Ethical AI for Public Participation and Transparency

Open Cities Summit 2018

—

Milena Pribic
AI Design Practices at IBM

@milenapribic
https://www.linkedin.com/in/milenapribic/


50 MILLION

02.13.18

Facial Recognition Systems Are Even More Biased Than We Thought

AI from Microsoft, IBM, and Face ++ is much less accurate when detecting dark-skinned female faces than light-skinned male faces.

The Switch

law

Intelligent Machines

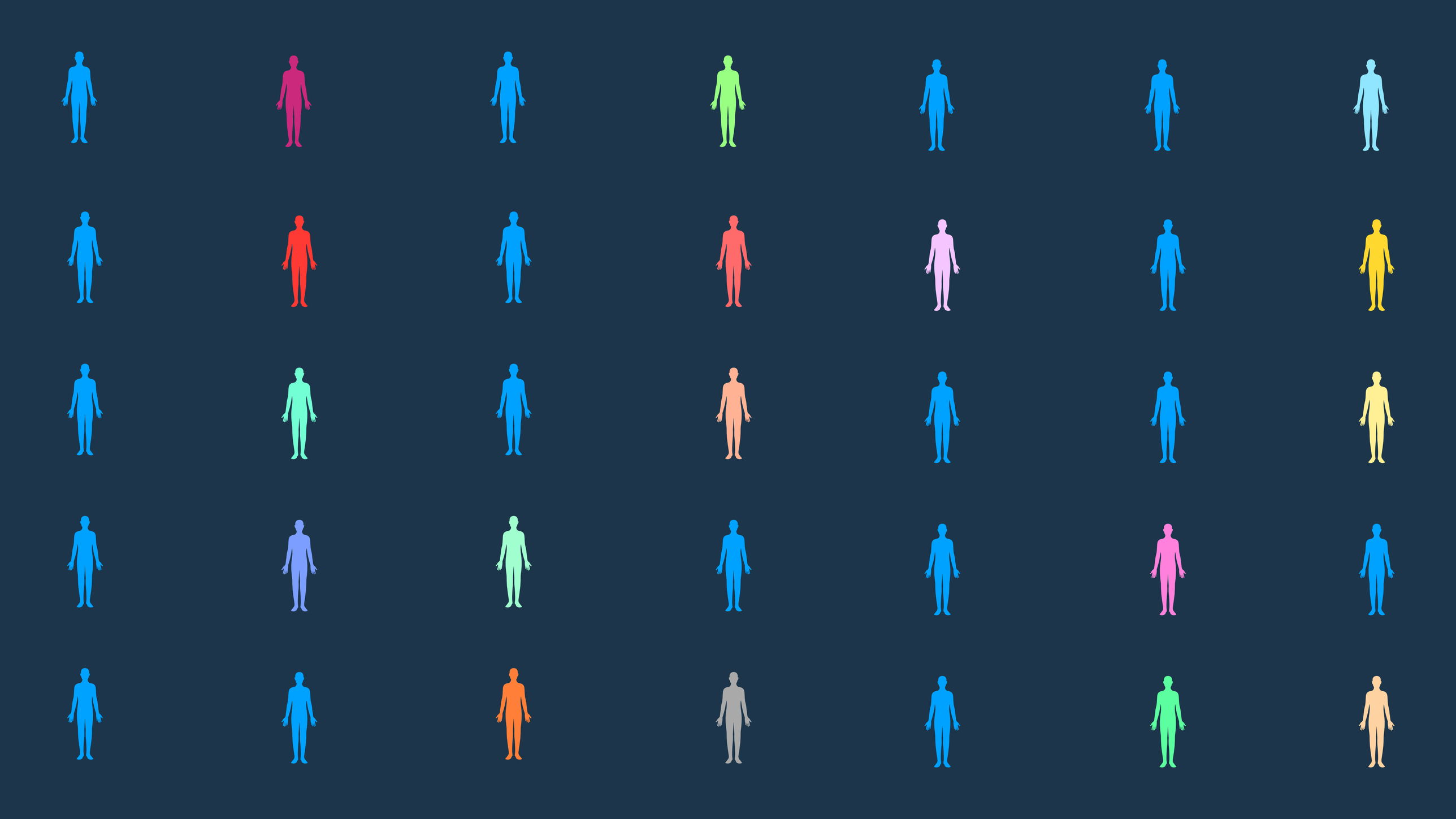
Forget Killer Robots—Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.


11832
11811
11816
11840
11862
11876
11863
11796
11808
11858
11845
11872
200 MILLION
11829
11786
11765
11750

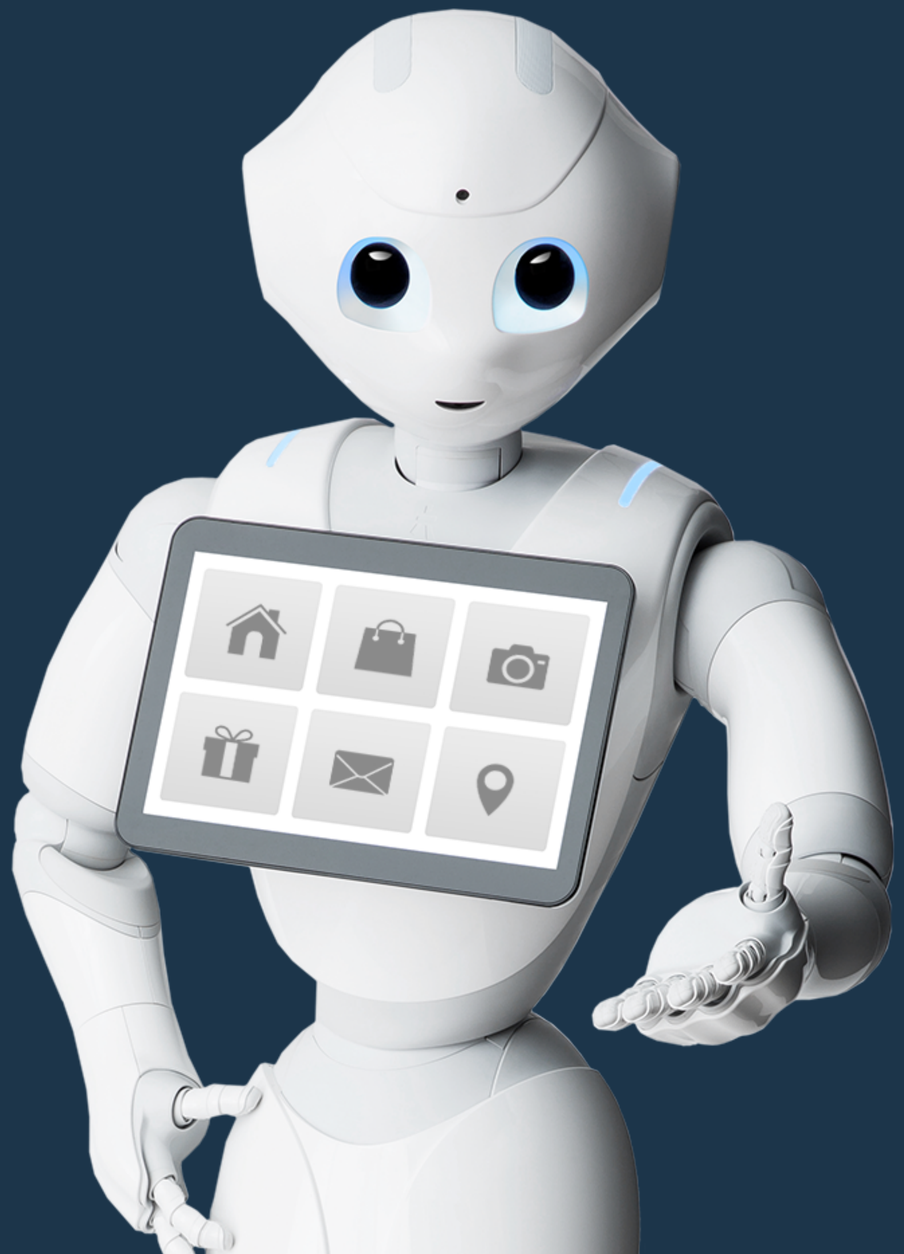
more data

less trust



As designers of programs, systems, and services, **we fail when we turn humans into statistics and data points.**

human/AI relationships



The foundation of a healthy relationship

Purpose

Value

Trust

How do we build **healthy** human/AI relationships in a smart city?

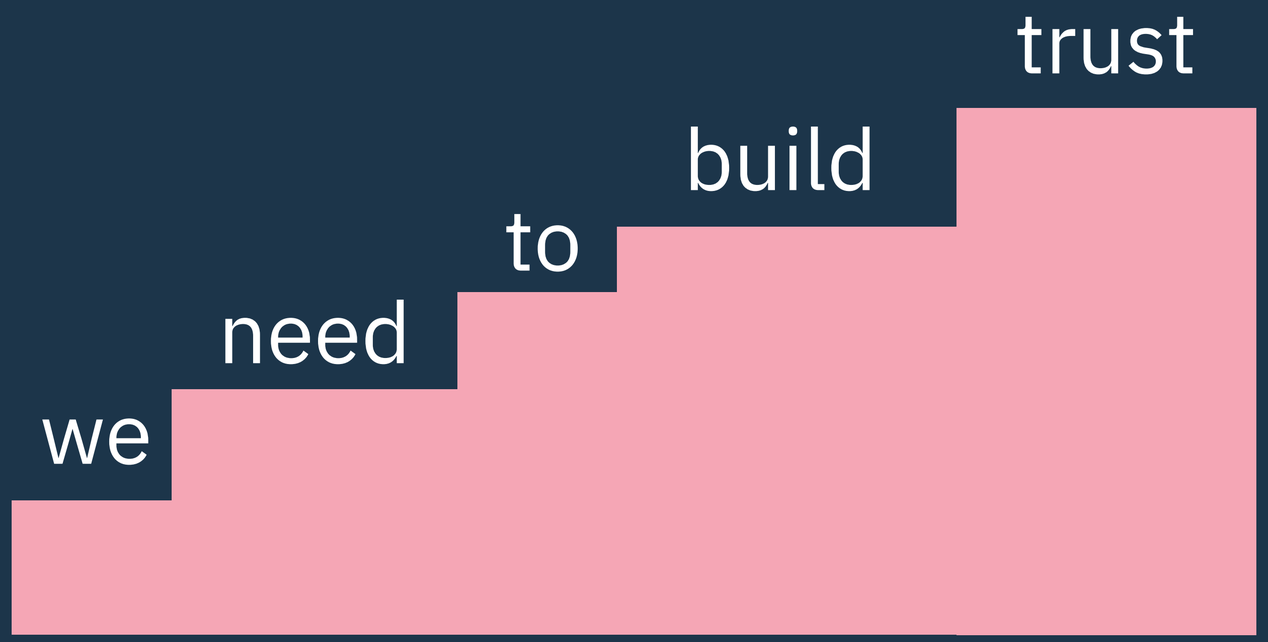
interaction

participation

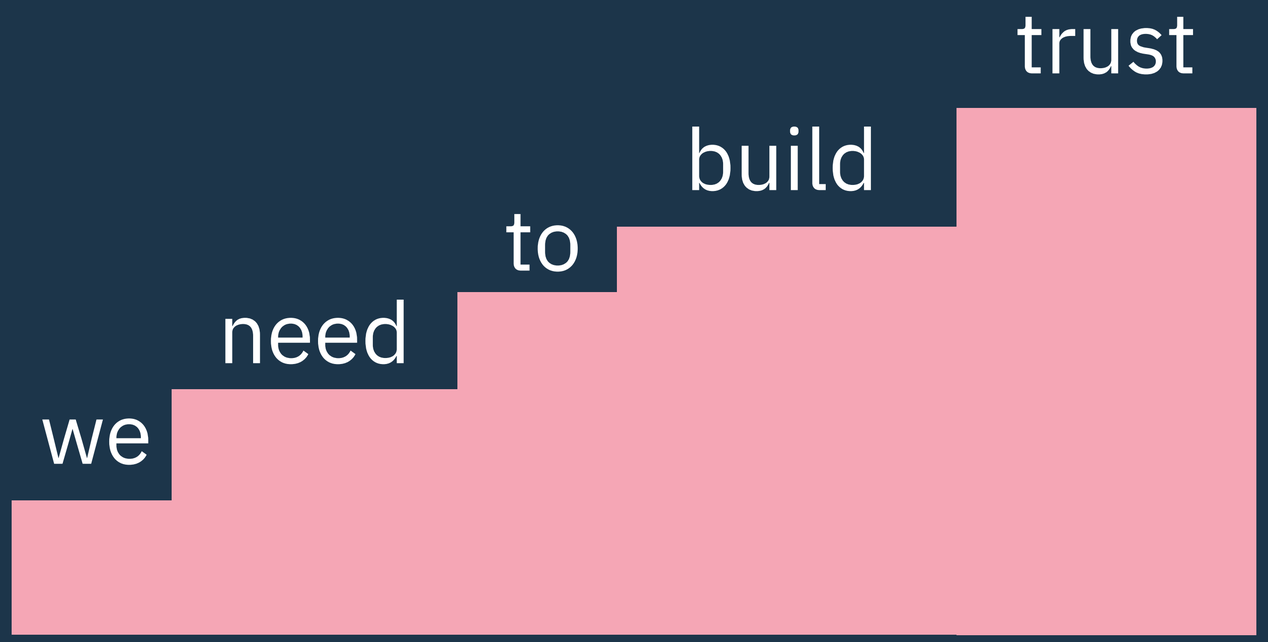
innovation

democratized AI:

a participatory system that operates on majority rule and individual rights


we
need
to
build
trust

okay but how



an ethical foundation for AI

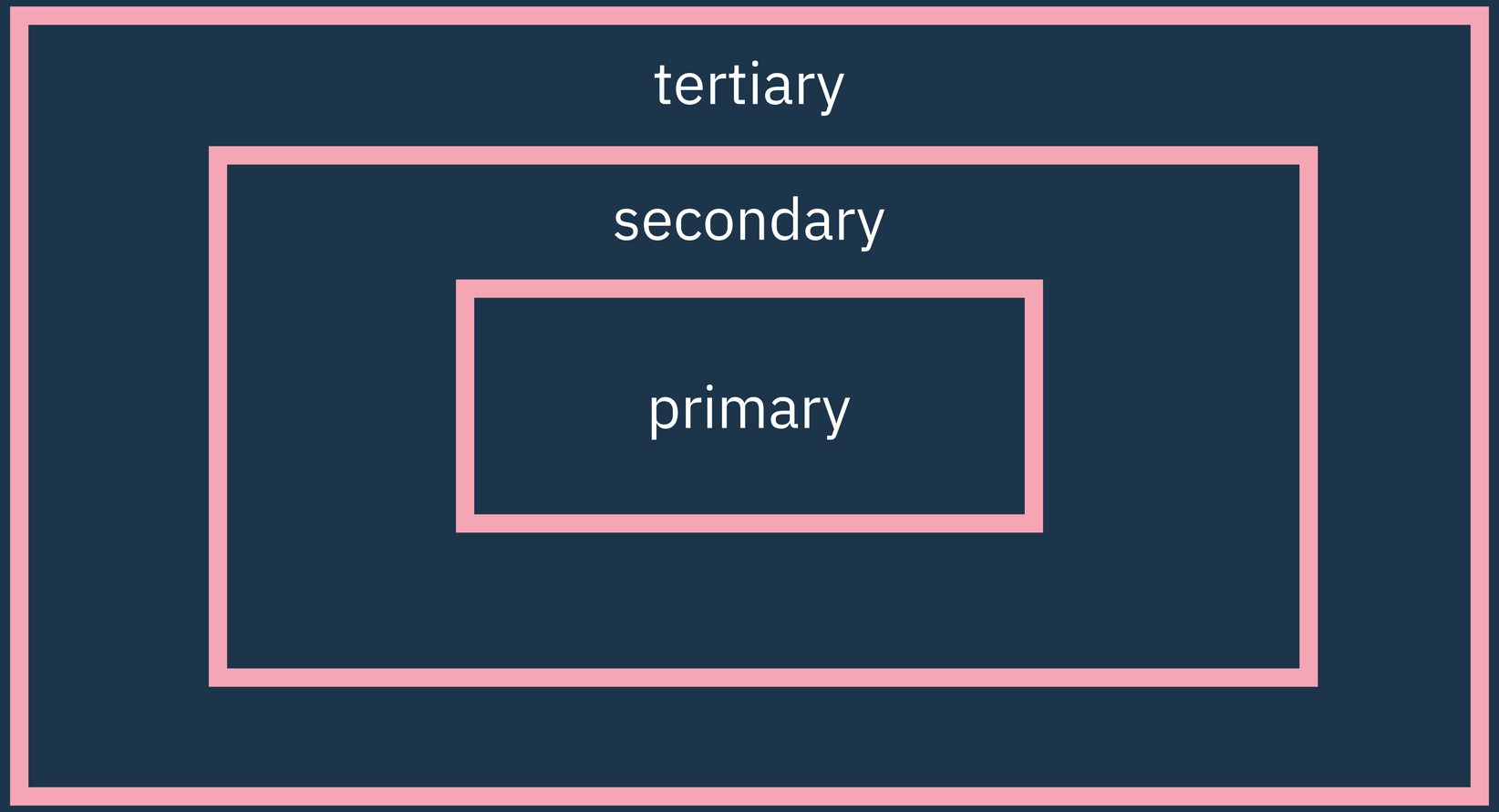
Five Focus Areas

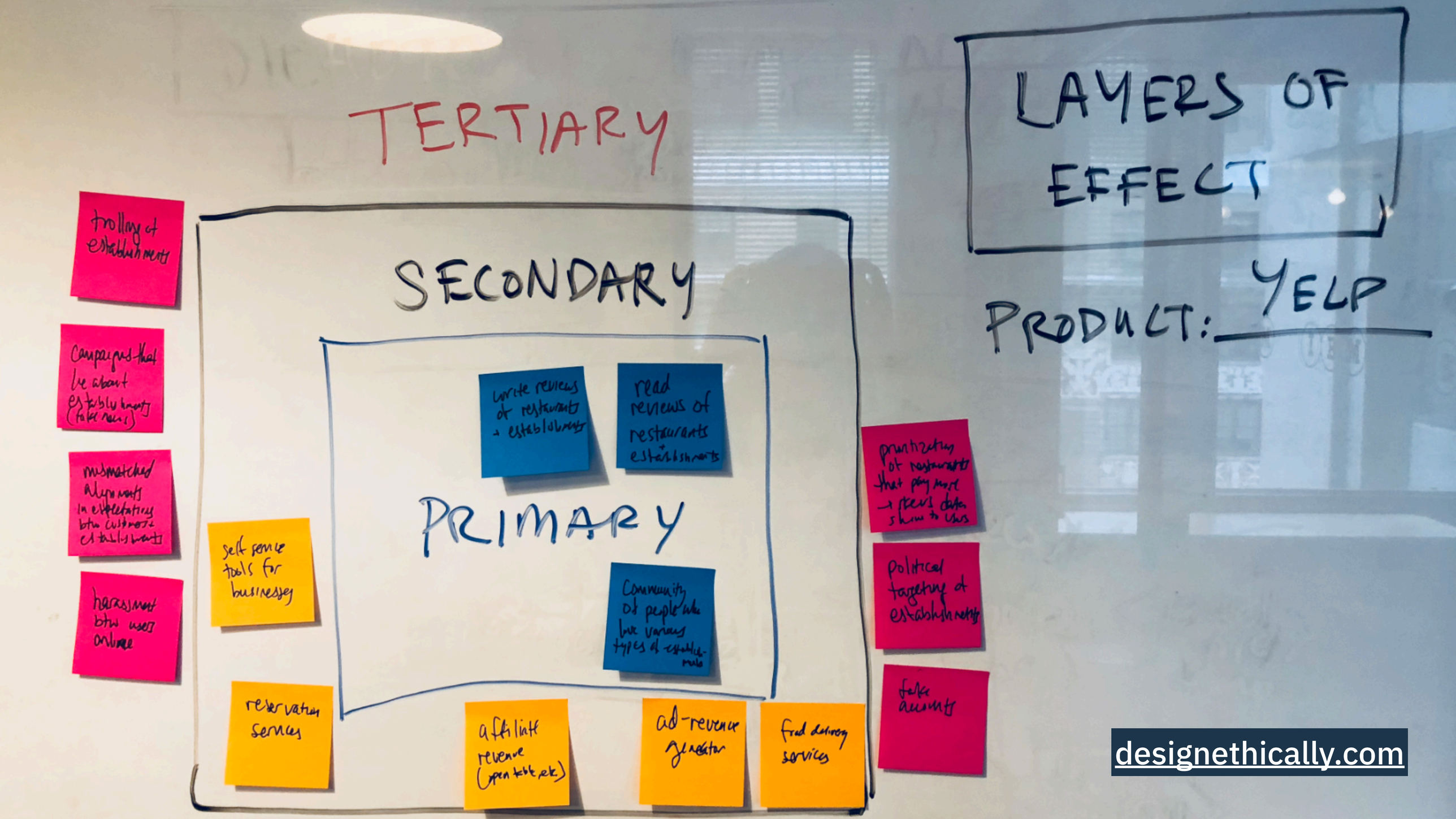
1. Accountability
2. Value Alignment
3. Explainability
4. User Data Rights
5. Fairness

Accountability

Every person involved in the creation of AI at any step is accountable for considering the system's impact in the world.

Accountability




LAYERS OF EFFECT
PRODUCT: YELP
TERTIARY
SECONDARY
PRIMARY
trolling of establishments
campaigns that lie about establishments (fake news)
mismatched alignments in expectations btw customers & establishments
harassment btw users online
write reviews of restaurants + establishments
read reviews of restaurants establishments
self service tools for businesses
community of people who love various types of establishments
reservation services
affiliate revenue (open table etc)
ad-revenue generator
food delivery services
prioritization of restaurants that pay more → steers data shown to users
political targeting of establishments
fake accounts
designethically.com

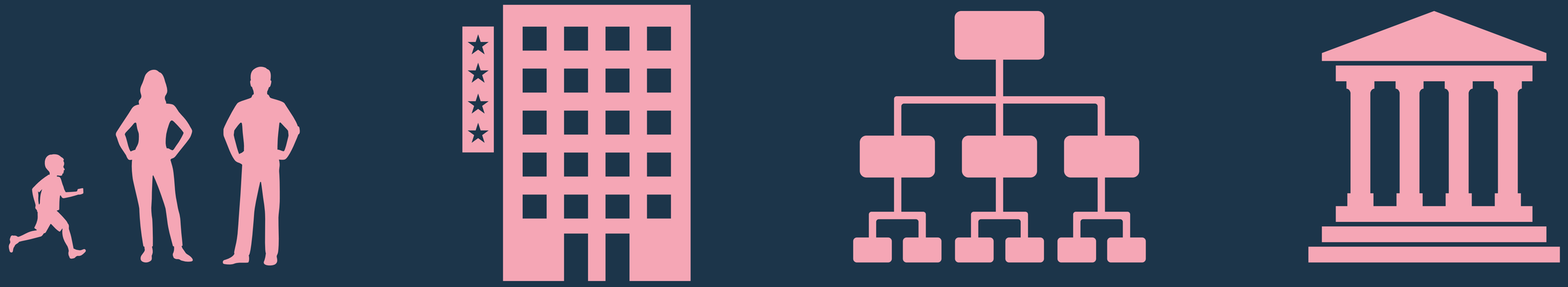
Accountability

“Everybody assumed somebody else knew how it worked.”

Value Alignment

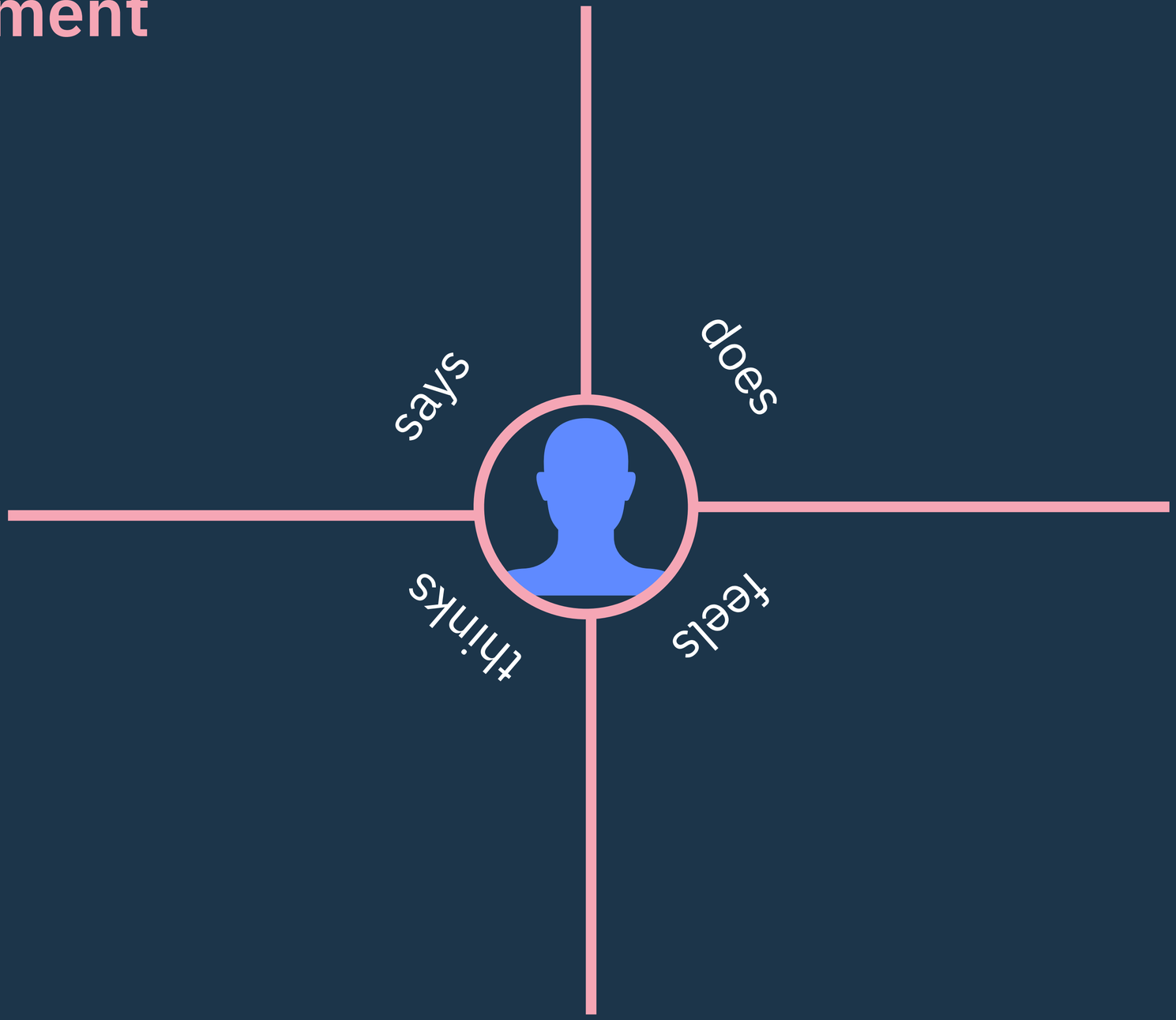
AI should be designed to align with the norms and values of your user group in mind.

Value Alignment



people businesses organizations public-sector

Value Alignment



Value Alignment

“How do we change or adjust the values reflected by our community as our values evolve over time?”

Explainability

AI should be designed for humans to easily perceive, detect, and understand its decision process.

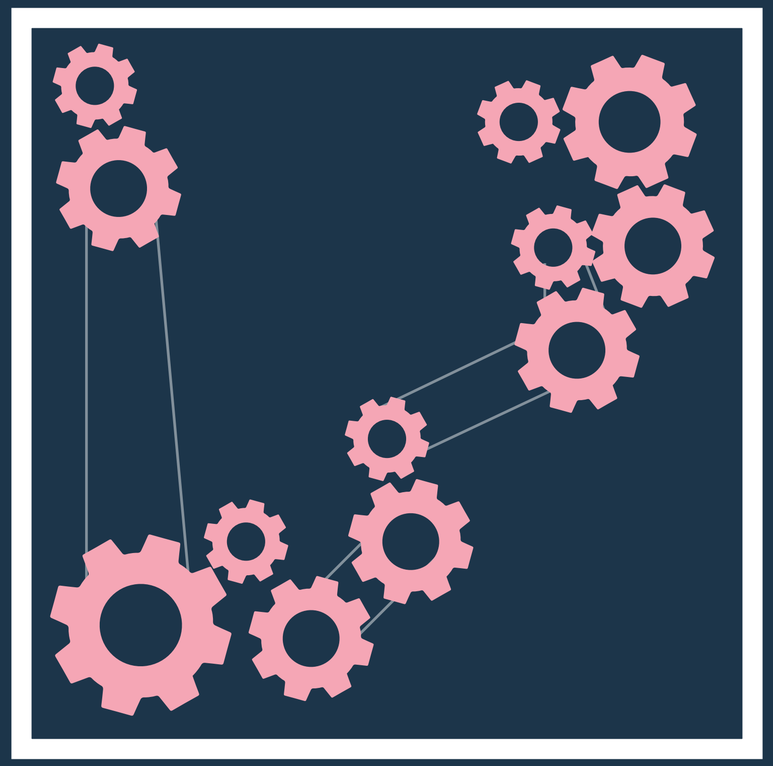
Explainability

input ->

-> output

Explainability

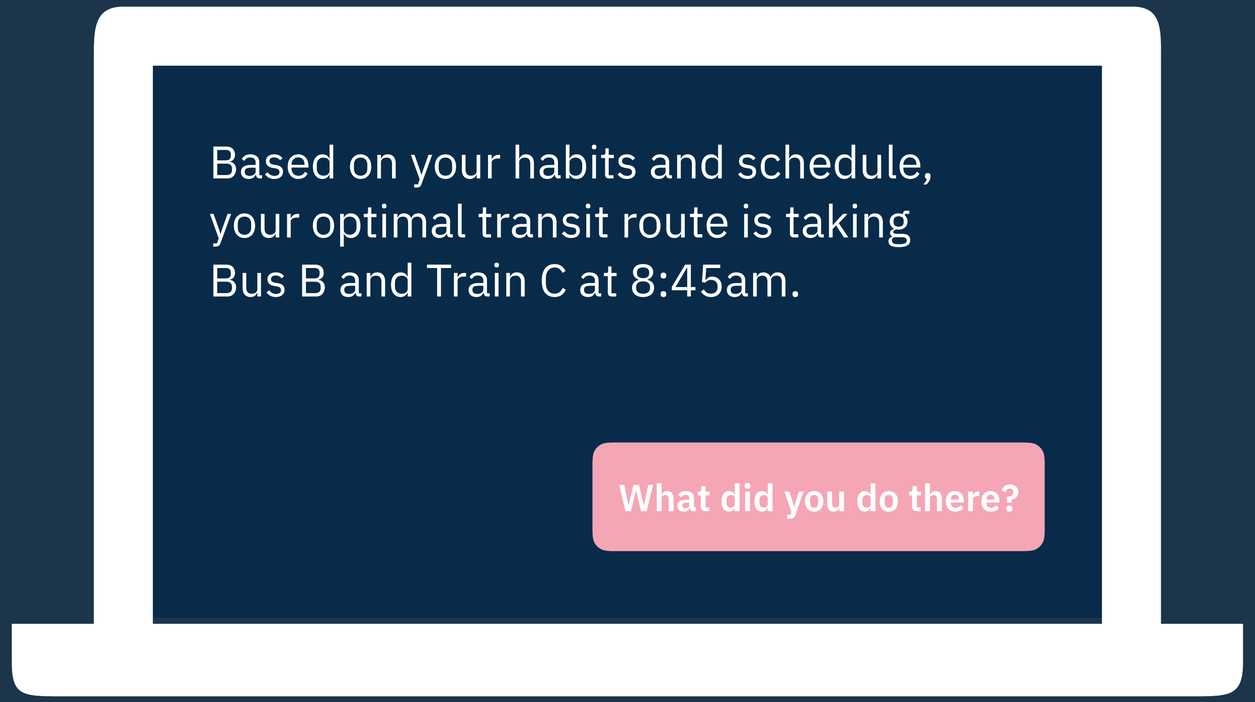
input ->



-> output

our responsibility

Explainability



SmartCity Transit: Official AI Factsheet

Category:

Purpose:

Reasoning:

Risks/ Potential Bias:

User Data Rights

AI must be designed to protect user data and preserve the user's power over access and uses.

User Data Rights

human

machine

User Data Rights

data point

machine

type of data we collect

User Data Rights

type of data we collect

why we need it

User Data Rights

type of data we collect

why we need it

how it empowers people

User Data Rights

experimentation platforms

User Data Rights

organic communities

User Data Rights

Fairness

AI must be designed to minimize bias and promote inclusive representation.

Unconscious Bias Definitions

The average knowledge worker is unaware of the many different types of biases. While this list is not all-encompassing, these biases are some of the more common types to be consciously aware of when designing and developing for AI.

Shortcut Biases

"I don't have the time or energy to think about this."

Availability Bias

Overestimating events with greater "availability" in memory — influenced by how recent, unusual, or emotionally charged the memories may be.

Base Rate Fallacy

The tendency to ignore general information and focus on specific information (a certain case).

Congruence Bias

The tendency to test hypotheses exclusively through direct testing, instead of testing alternative hypotheses.

Empathy Gap Bias

The tendency to underestimate the influence or strength of feelings, in either ones' self or others.

Stereotyping

Expecting a member of a group to have certain characteristics without having actual information about that individual.

Impartiality Biases

"I know I'm wrong sometimes, but I'm right about this."

Anchoring Bias

To rely too much on one trait or piece of information when making decisions (usually the first piece of information that we acquire on that subject.

Bandwagon Bias

The tendency to do or believe things because many other people do. (Groupthink)

Bias Blind Spot

The tendency to see oneself as less biased than others, or to be able to identify more cognitive biases in others than in oneself.

Confirmation Bias

The tendency to search for, interpret, or focus on information in a way that confirms one's preconceptions.

Halo Effect

The tendency of an overall impression to influence the observer. Positive feelings in one area causes ambiguous or neutral traits to be viewed positively.

Self-Interest Biases

"We contributed the most. They weren't very cooperative."

Ingroup / Outgroup Bias

The tendency or patter of favoring members of one's ingroup over outgroup members.

Sunk Cost Bias

The tendency to justify past choices, even though they no longer seem valid.

Status Quo Bias

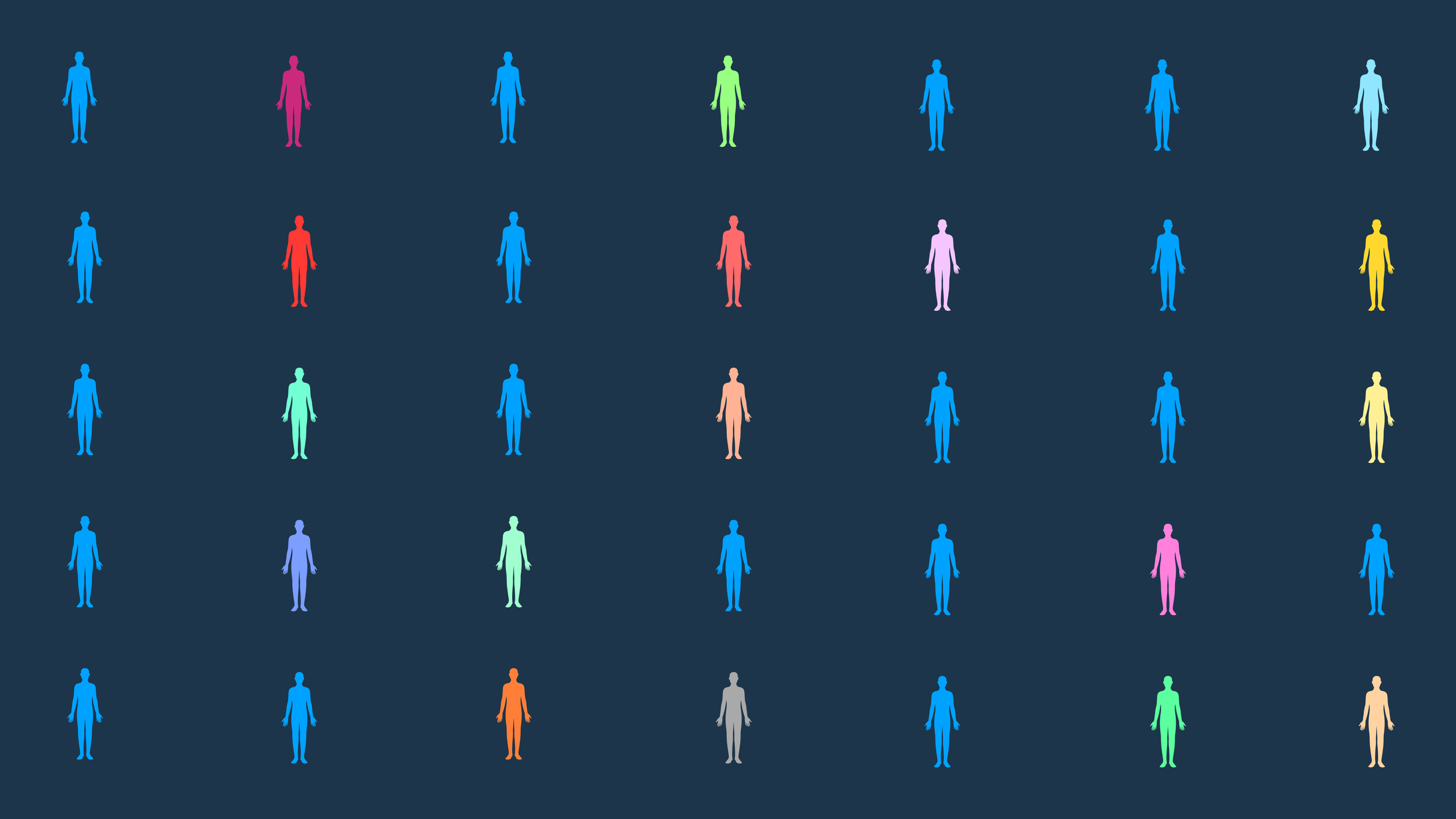
The tendency to maintain the current situation — even when better alternatives exist.

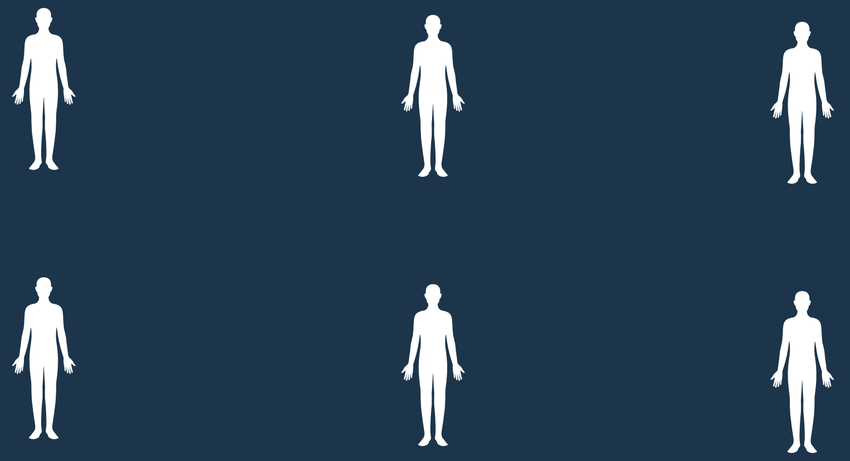
Not Invented Here Bias

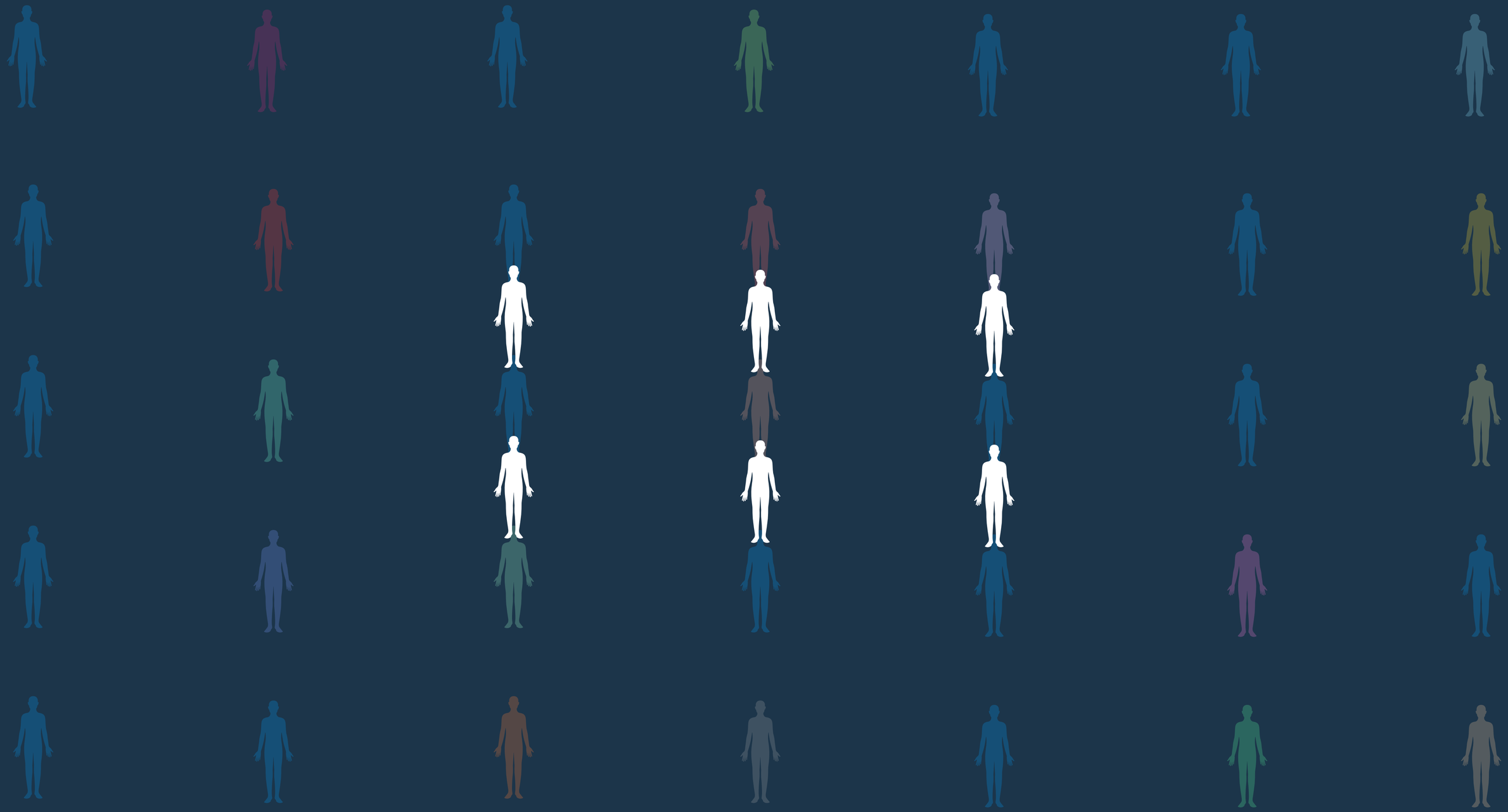
Aversion to contact with or use of products, research, standards, or knowledge developed outside a group.

Self-Serving Bias

The tendency to focus on strengths/ achievements and overlook faults/failures. To take more responsibility for their group's work that they give to other groups.







Fairness

equal representation

Fairness

equal representation
+
equal participation

Five Focus Areas

1. Accountability
2. Value Alignment
3. Explainability
4. User Data Rights
5. Fairness

ibm.biz/everydayethics

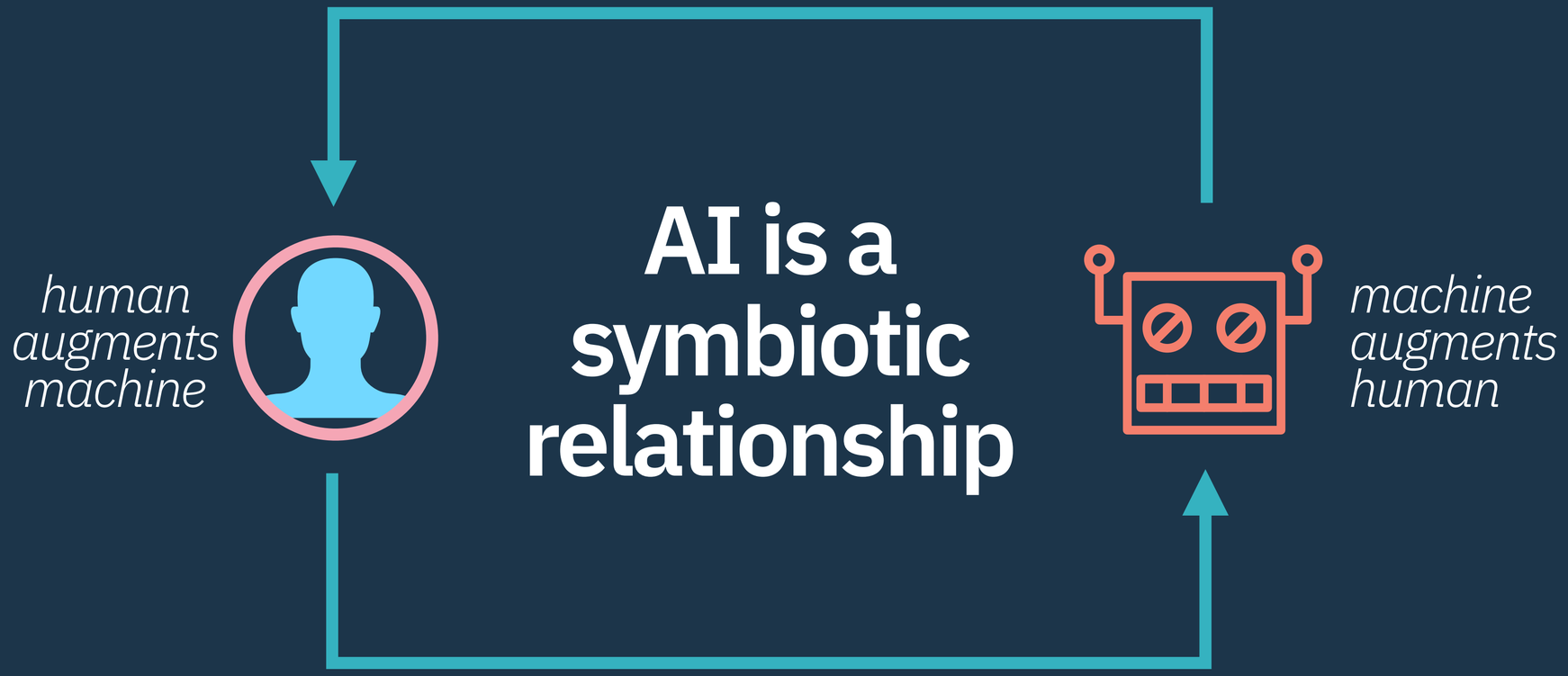
There’s a **market advantage** in building trust.

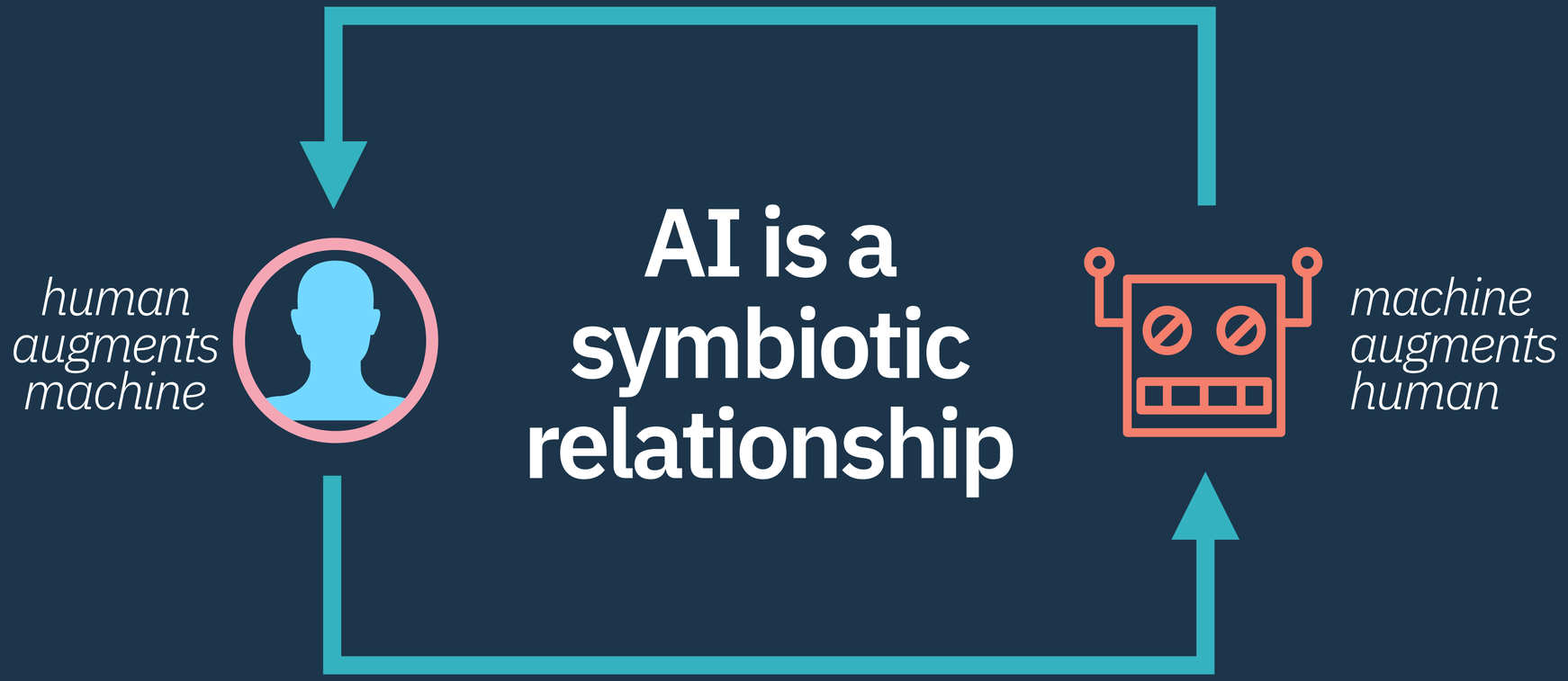
inclusivity

bias reduction

efficiency

scalability


human
augments
machine
AI is a
symbiotic
relationship
machine
augments
human


human
augments
machine
AI is a
symbiotic
relationship
machine
augments
human

Thank you!

@milenapribic
https://www.linkedin.com/in/milenapribic/