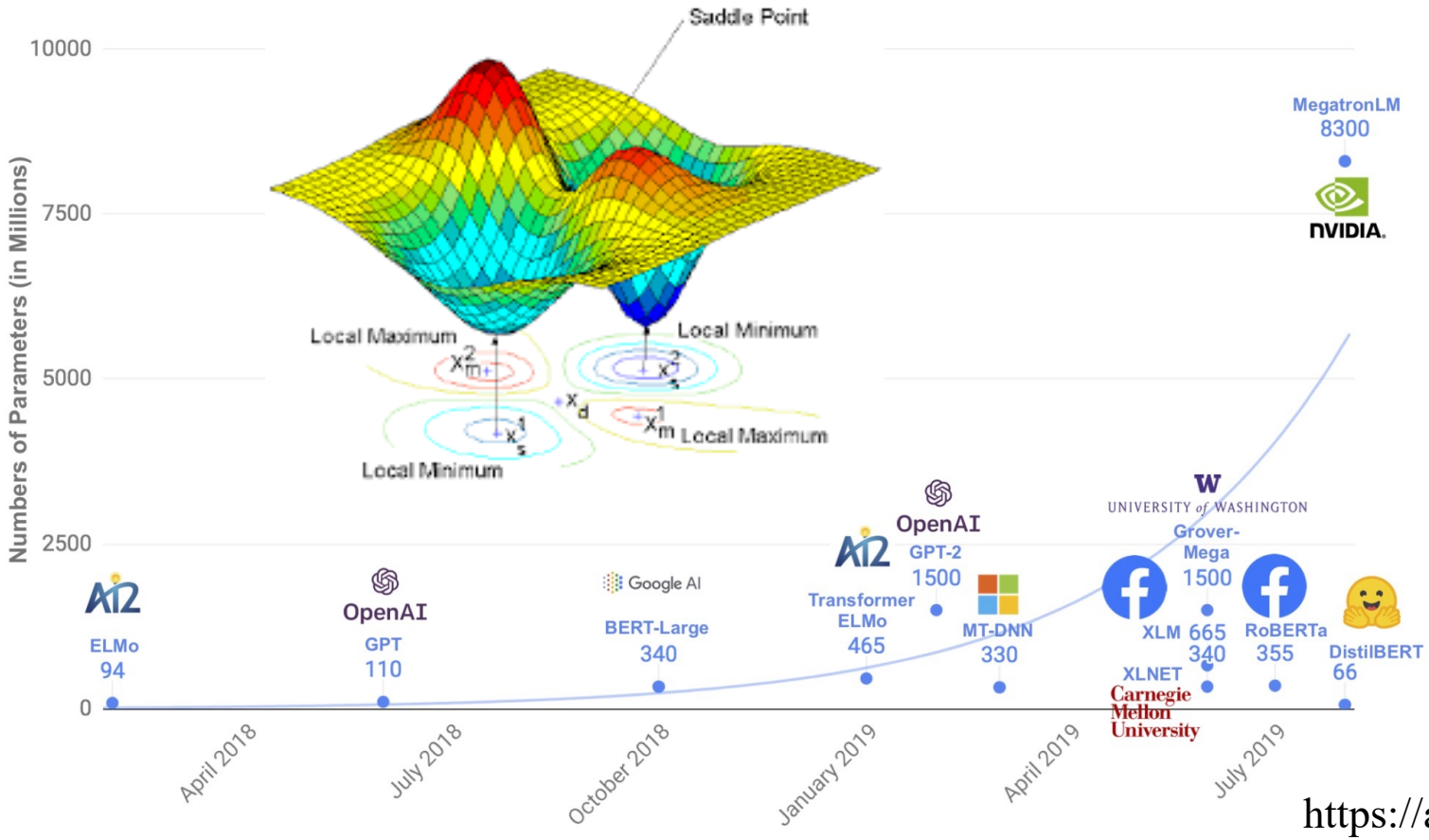
Motivation

- Managing AI risk towards **Trustworthy** and **Responsible AI**
  - **Trustworthy AI** is valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced
  - **Responsible use and practice of AI systems** is a counterpart to AI system trustworthiness

- Risks to any software or information-based system apply to AI
  - including concerns related to cybersecurity, privacy, safety, and infrastructure

Motivation

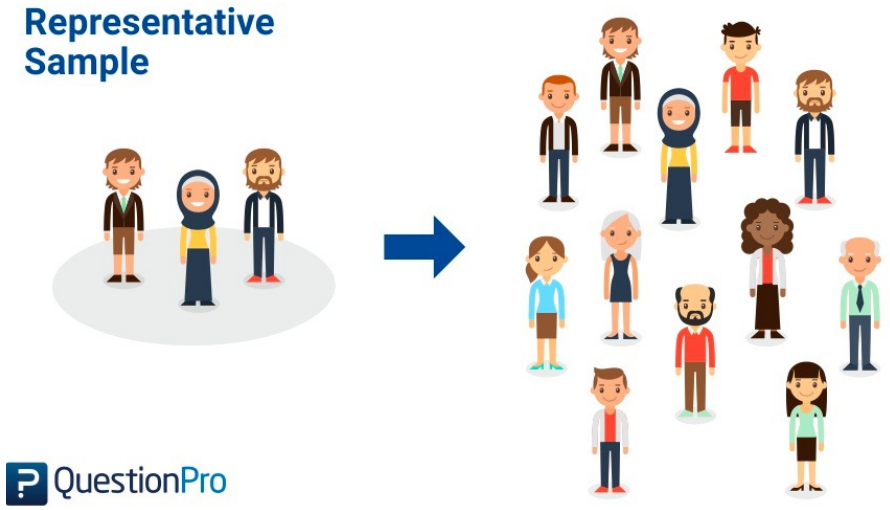
- New challenges:
  - A useful mathematical representation of the data interactions that drive the AI system's behavior is **not fully known**



https://aman.ai/papers/

Motivation

- Sources of Risks including
  - **<u>Data used to train the AI system</u>**
    - Data quality: inaccurate
    - Data not appropriate representation of the context

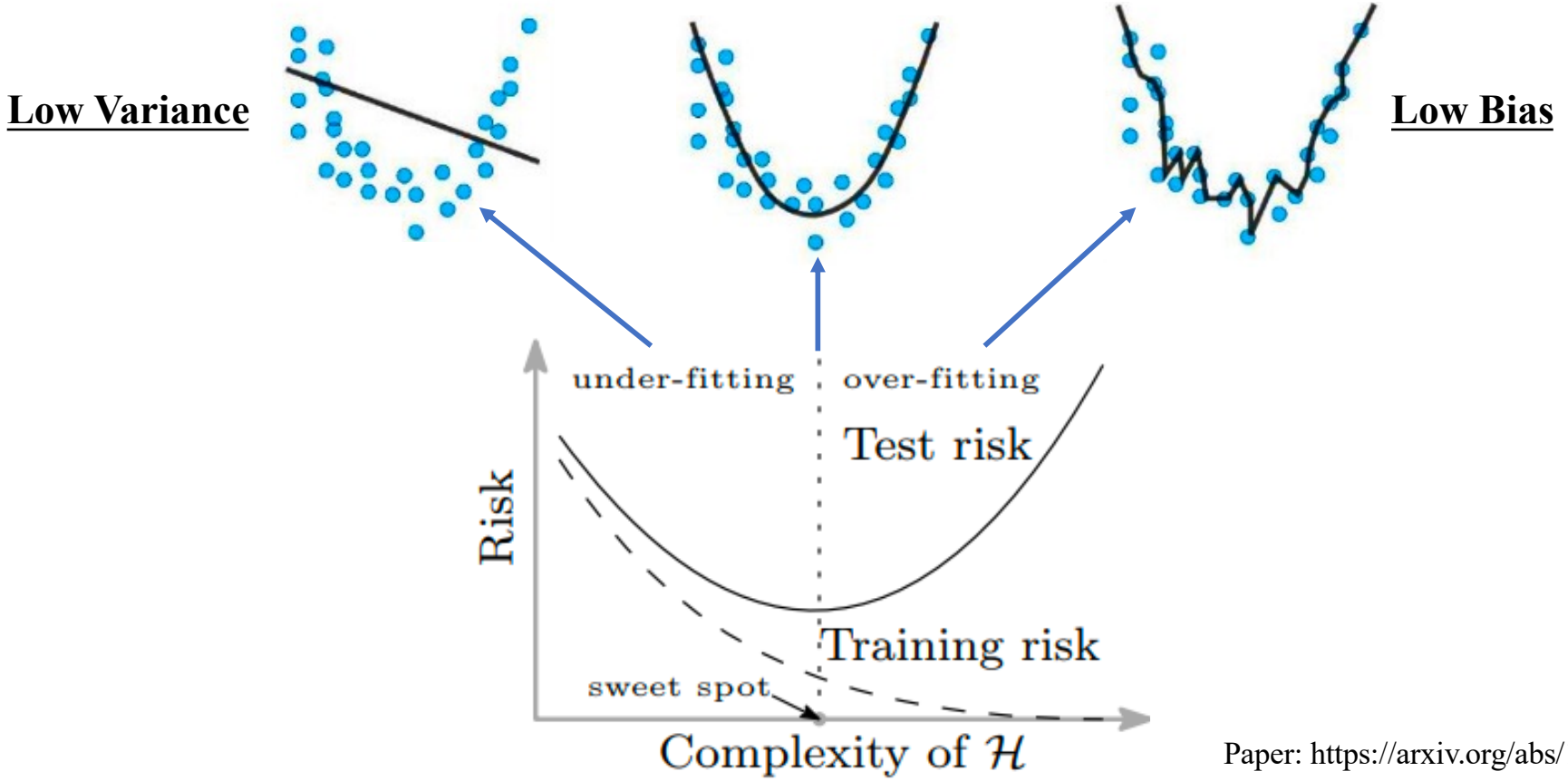


https://www.questionpro.com/blog/representative-sample/

L. Yang et al, doi: 10.1109/LRA.2020.2969932

Motivation

- Sources of Risks including
  - **AI system itself**
    - Overfitting/Underfitting, Instability

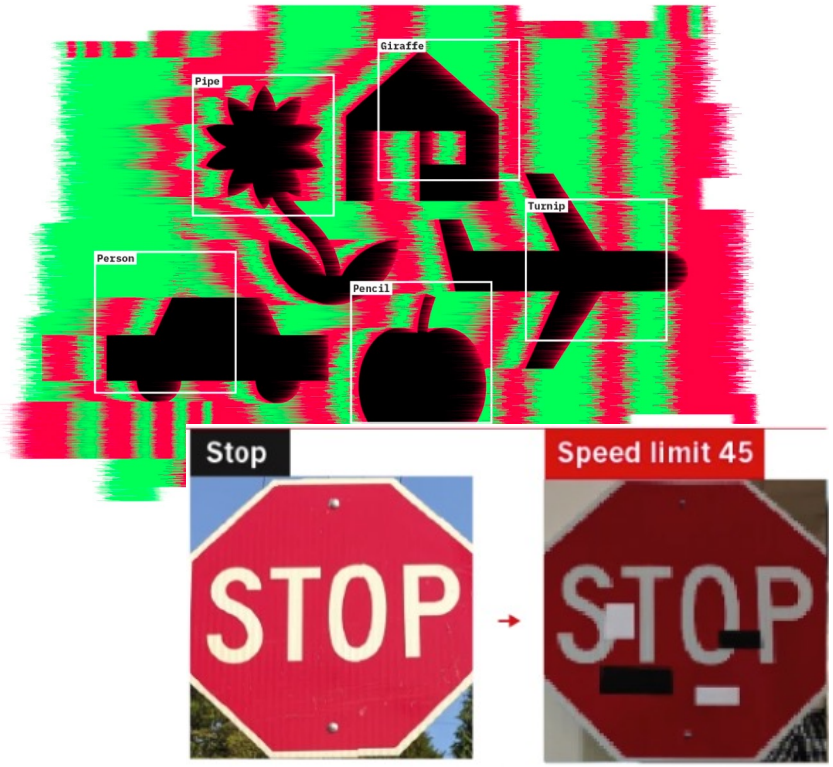


Paper: https://arxiv.org/abs/1812.11118

Motivation

- Sources of Risks including
  - **Use of the AI system, or interaction of people with the AI system**

Inappropriate use of narrow AI



https://www.nature.com/articles/d41586-019-03013-5

Unintended, Malicious uses

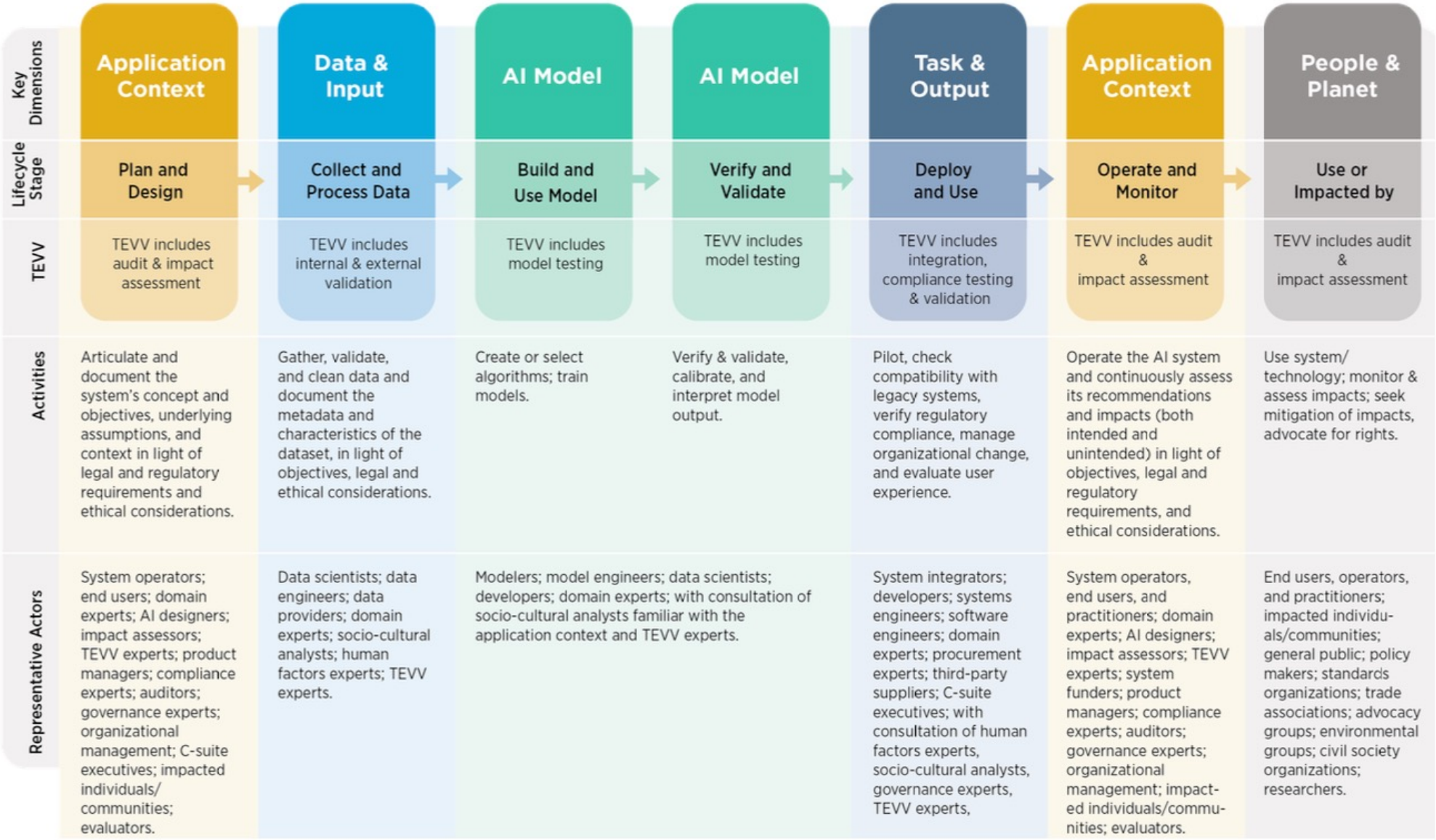
Who should be involved?

AI actors *who play an active role in the **AI system lifecycle**, including organizations and individuals that deploy or operate AI*

TEVV: test, evaluation, verification, and validation

- The two inner circles show AI systems' key dimensions and the outer circle shows AI lifecycle stages.
- Ideally, risk management efforts start with the Plan and Design function in the application context and are performed throughout the AI system lifecycle.

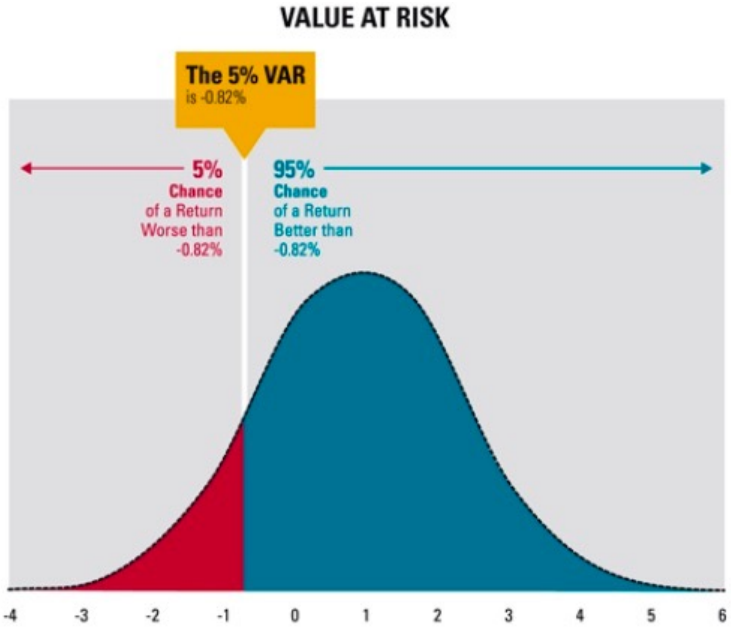
Who should be involved?



Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts,	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impact-ed individuals/commu-nities; evaluators.	End users, operators, and practitioners; impacted individu-als/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

Understanding Risk, Impacts, and Harms

- Risk refers to the composite measure of
  - an event's probability of occurring
  - the magnitude of the consequences

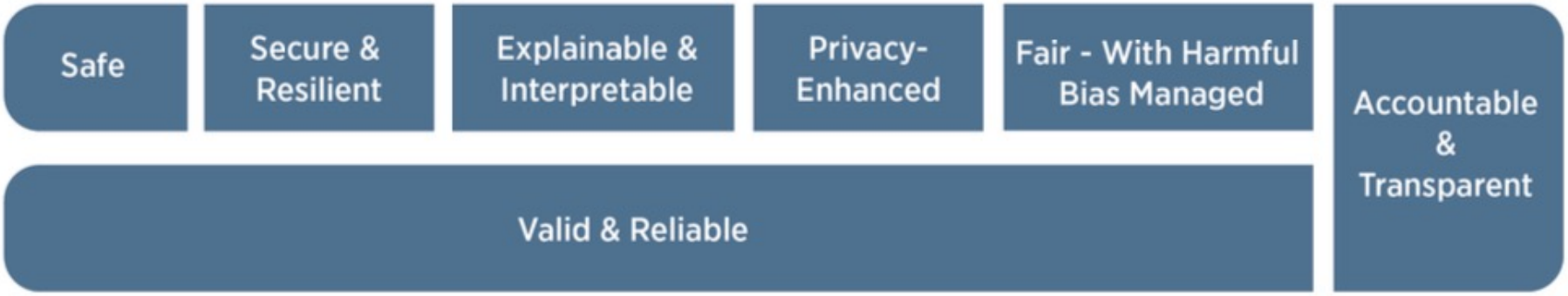


- Risk management processes address negative impacts

- This framework offers approaches to
  - minimize anticipated negative impacts
  - identify opportunities to maximize positive impacts

Examples of potential harms related to AI systems

AI Risk and Trustworthiness

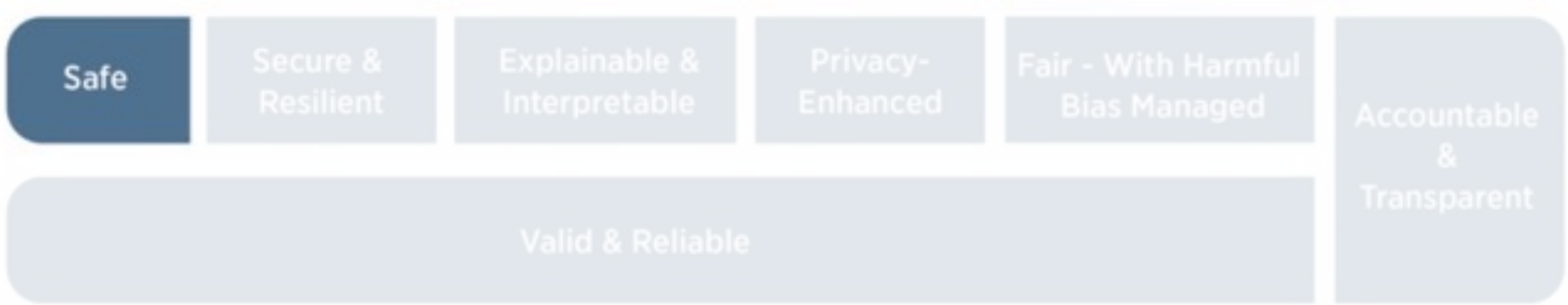


- These characteristics are tie to
  - Social and organizational behavior
  - Datasets used by AI systems
  - Selection of AI models and algorithms and the decisions made by those who build them
  - the interactions with the humans who provide insight from and oversight of such systems
- Trustworthiness characteristics are **interrelated**
  - Tradeoffs are usually involved
  - Highly secure but unfair, accurate but opaque and uninterpretable, and inaccurate but secure, privacy-enhanced, and transparent

Definition of AI trustworthy characteristics

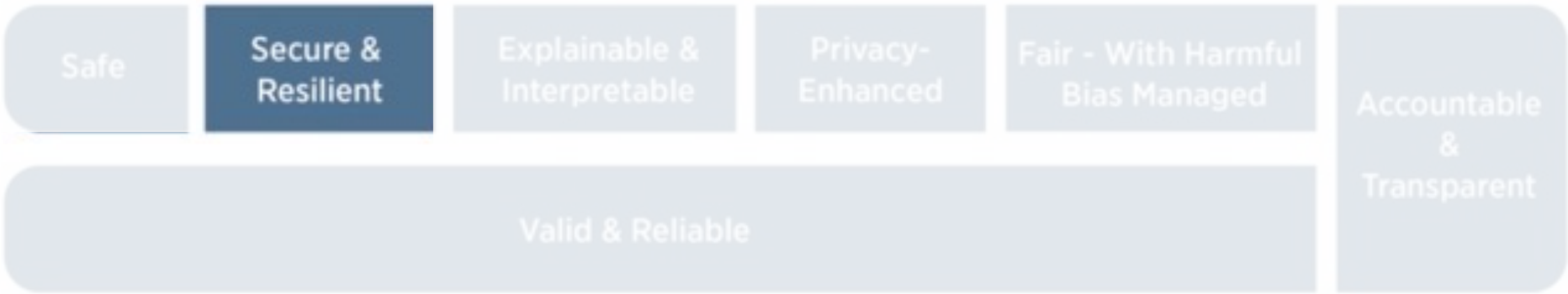
Characteristics	Definition
Validation	Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled
Reliability	Ability of an item to perform as required, without failure, for a given time interval, under given conditions
Accuracy	Closeness of results of observations, computations, or estimates to the true values or the values accepted as being true
Robustness	Ability of an AI system to maintain its level of performance under a variety of circumstances

Definition of AI trustworthy characteristics



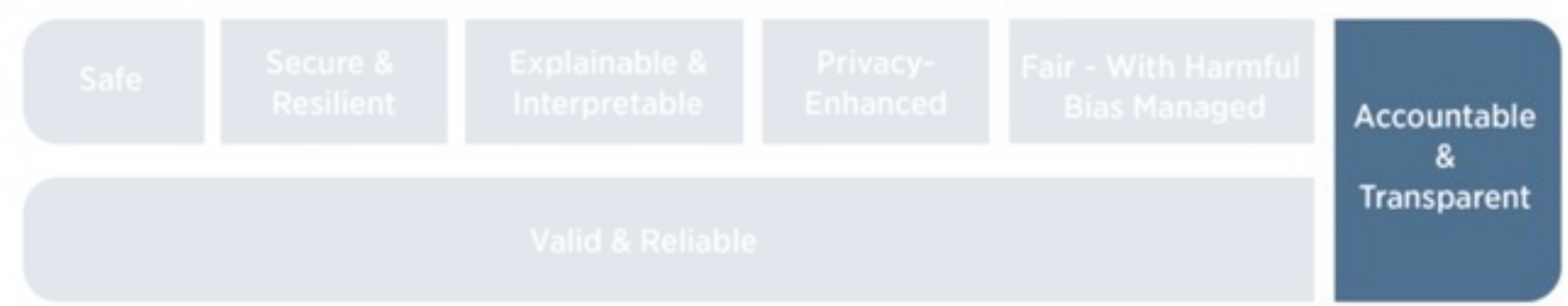
Characteristics	Definition
Safety	AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered

Definition of AI trustworthy characteristics



Characteristics	Definition
Security	Maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use
Resilient	Withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary

Definition of AI trustworthy characteristics



Characteristics	Definition
Accountability & transparency	Accountability presupposes transparency. Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so.

Definition of AI trustworthy characteristics

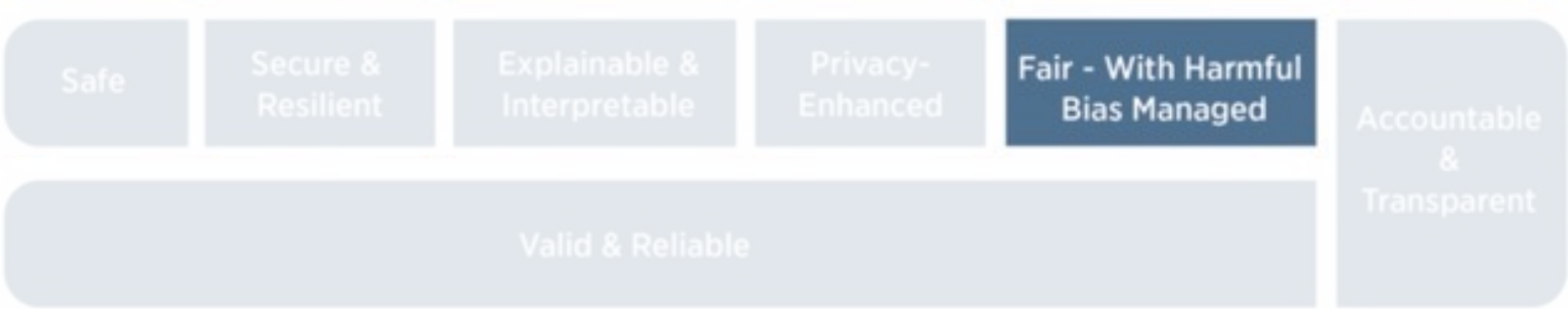
Characteristics	Definition
Explainability	a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes

Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of "what happened" in the system. Explainability can answer the question of "how" a decision was made in the system. Interpretability can answer the question of "why" a decision was made by the system and its meaning or context to the user.

Definition of AI trustworthy characteristics

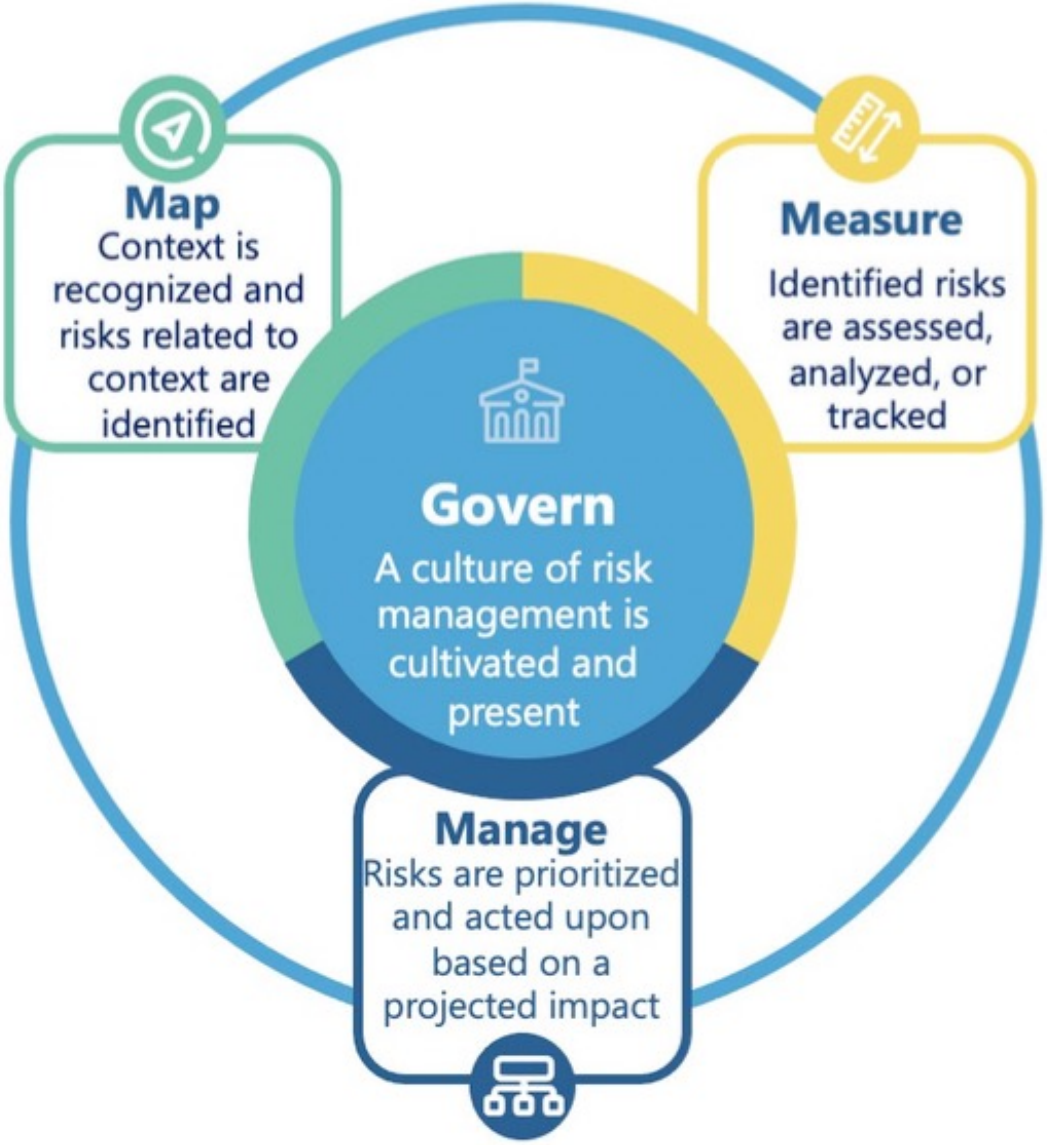
Characteristics	Definition
Privacy	refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity

Definition of AI trustworthy characteristics



Characteristics	Definition
Fairness	concerns for equality and equity but can be complex and difficult to define

AI RMF Core


Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact

DS323: AI in Design

MAP-1: Context is estabished and understood. +

MAP-2: Classification of the AI system is performed. −

🔗 MAP 2.1: The specific task, and methods used to implement the task, that the AI system will support is defined (e.g., classifiers, generative models, recommenders, etc.).

🔗 MAP 2.2: Information is documented about the system's knowledge limits, and how output will be utilized and overseen by humans.

🔗 MAP 2.3: Scientific integrity and TEVV considerations are identified and documented including related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), and construct validation.

MAP-3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with the status quo are understood. +

MAP-4: Risks and benefits are mapped for third-party software and data. +

MAP-5: Impacts to individuals, groups, communities, organizational, or society are assessed. +

Source

Exercise: What if tool from Google

"Diagnostic tool lets users try on five different types of fairness."

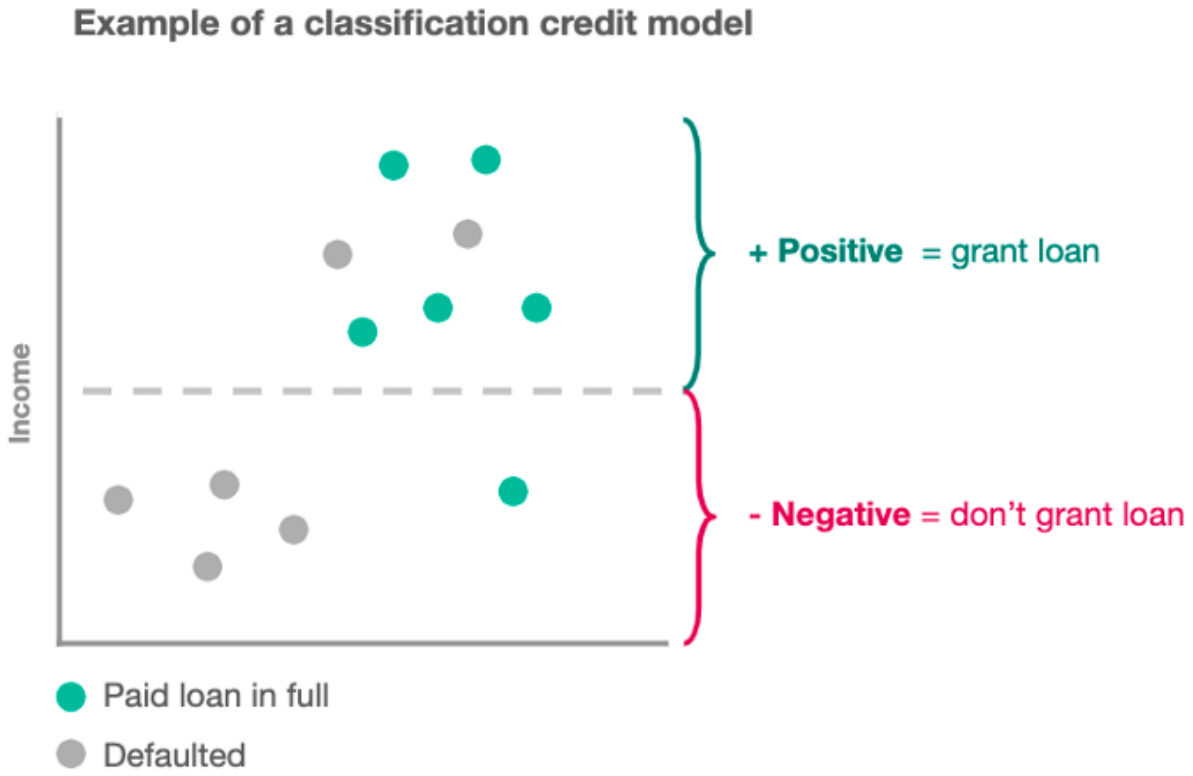
Web Demos: https://pair-code.github.io/what-if-tool/explore/#web

Types of fairness	Description
Group unaware	Disregard the different slices/groups
Group thresholds	Optimize a separate threshold for each slice based on the specified cost ratio.
Demographic parity	Similar percentages of datapoints from each slice are predicted as positive classifications.
Equal opportunity	Among those datapoints with the positive ground truth label, there is a similar percentage of positive predictions in each slice.
Equal accuracy	There is a similar percentage of correct predictions in each slice.

https://pair-code.github.io/what-if-tool/ai-fairness.html

Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy



Set a single income threshold in our training set to decide who gets a loan in the future (the dotted line)

Source

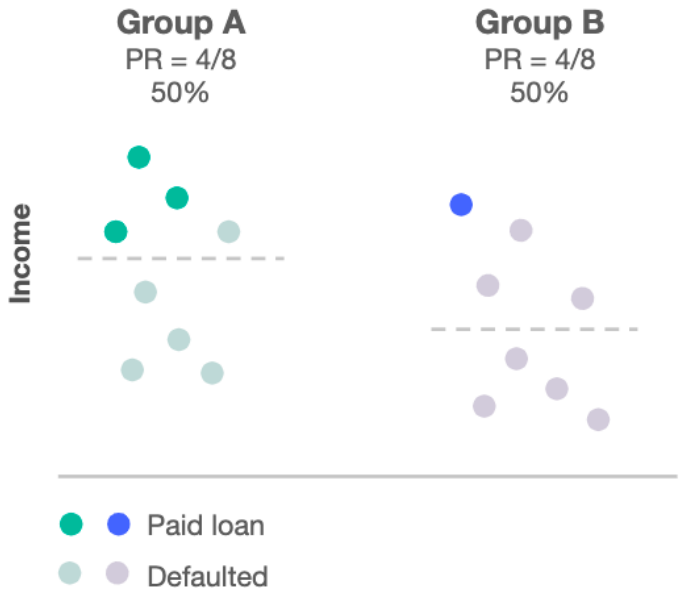
Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy

Source

Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy



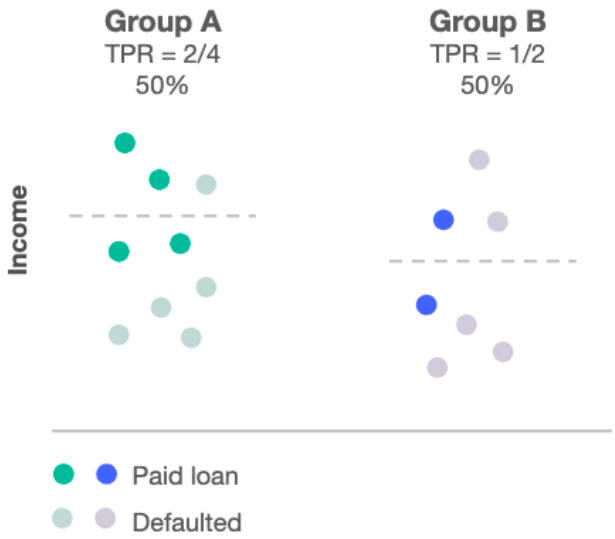
When to use Demographic Parity

- We want to change the state of our current world to improve it (e.g.: we want to see more minority groups getting to the top)
- We are aware of historical biases may have affected the quality of our data (e.g.: ML solution trained to hire software engineers, where nearly no women was hired before)
- We have a plan in place to support the unprivileged group

Source

Exercise: What if tool from Google

Types of fairness
Group unaware
Group thresholds
Demographic parity
Equal opportunity
Equal accuracy



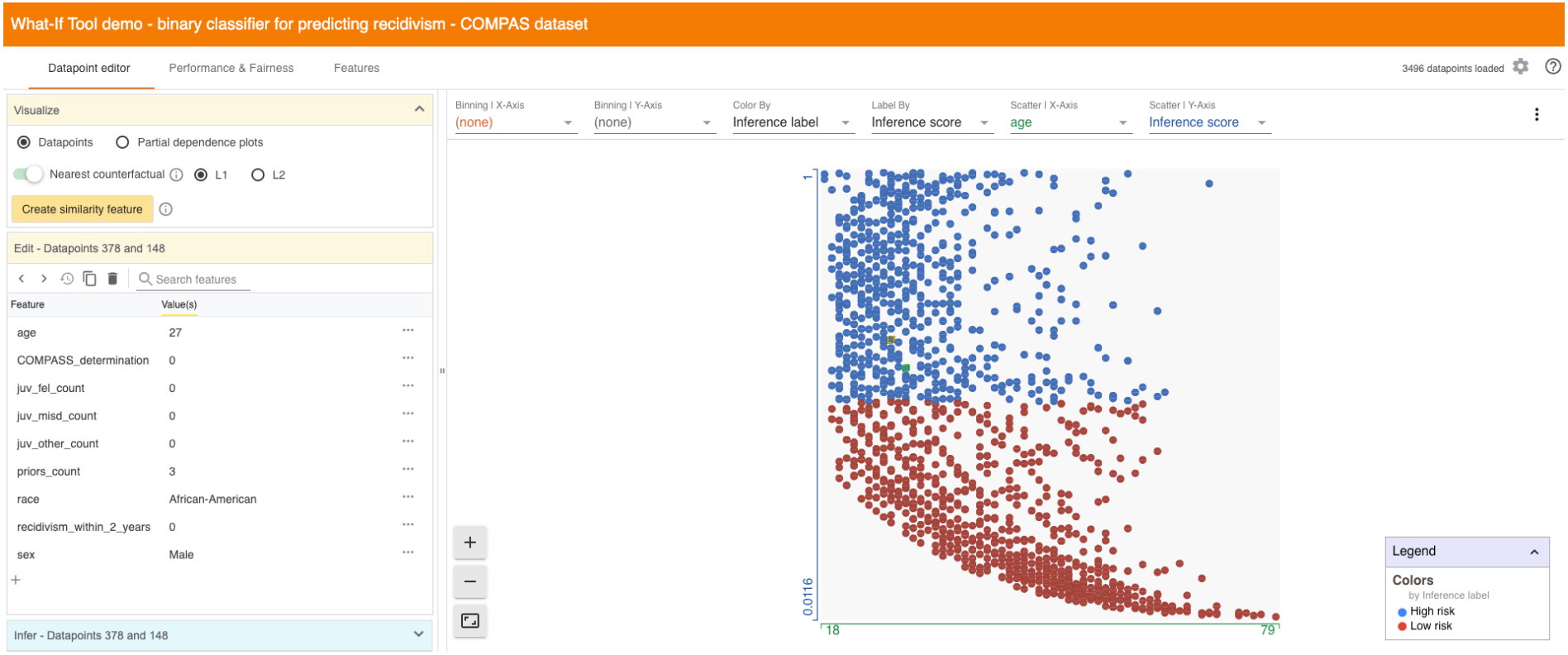
When to use Demographic Parity

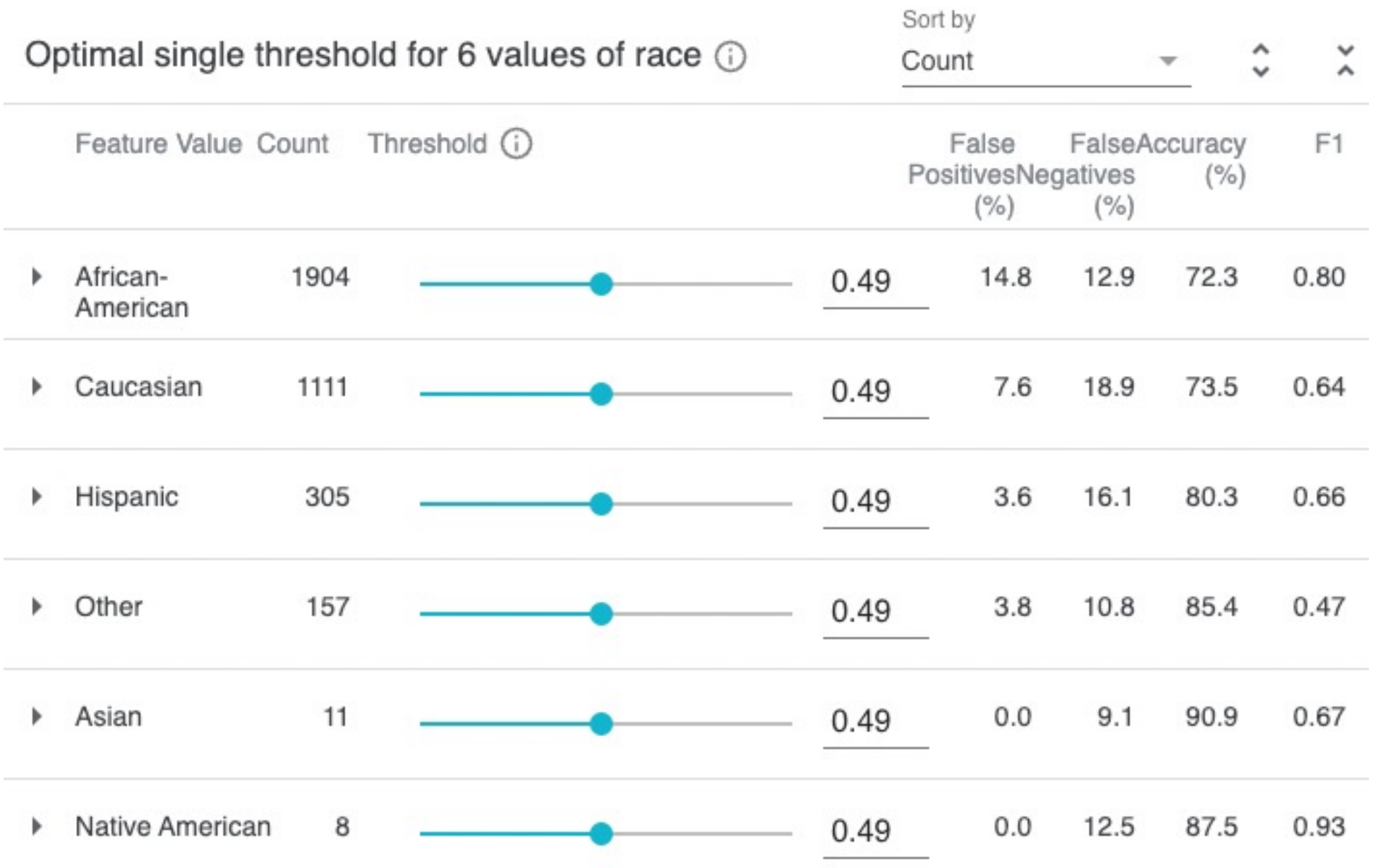
- There is a strong emphasis on predicting the positive outcome correctly (e.g.: we need to be very good at detecting a fraudulent transaction)
- Introducing False Positives are not costly to the user nor the company (e.g.: wrongly notifying a customer about fraudulent activity will not be necessarily expensive to the customer nor the bank sending the alert)

Source

Investigate fairness on recidivism classification:

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending (recidivism).
- Website: https://pair-code.github.io/what-if-tool/demos/compas.html





Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
African-American	1904	0.49	14.8	12.9	72.3	0.80
Caucasian	1111	0.49	7.6	18.9	73.5	0.64
Hispanic	305	0.49	3.6	16.1	80.3	0.66
Other	157	0.49	3.8	10.8	85.4	0.47
Asian	11	0.49	0.0	9.1	90.9	0.67
Native American	8	0.49	0.0	12.5	87.5	0.93