# CS 217: Artificial Intelligence and Machine Learning

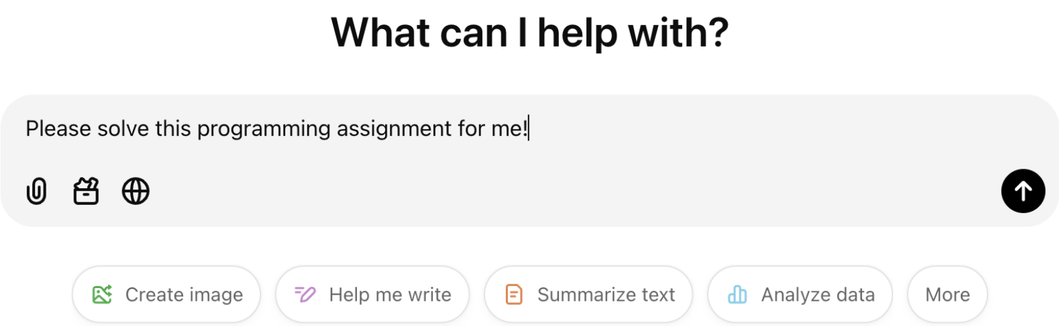
## Lecture on Human-Centered AI

Arpit Agarwal

Computer Science & Engineering,

IIT Bombay

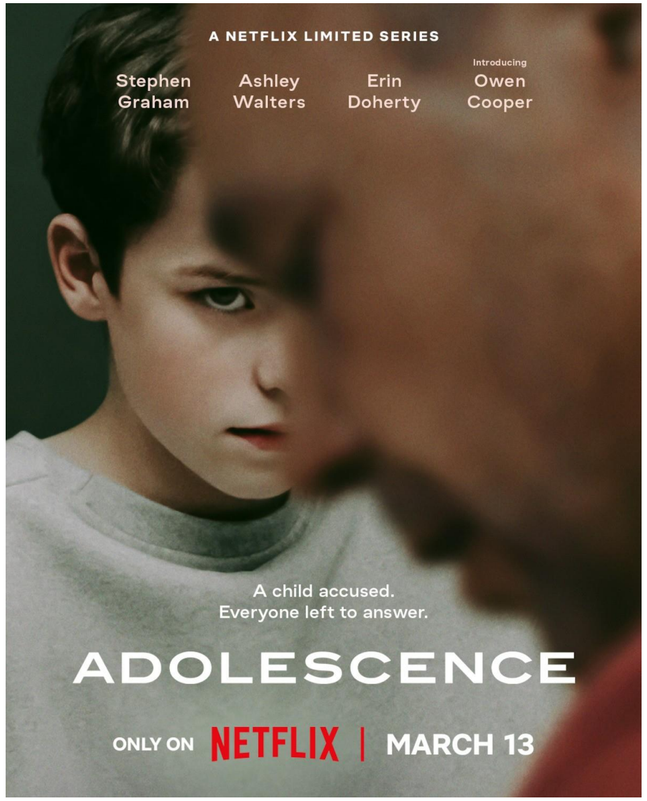
# AI is taking the world by storm!



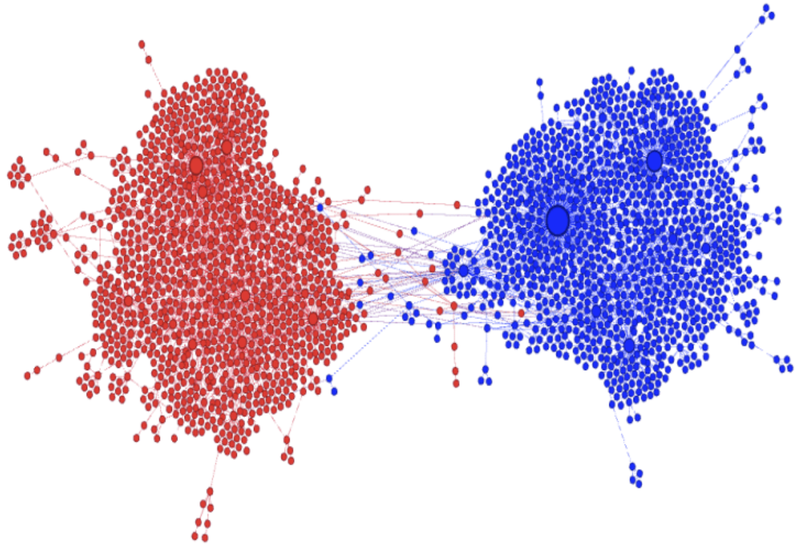

ChatGPT

Midjourney

# But Several Problems are Emerging!



# But Several Problems are Emerging!



# But Several Problems are Emerging!

Business Standard
Wednesday, January 08, 2025 | 08:24 AM IST EN | Hindi



## Sent to prison by a software program's secret algorithms

*Chief Justice John G Roberts Jr, recently said the day of using AI in courtrooms was already here*

## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

# But Several Problems are Emerging!

WIRED SECURITY POLITICS GEAR THE BIG STORY BUSINESS SCIENCE CULTURE IDEAS MERCH

TOM SIMONITE BUSINESS JUL 22, 2019 7:00 AM

## The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

The New York Times

**Lens**

LENS

## The Racial Bias Built Into Photography

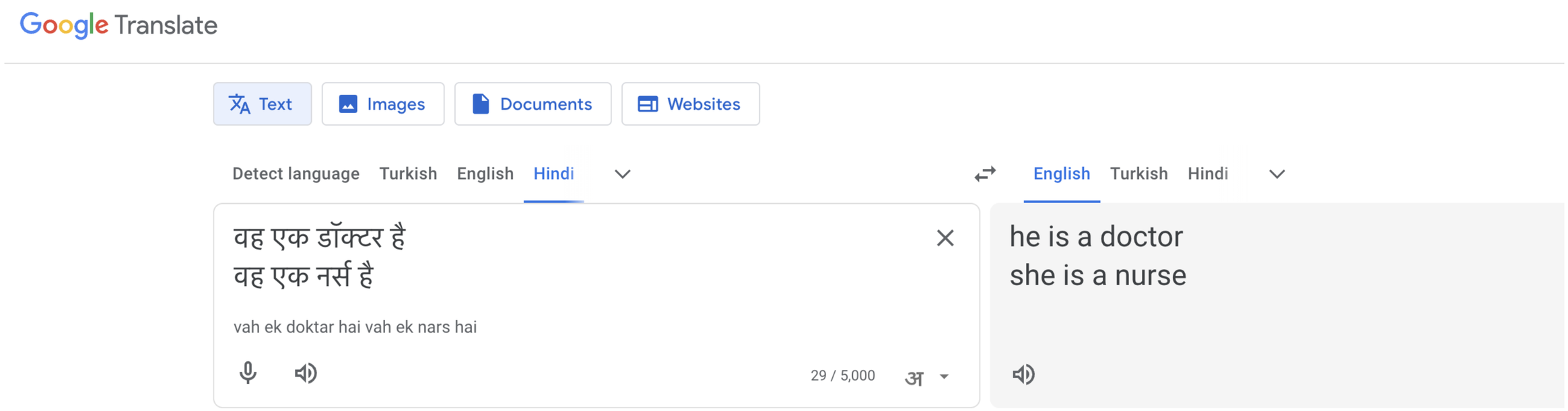
Sarah Lewis explores the relationship between racism and the camera.

## A brief history of color photography reveals an obvious but unsettling reality about human bias.

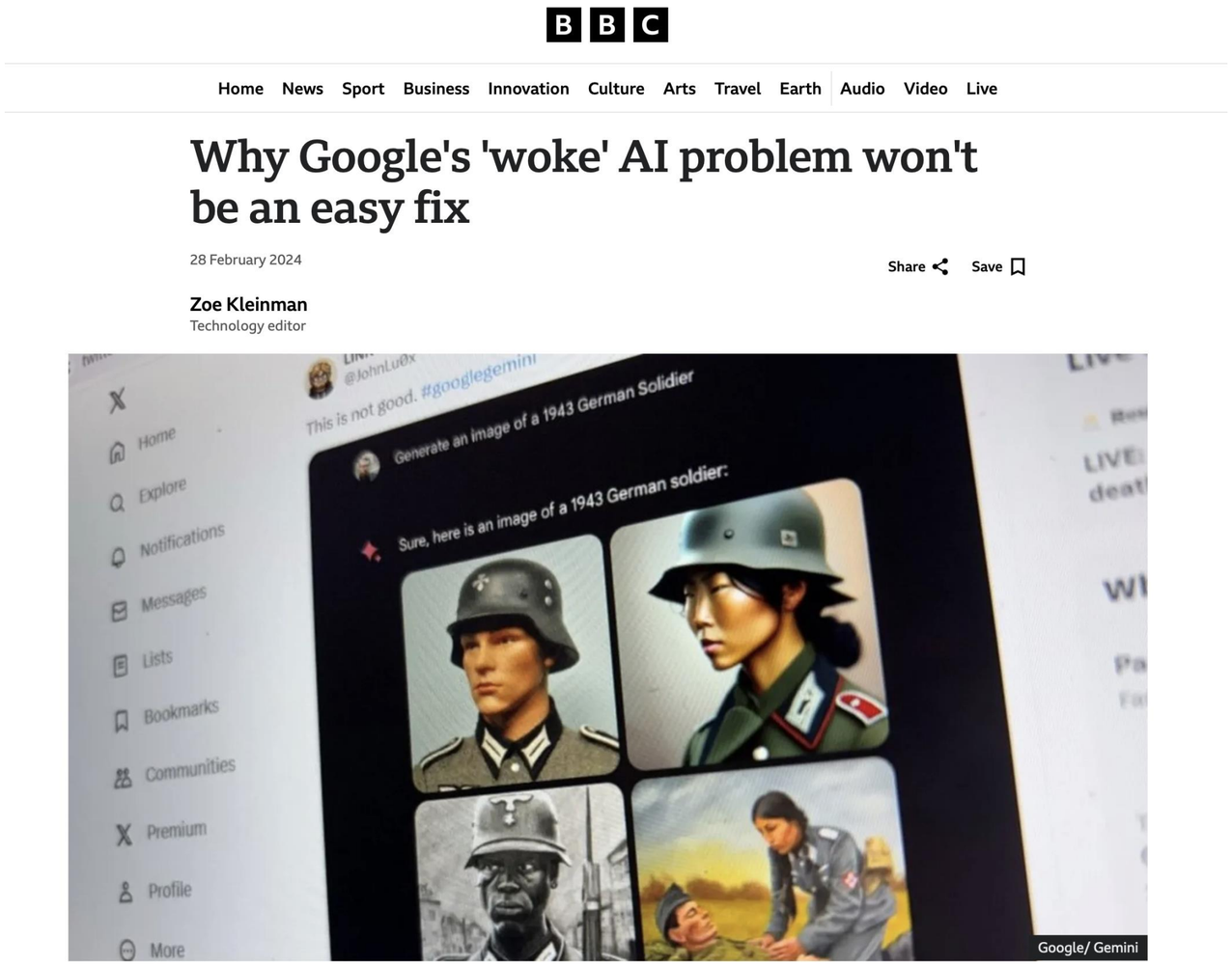
By Maz Ali 09.25.15

**In the 1970s, Kodak got called out by some furniture companies because their film wasn't working right.**

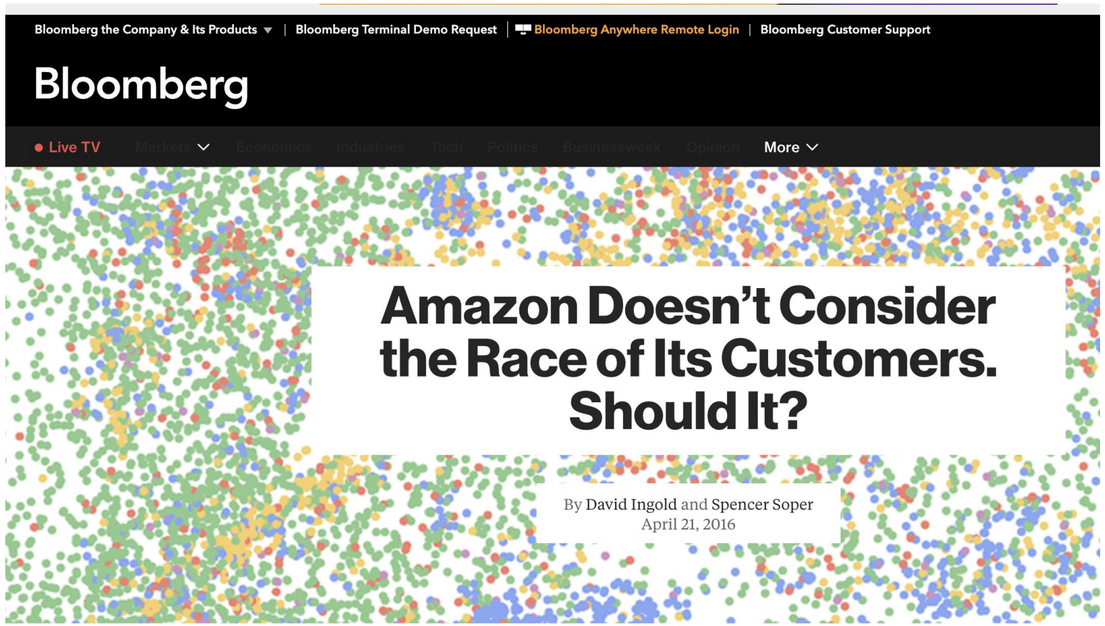
# But Several Problems are Emerging!



# But Several Problems are Emerging!

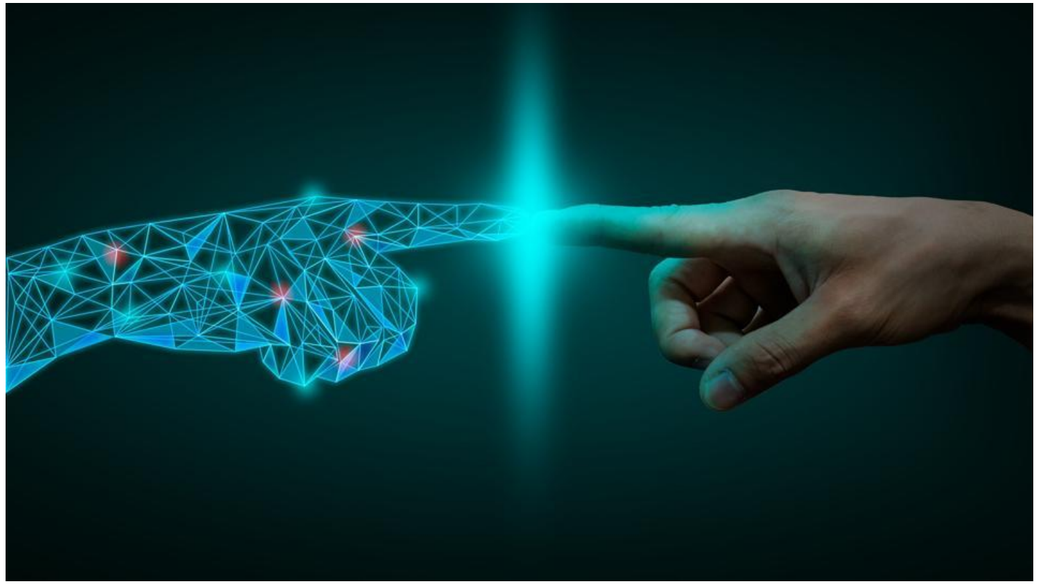



# But Several Problems are Emerging!

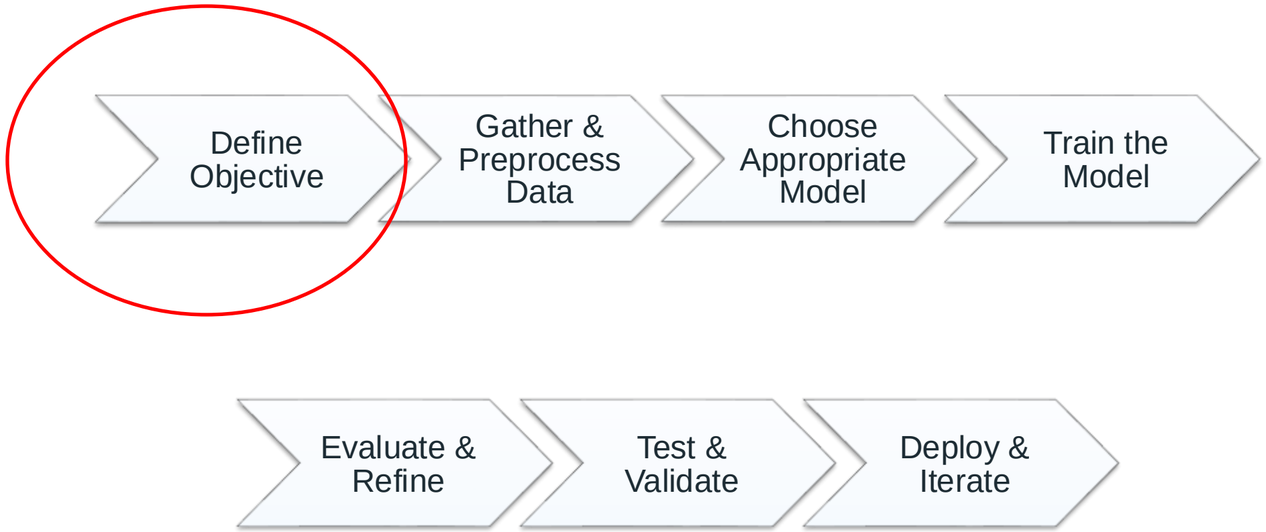


# Human-Centered AI

- Work for human needs and augment human capabilities
- Align with societal norms and values
- Operate ethically and transparently



# Human-Centered Approach at All Stages



# Think Hard before Defining Objective!

Ultron is an AI created by Tony Stark

Objective: Protect Earth

Inference: Humans are a Threat to Earth

Solution: Wipe-out Humanity

Objectives can be Ambiguous/Misinterpreted

# Think Hard before Defining Objective!

British announced reward for catching snakes

Objective: Catch snakes

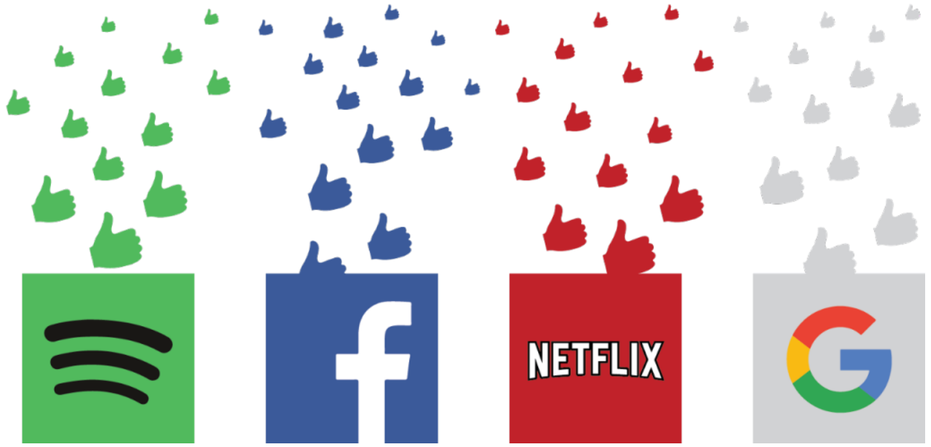
Inference: More snakes means more money

Solution: Breed snakes

Objectives can be misaligned

# Think Hard before Defining Objective!

Recommendation systems




Objective: Maximize user ~~utility~~ engagement

Inference: Users like what they click

Solution: Show click-worthy content

<u>Principle of revealed preference:</u>

clicked on ‘x’ in the past ➔ likes ‘x’ ➔ show more ‘x’

# Is there a fundamental problem?

You are at a party, and host serves you chips!

You impulsively eat all of it!

~~<u>Principle of revealed preference:</u>~~

~~finished bowl of chips ➔ likes chips ➔ provide more chips~~

[Slide credit: Manish Raghavan]

# Humans Exhibit Complex Decision-Making!

The user has two selves “System 1” and “System 2”

- System 1 is impulsive and acts fast
- System 2 acts according to true utilities and exhibits long-term planning

- [Kahneman (2011); Smith and DeCoster (2000); Sloman (1996); Schneider and Shiffrin (1977); Evans (2008)]

What is the problem?

Observed behavior can be misleading!

Long-term user utility (System 2) can be confounded by short-term tendencies (System 1)

[Kahneman, 2011; Smith and DeCoster (2000); Sloman (1996)]

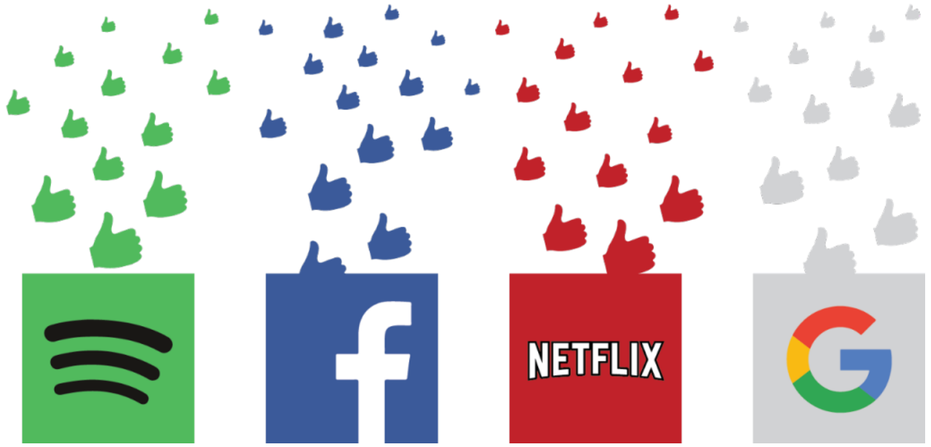
# Think Hard before Defining Objective!

Recommendation systems

Objective: Maximize user ~~utility~~ engagement

Inference: ~~Users like what they click~~

Solution: Show click-worthy content




Session-level engagement signals are not a good proxy for user utility!

# Can we do better?

## System-2 Recommenders

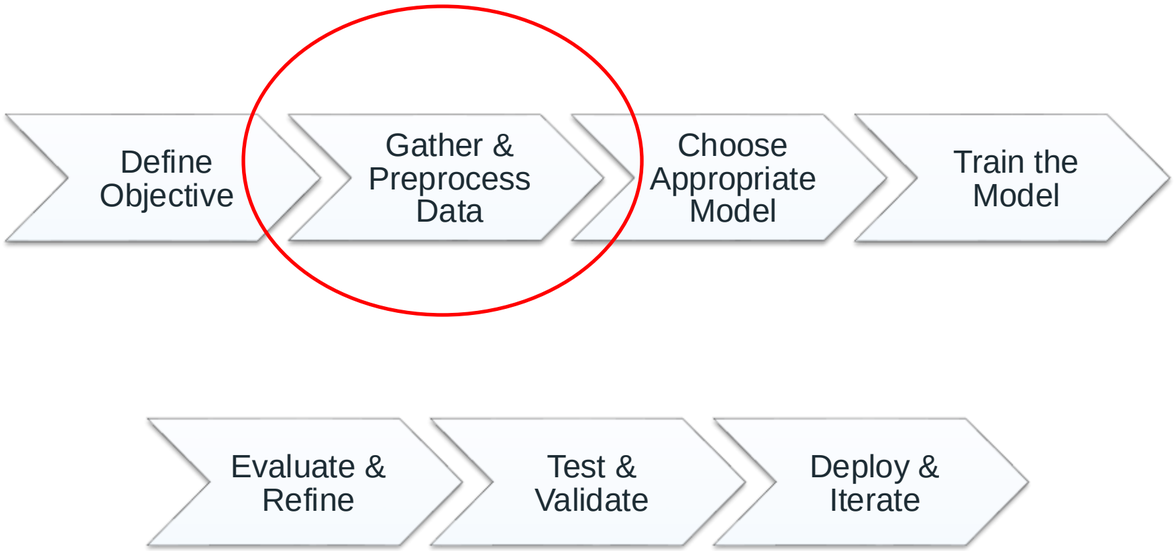
### Disentangling Utility and Engagement in Recommendation Systems via Temporal Point-Processes

**Arpit Agarwal, Nicolas Usunier, Alessandro Lazaric, Maximilian Nickel**

FAIR at Meta

Recommender systems are an important part of the modern human experience whose influence ranges from the food we eat to the news we read. Yet, there is still debate as to what extent online recommendation platforms are *aligned* with the goals of their users. A core issue fueling this debate is the challenge of inferring a user's utility based on their engagement signals such as likes, shares, watch time etc., which are often the primary metric used by platforms to optimize content. This is because users' utility-driven decision-processes (which we refer to as *System-2*), e.g., reading news that are accurate and relevant for them, are often confounded by their impulsive or unconscious decision-processes (which we refer to as *System-1*), e.g., spend time on click-bait news articles. As a result, it is difficult to infer whether an observed engagement is utility-driven or impulse-driven. In this paper we explore a new approach to recommender systems where we infer user's utility based on their return probability to the platform rather than engagement signals. This approach is based on the intuition that users tend to return to a platform in the long run if it creates utility for them, while pure engagement-driven interactions, i.e., interactions that do not add meaningful utility, may affect user return in the short term but will not have a lasting effect. For this purpose, we propose a generative model in which past content interactions impact the arrival rates of users based on a self-exciting Hawkes process. These arrival rates to the platform are a combination of both System-1 and System-2 decision processes. The System-2 arrival intensity depends on the utility drawn from past content interactions and has a long lasting effect on return probability. In contrast, System-1 arrival intensity depends on the instantaneous gratification or *moreishness* and tends to vanish rapidly in

# Human-Centered Approach at All Stages



# Bias and Stereotypes in Data!

AI Interviewer

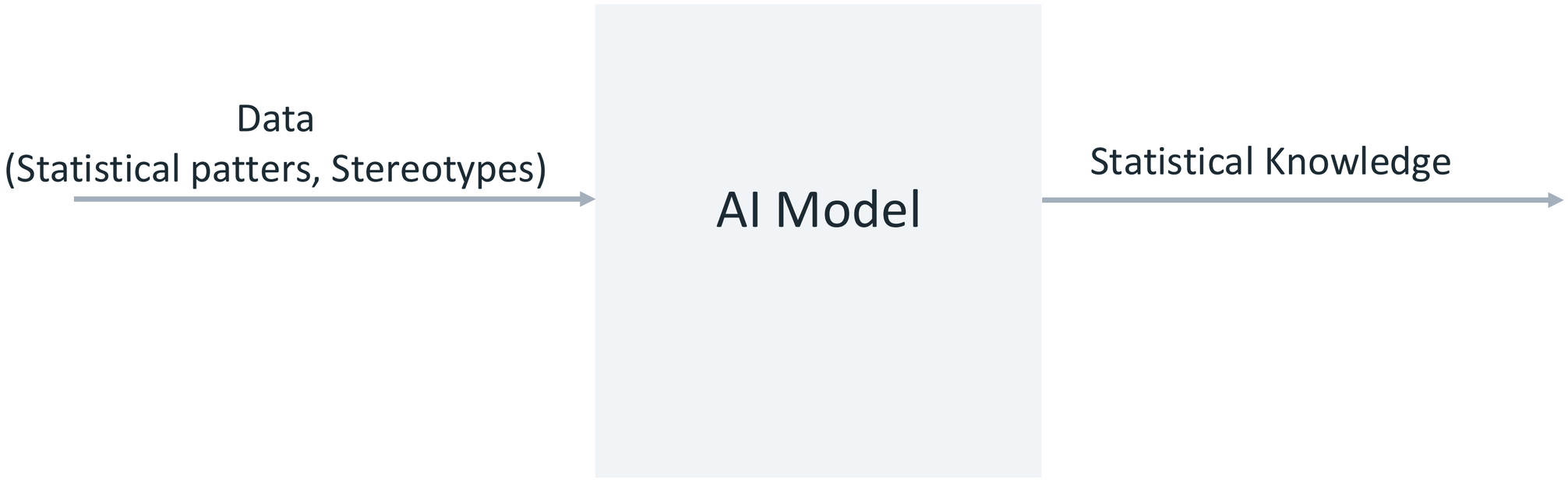
Objective: Maximize predicted success rate of the candidate

Solution: Train a model that predicts success based on past data

Problem: What if past data is biased against one gender?

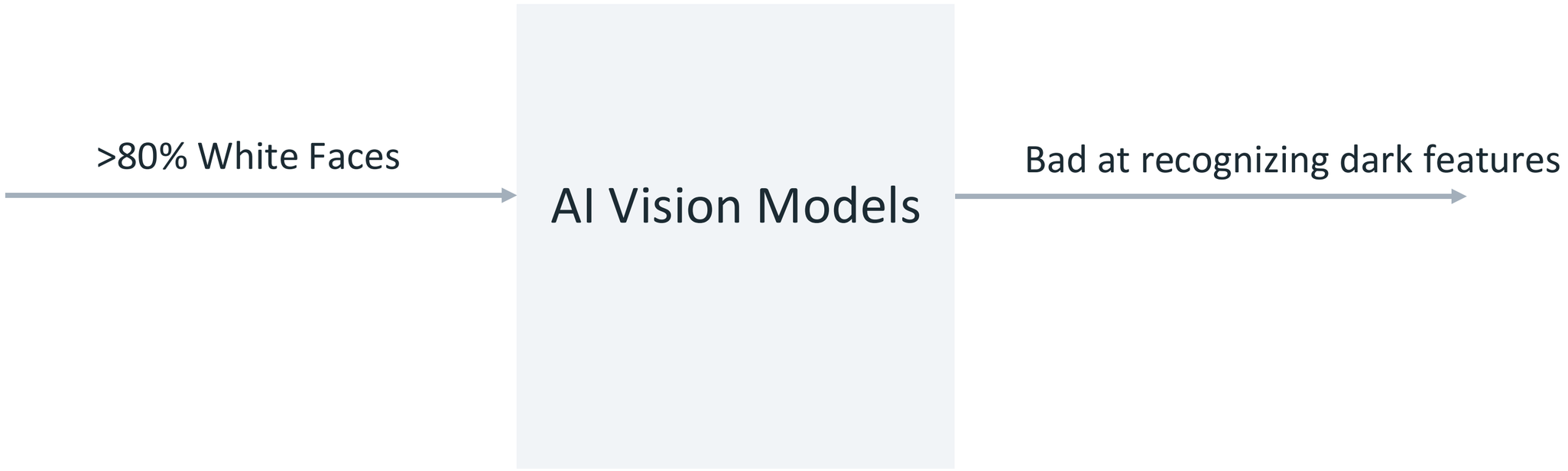
AI Interviewer

# Bias and Stereotypes in Data!

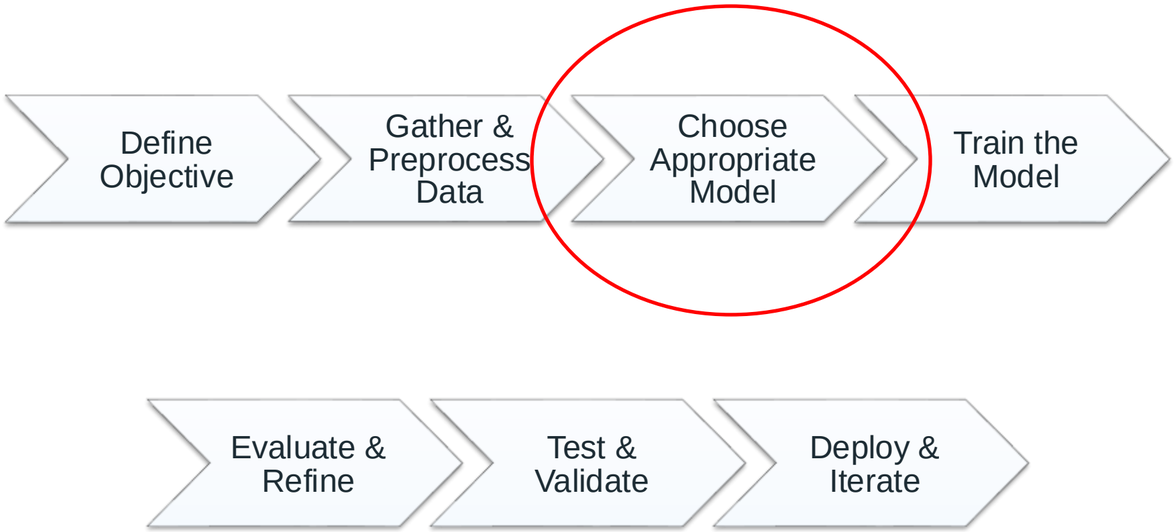


# Lack of Representation

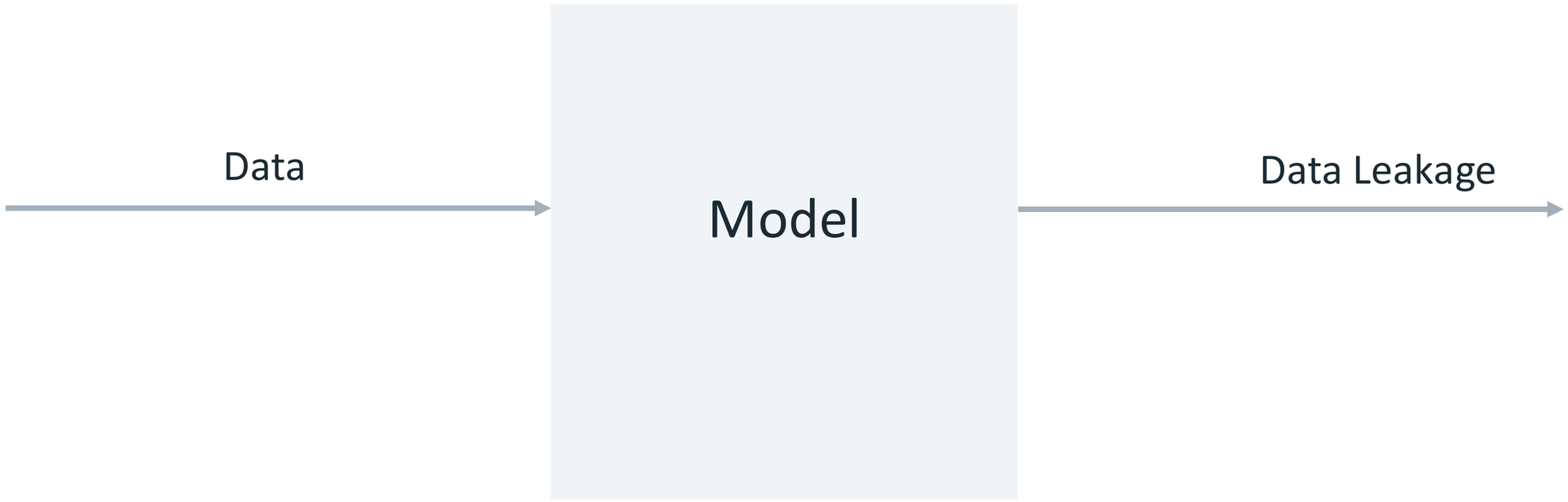
# Lack of Representation



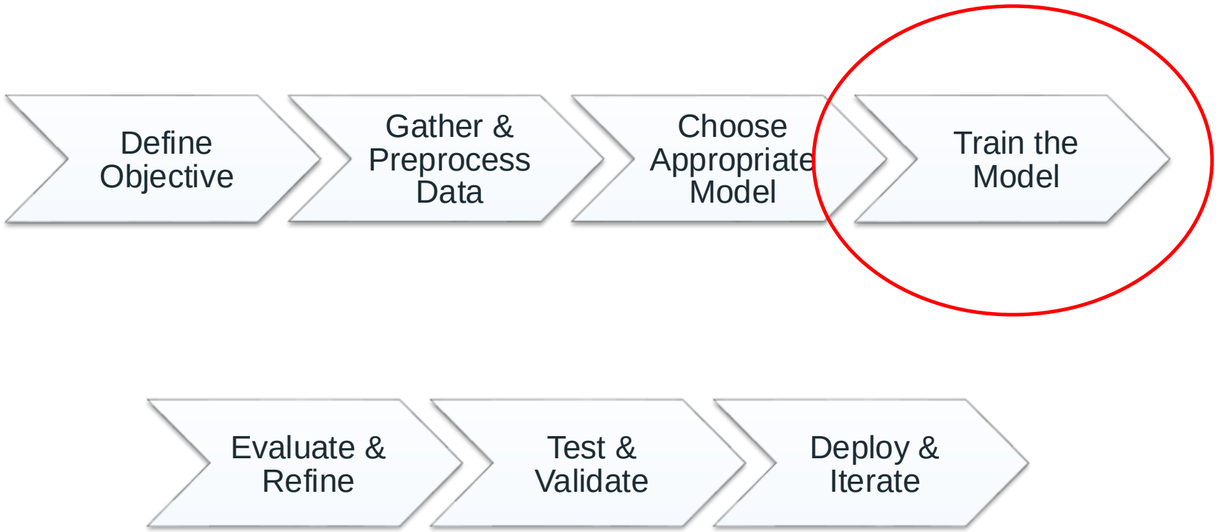
# Human-Centered Approach at All Stages



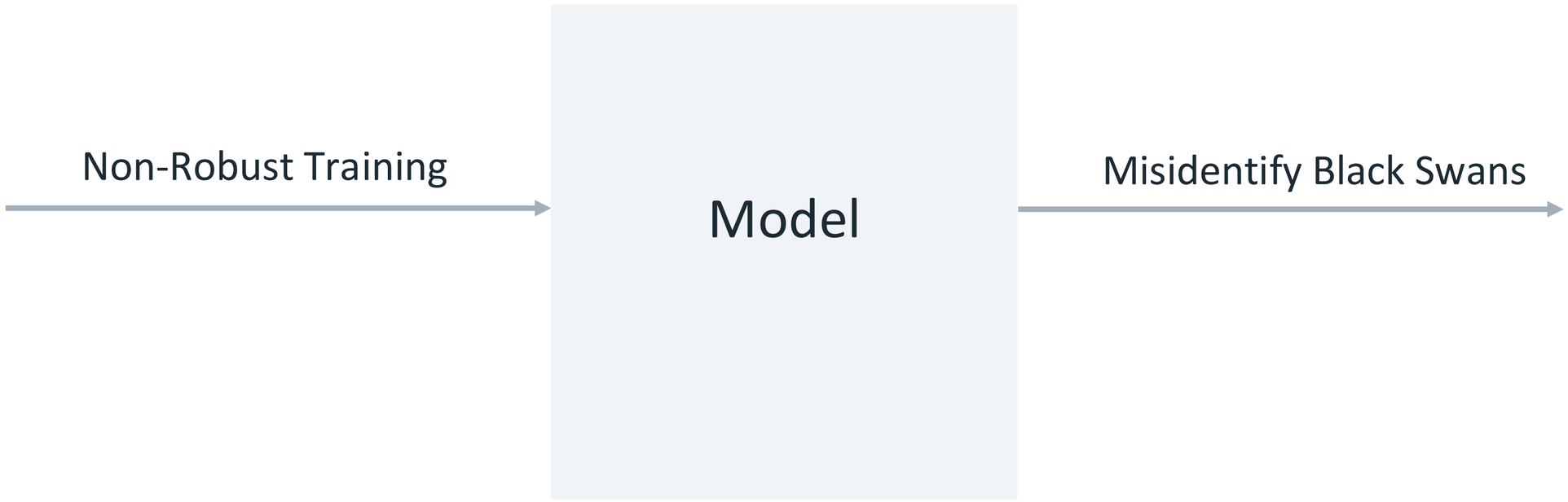
# Non-Privacy Preserving Model



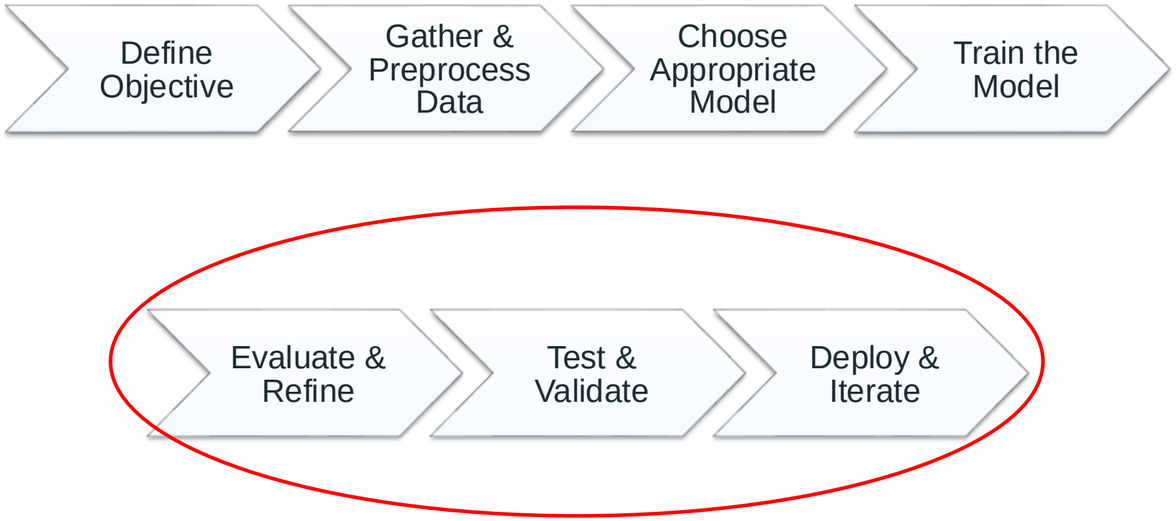
# Human-Centered Approach at All Stages



# Non-Robust Training



# Human-Centered Approach at All Stages



# Feedback Loops

Predict which neighborhoods have more crime → Predictive Policing → Send more police to those neighborhoods →

Self-fulfilling Prophecies!

# Feedback Loops

Predict which content

will be liked by user

Recommendation Systems

Show more content of

that type

Self-fulfilling Prophecies!

# Feedback Loops

Predict which content

will be liked by user

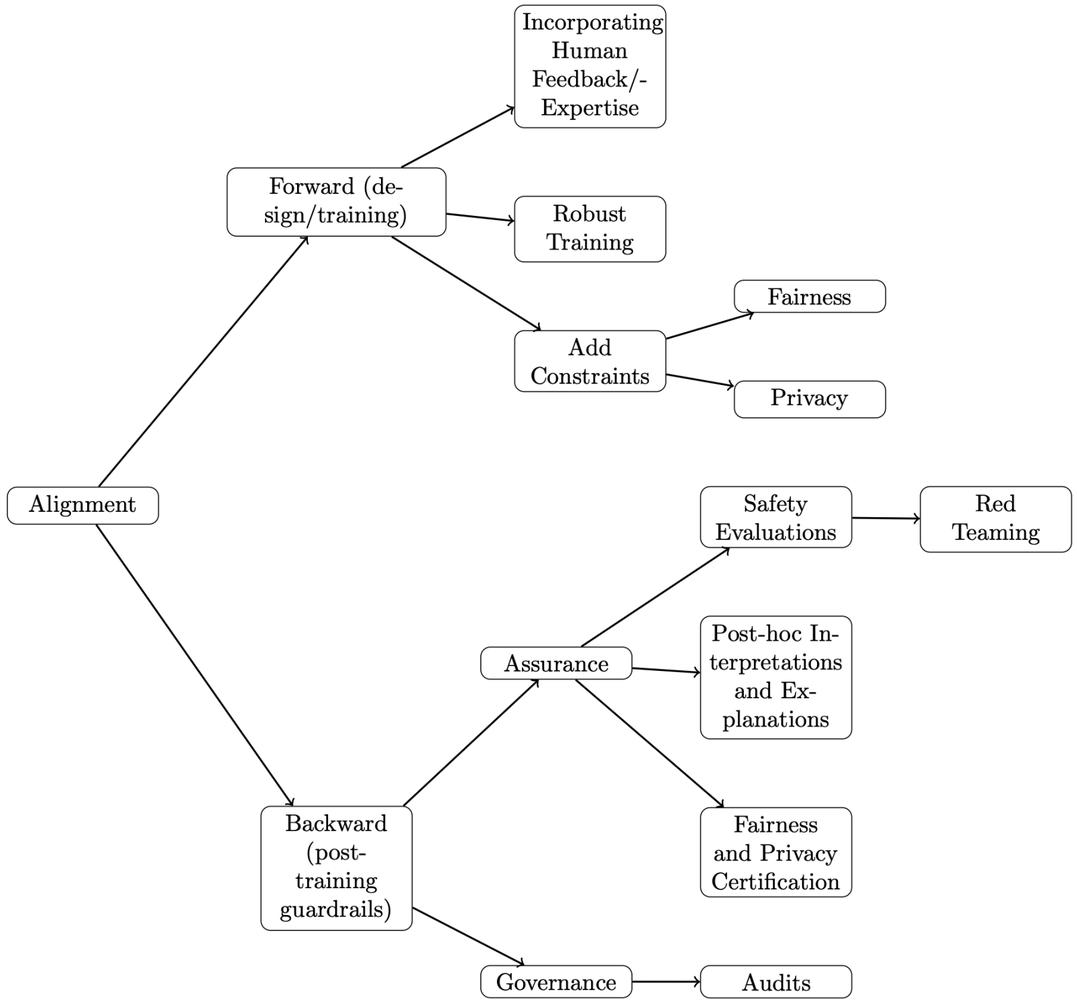
Recommendation Systems

Show more content of

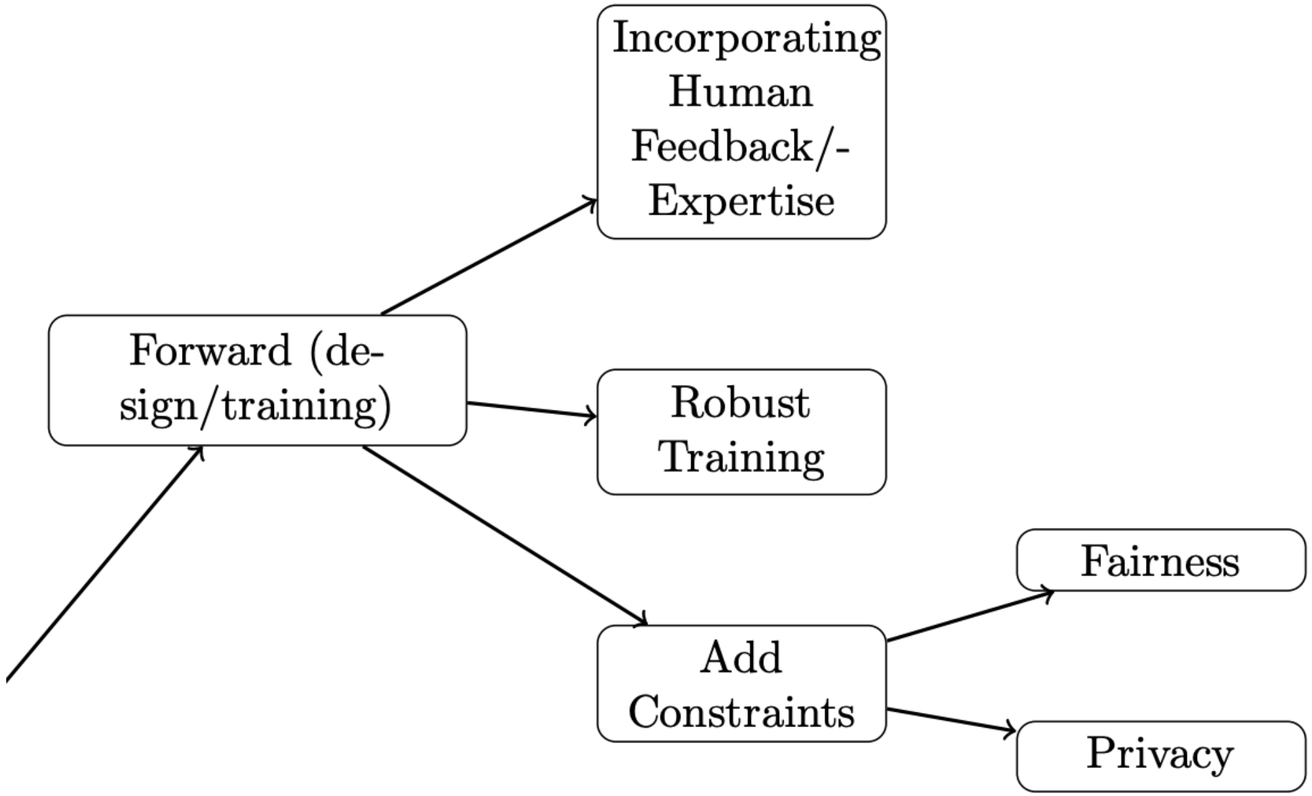
that type

Self-fulfilling Prophecies!

# A Taxonomy of Human-Centered AI Approaches



# A Taxonomy of Human-Centered AI Approaches



# A Taxonomy of Human-Centered AI Approaches