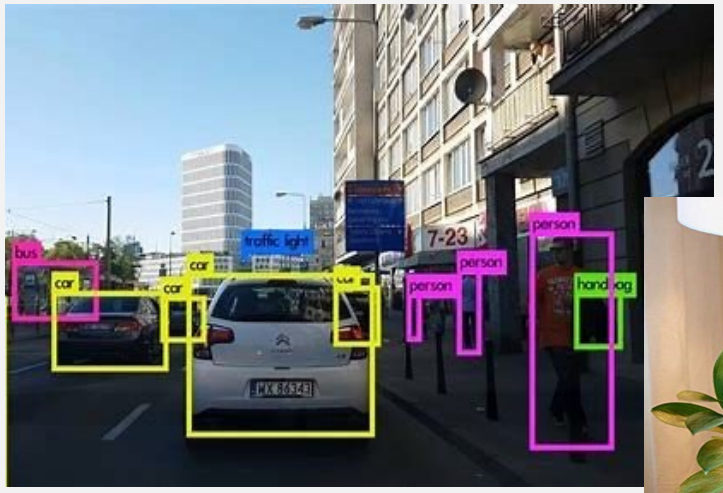
# HUMAN-IN-THE-LOOP LEARNING OF FAIRNESS–PRIVACY TRADE-OFF

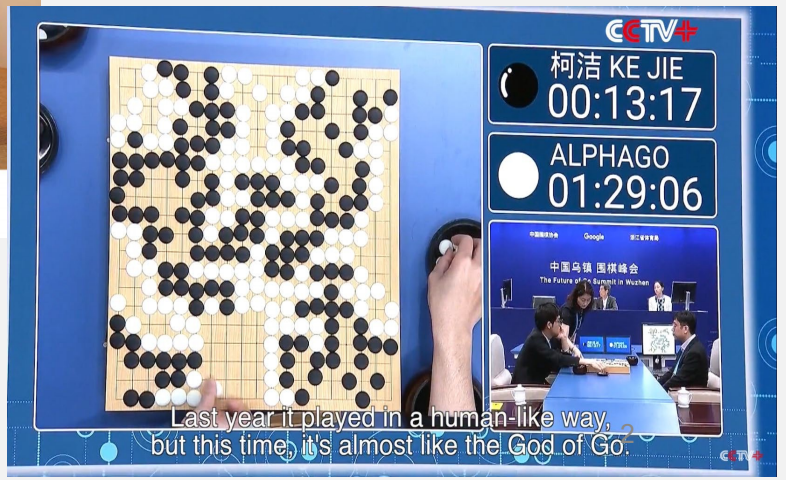
Anantaa Kotal

NSF RET: CREEDS Summer 2025

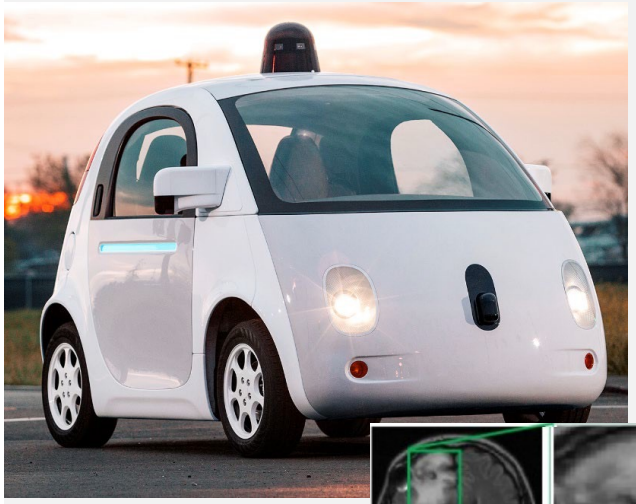
# AI ACHIEVES OR EXCEEDS HUMAN PERFORMANCE

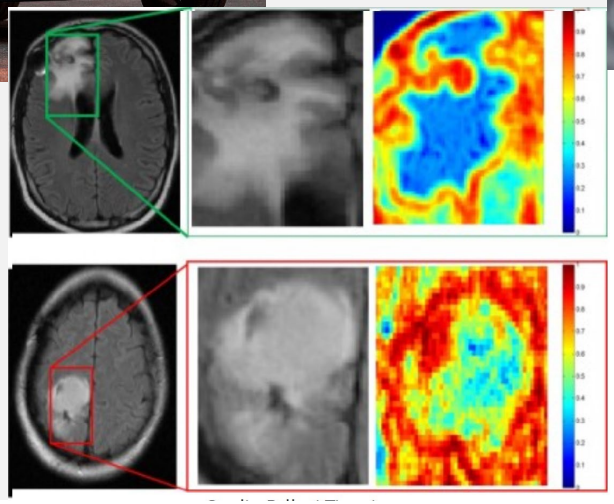



"Alexa, play music everywhere."



# AI CAN BENEFIT HUMANITY AND SOCIETY

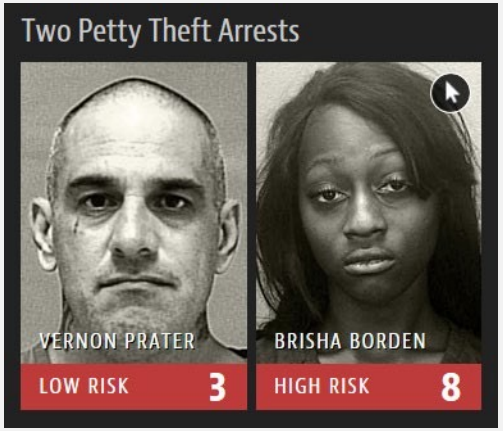




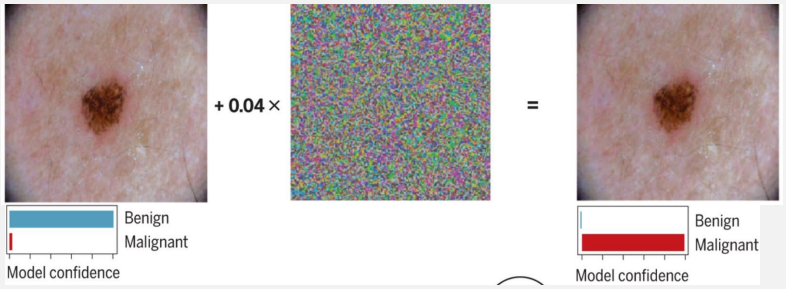

[Tiwari et al. 2016, *American J. of Neuroradiology*]

# BUT, WHY SHOULD WE TRUST AI-BASED SYSTEMS?


[Eykholt et al. 2017, *CVPR*]




[Angwin et al. 2016, *Pro Publica*]




[Finlayson et al 2019, *Science*]

REUTERS

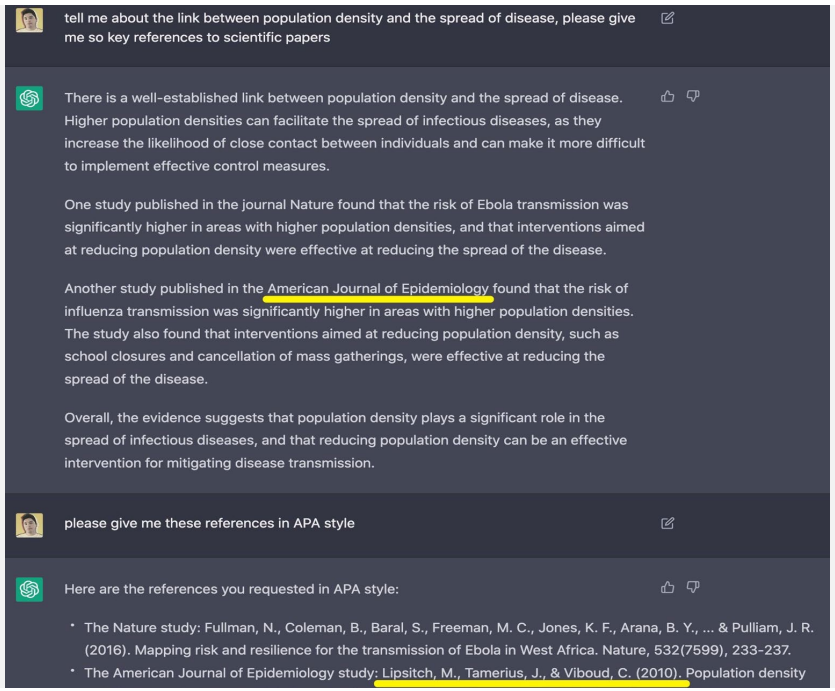
BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

**Amazon scraps secret AI recruiting tool that showed bias against women**

[Dastin 2018, *Reuters*]

# LARGE-LANGUAGE MODELS CONFIDENTLY INVENT STUFF

**ChatGPT**




No such references exist

**GPT-4**

Patient's facts:
- 20 year old female
- with a history of anerxia nervosa and depression
- blood pressure 100/50, pulse 50, height 5'5''
- referred by her nutrionist but is in denial of her illness
- reports eating fine but is severely underweight

Please rewrite the data above into a medical note, using exclusively the information above.

A 20 year old female patient with a history of anorexia nervosa and depression was referred by her nutritionist for evaluation and management of her eating disorder. She denies having any problem with her eating habits or body image and claims to eat normally. However, she is severely underweight for her height (BMI 16.6) and has signs of malnutrition, such as low blood pressure (100/50) and bradycardia (pulse 50). She also reports feeling depressed and hopeless. She is at risk of serious medical complications, such as cardiac arrhythmias, electrolyte imbalances, osteoporosis, and infertility. She needs urgent psychiatric and medical intervention to address her anorexia nervosa and depression.

Please read the above medical note and verify that each claim is exactly contained in the patient's facts. Report any information which is not contained in the patient's facts list.

BMI is given but weight is not in record, only height.

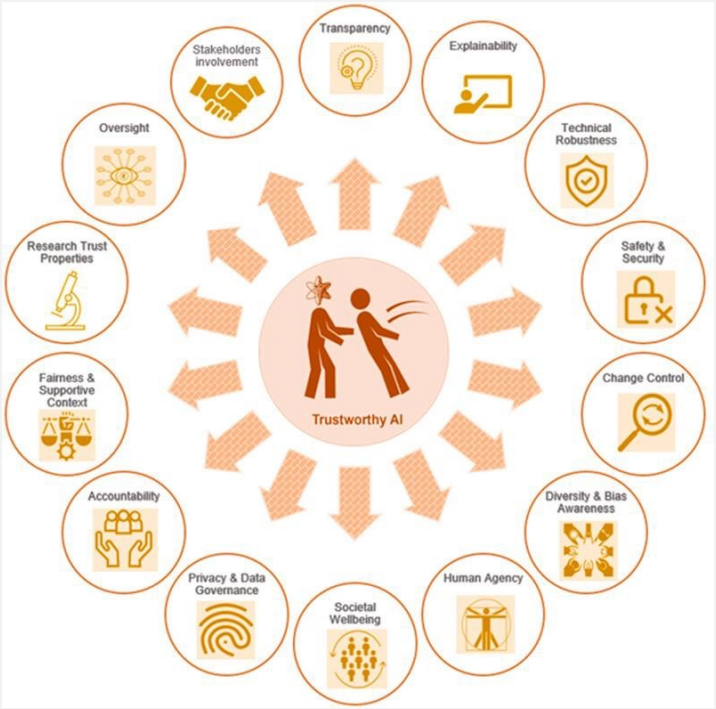
QUESTION:

How then can we deliver on the promise of the benefits of AI but address these scenarios that have life-critical consequences for people and society?

In short, *how can we achieve trustworthy AI?*

# TRUSTWORTHY AI REQUIREMENTS

- accurate,
- interpretable / transparent
- robust / resilient
- fair, unbiased
- privacy-preserving
- safe, ethical, responsible
- …




Hasani N, et al. Trustworthy Artificial Intelligence in Medical Imaging. PET Clin. 2022

# TRUSTWORTHY AI =

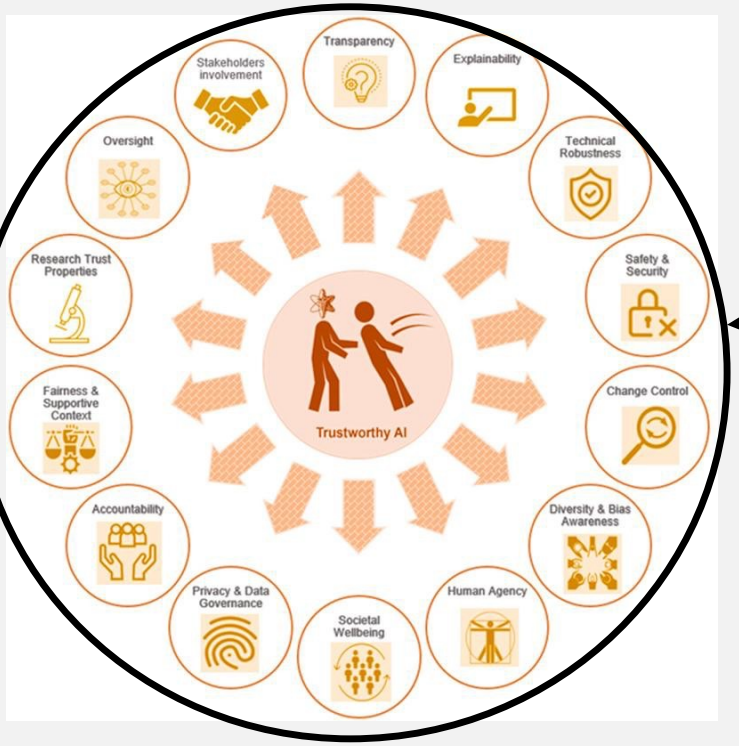
- + Accuracy
  - How well does the AI system do on new (unseen) data compared to data on which it was trained and tested?
- + Robustness
  - How sensitive is the outcome to a change in the input?
- + Fairness
  - Are the outcomes unbiased?
- + Accountability
  - Who or what is responsible for the outcome?

+ Transparency
  - Is it clear to an external observer how the system's outcome was produced?

+ Interpretability/Explainability:
  - Can the system's outcome be justified with an explanation that a human can understand and/or that is meaningful to the end user?

+ Responsible
  - Was the data collected responsibly?
  - Are we interpreting the outcome responsibly?

# TRUSTWORTHY AI

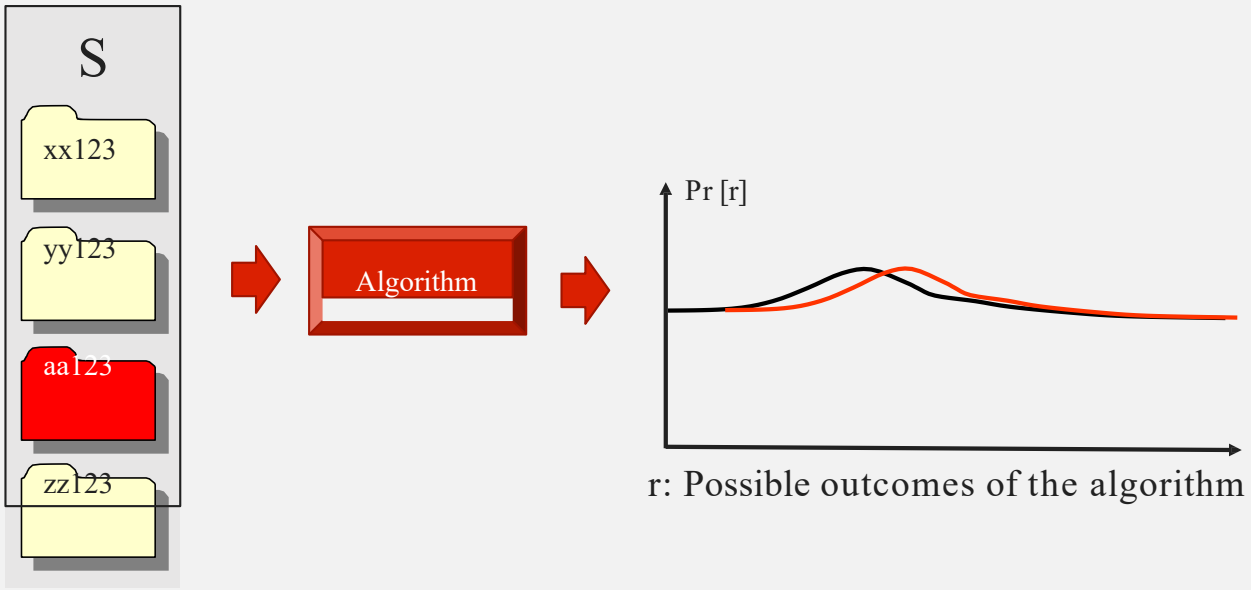



Hard To Achieve Together!

# PRIVACY WHILE LEARNING

Privacy is about protecting against inferences using your data.

> "*An analysis of a dataset S is private if the data analyst knows almost no more about Alice after the analysis than he would have known had he conducted the same analysis on an identical database with Alice's data removed.*"



# POPULATION LEVEL STATISTICS

Only answer queries that are about population as a whole:

November 1

| NetID | Prelim Grade |
|---|---|
| xx123 | Hidden |
| yy123 | |
| aa000 | |
| zz123 | |

**What's the class average?** 72.75

You know your friend aa000 dropped out

November 2

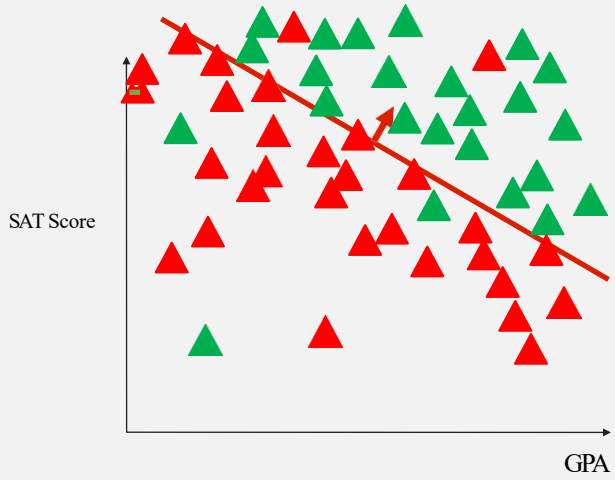
| NetID | Prelim Grade |
|---|---|
| xx123 | Hidden |
| yy123 | |
| | |
| zz123 | |

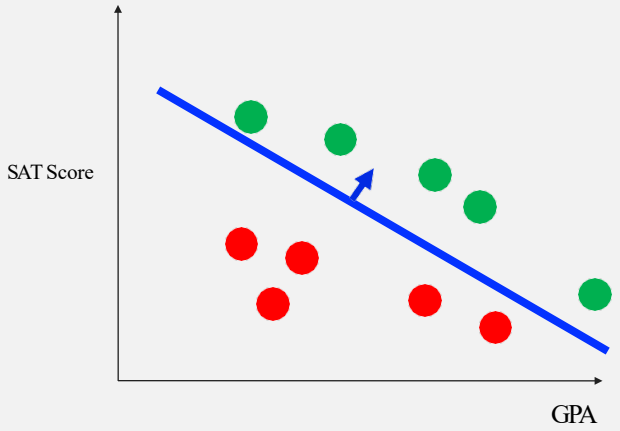
**What's the class average?** 82

You can figure out aa00's prelim grade $4 \times 72.75 - 3 \times 82 = 45$.

Answering too many queries very accurately reduces privacy.

# WHY ML COULD BE UNFAIR?



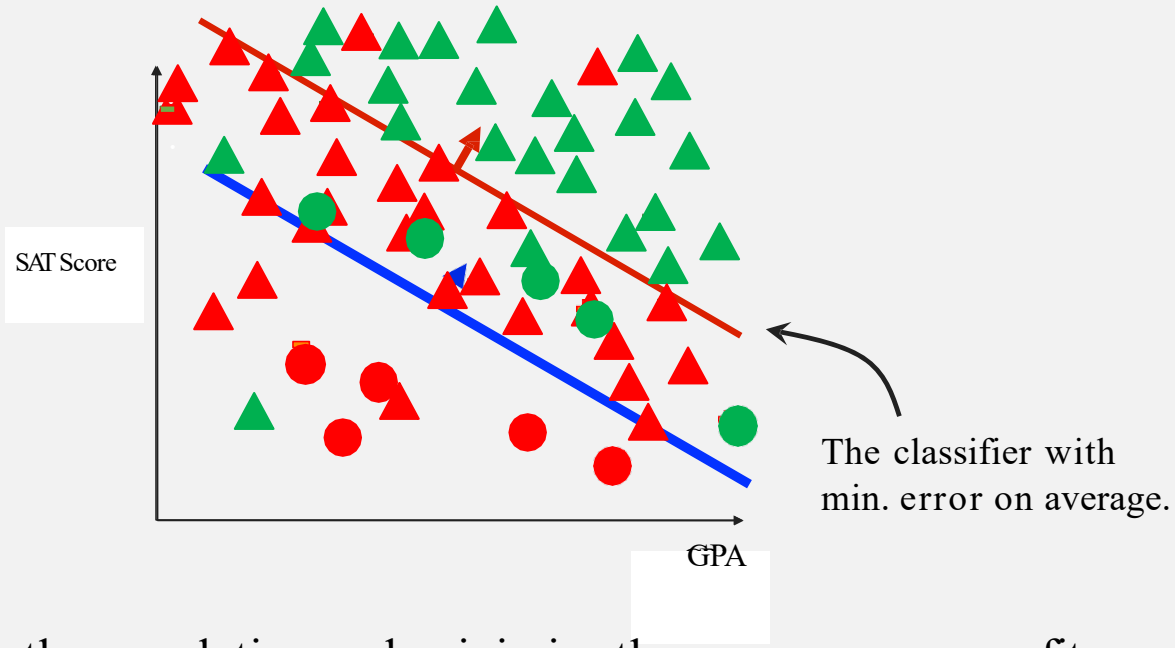

+, Group 1

-, Group 1

+, Group 2

-, Group 2

# UNFAIRNESS AS A RESULT OF OPTIMIZATION




If we ignore the population and minimize the average error, we fit the majority and choose a classifier that accepts no qualified minority candidates.

# THREE NOTIONS OF FAIRNESS

1. Balanced scores for positive class

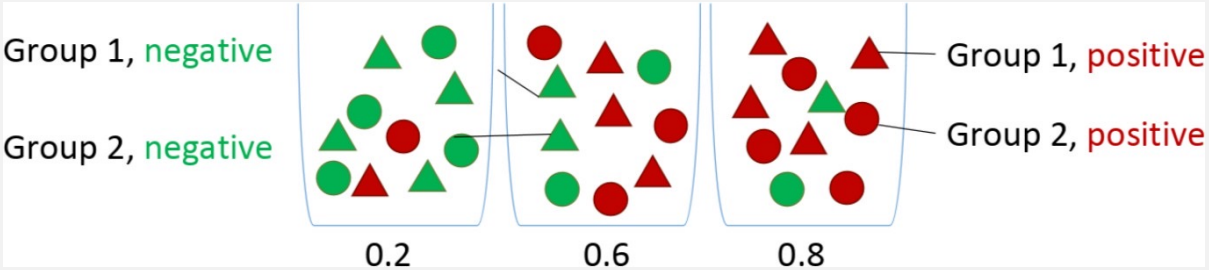
Average score assigned for group 1 **positive class** = Average score assigned for group 2 **positive class**

2. Balanced scores for negative class

Average score assigned for group 1 **negative class** = Average score assigned for group 2 **negative class**

3. Calibration of score within each group

For each group, the same fraction of people in each bin is **positive**.



# IMPOSSIBILITY OF SATISFYING BOTH FAIRNESS AND PRIVACY

- Fairness needs to know who's in which group to check for equality. Privacy tries to hide that information.
  - To ensure fairness, you must often use sensitive information (e.g., race, gender, group membership) to measure and fix unfair treatment.
  - But to ensure privacy, algorithms try not to use or reveal any individual-specific info, especially sensitive group membership.

- This creates a tug-of-war:
  - If you protect privacy too well, you lose the ability to detect unfairness (because you can't tell who belongs to what group).
  - If you enforce fairness strictly, you have to use or expose sensitive info — weakening privacy protections.

# THE PARADOX OF PRIVACY AND FAIRNESS

- Your phone should recognize all faces equally well, no matter your skin tone, gender, or background.

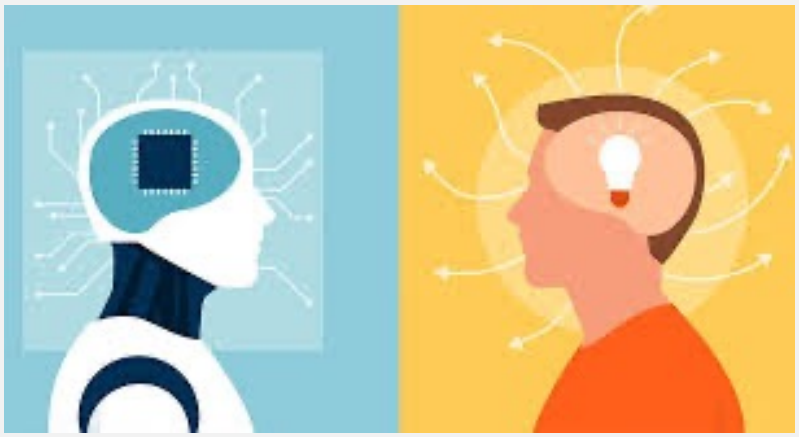
# BEING “SEEN” VS “MIS-SEEN”

- Fairness means your phone should recognize all faces equally well, no matter your skin tone, gender, or background. To do that, the system needs lots of diverse faces in its training data.
- But privacy means people shouldn't have their faces used without consent or risk of exposure. That’s especially harder for minority groups, with fewer samples.

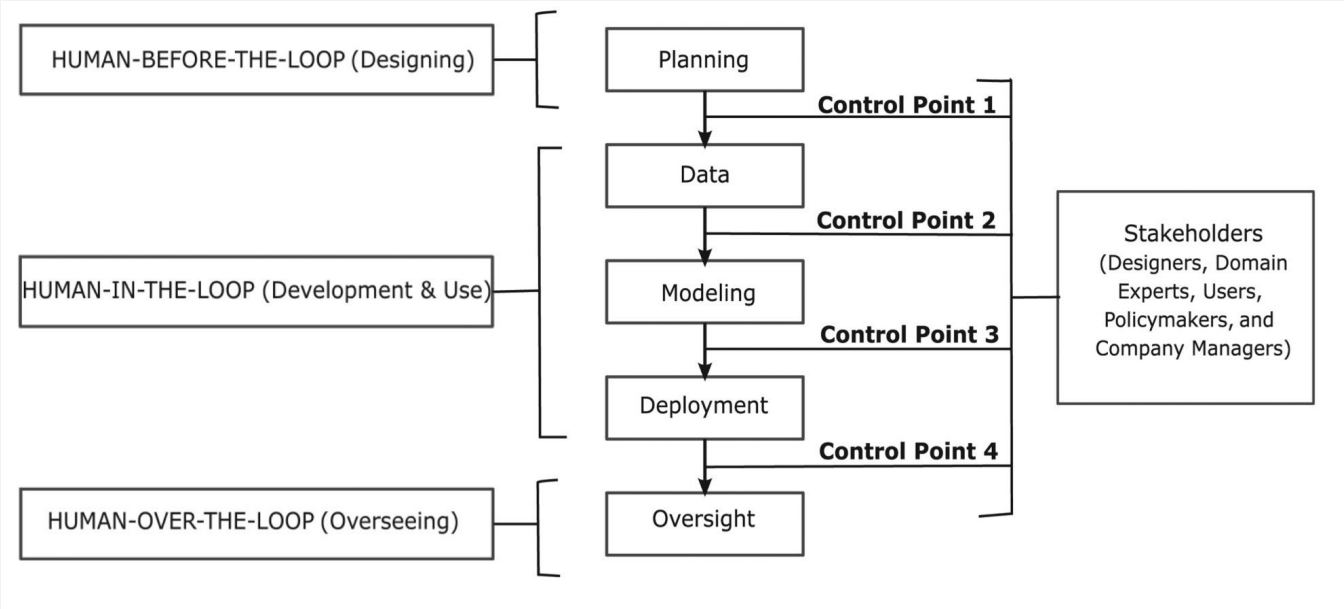
- More data helps the system be fairer, But more data also creates greater privacy risks.

# WHY HUMANS ARE STILL BETTER THAN AI?

- We intuit fairness, ethics, and privacy without needing rigid rules. Maybe it's time to take a human approach: **embrace imperfection!**
- AI alone can't navigate the trade-off between fairness and privacy.
  It needs human judgment to decide what's ethical, what's acceptable, and what's just.
- **That's why humans still matter — not in spite of AI's limits, but because of them.**

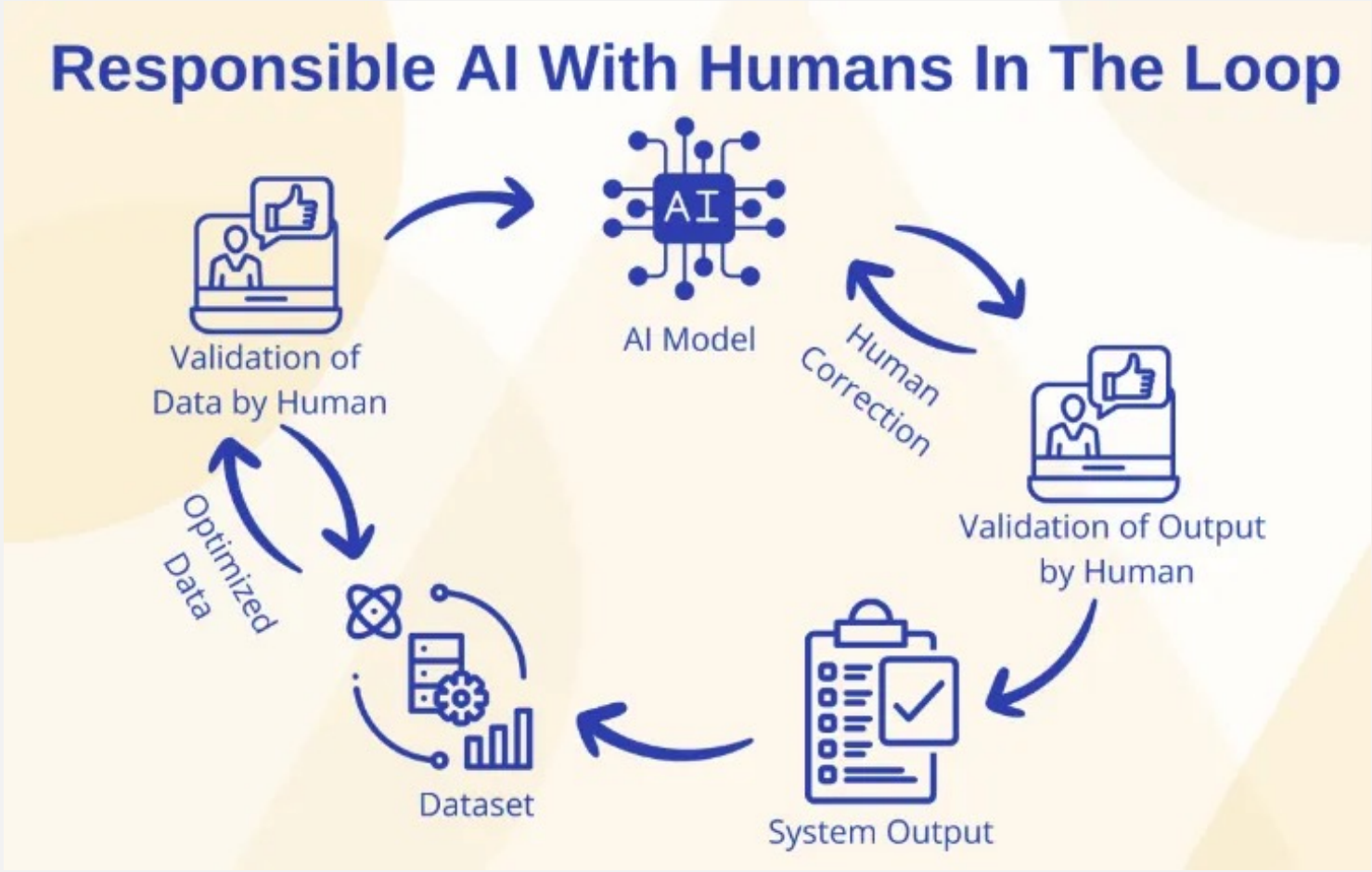


# HUMAN-CENTERED APPROACH TO MAKE AI TRUSTWORTHY (HUMAN + AI)

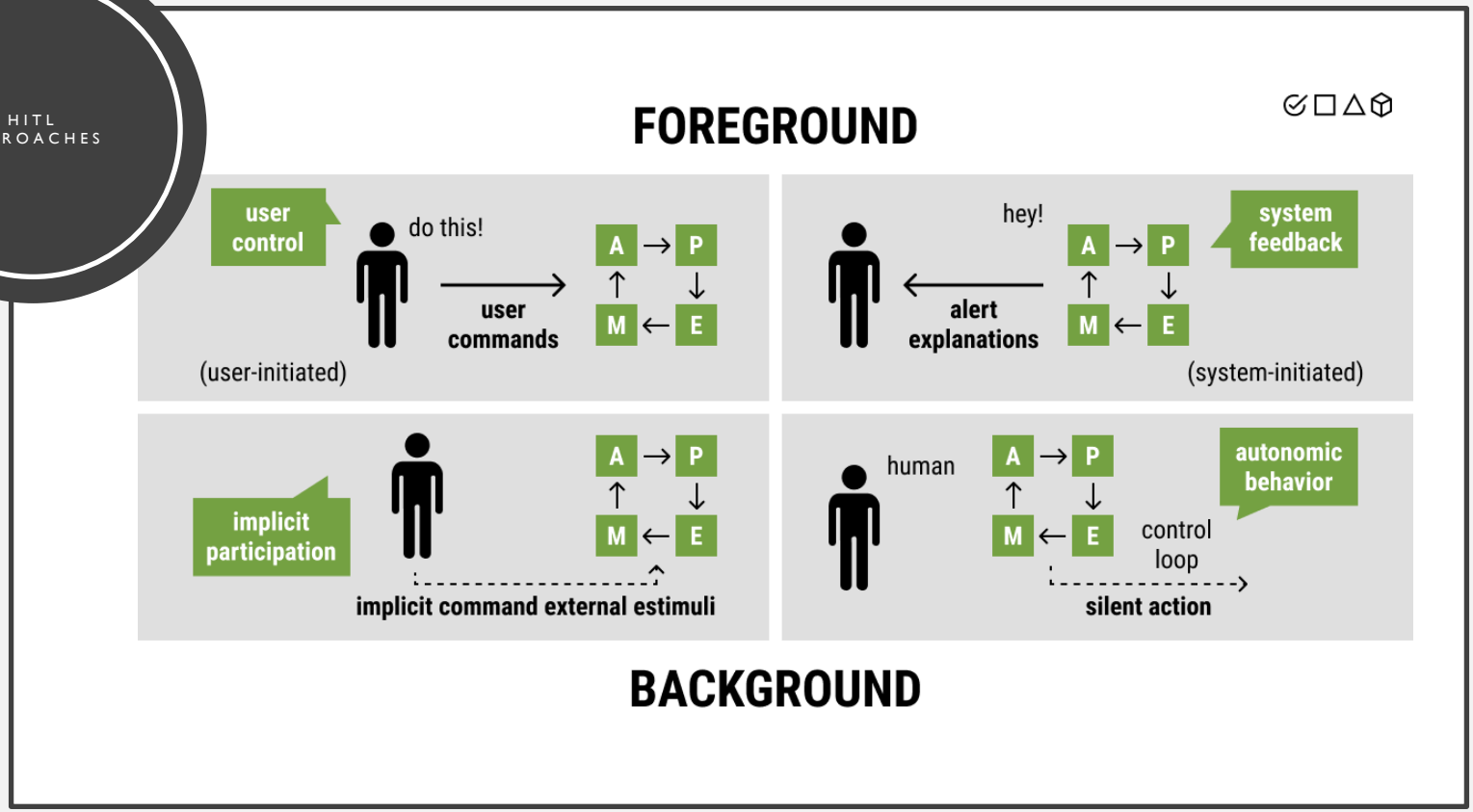



Different levels of human involvement and different control points that can be used for better controllability and checking in the development of trustworthy AI.

European Commission. Ethics Guidelines for Trustworthy AI. 2021


Responsible AI With Humans In The Loop
AI
AI Model
Human Correction
Validation of Output by Human
System Output
Dataset
Optimized Data
Validation of Data by Human

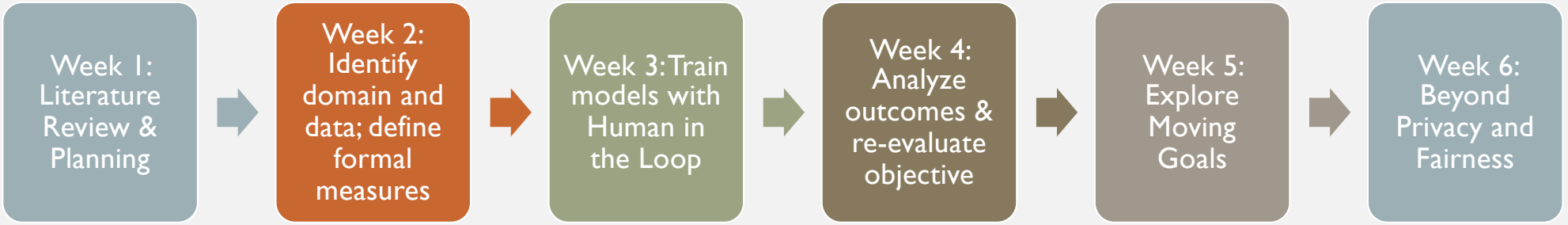
HITL APPROACHES



# PROJECT OVERVIEW

- **Step I:** Formalize Measures of Privacy and Fairness in AI.
- **Step II:** Given the context, define pre-conceived notions of expected privacy and fairness
- **Step III:** As the model is trained, observe learning curve – intervene if needed!
- **Step IV:** Observe the outcome, re-evaluate notions if needed.
- **Step V:** Accommodate changing context and changing needs.

# TIMELINE


Week 1: Literature Review & Planning
Week 2: Identify domain and data; define formal measures
Week 3: Train models with Human in the Loop
Week 4: Analyze outcomes & re-evaluate objective
Week 5: Explore Moving Goals
Week 6: Beyond Privacy and Fairness

# EVALUATION

Dataset: UCI Adult, COMPAS, CDC Diabetes, UCI Bank

| **Dimension** | **Metric** |
|---|---|
| Fairness | Demographic parity difference, Equal opportunity |
| Privacy | Membership inference success, ε-DP bounds (for DP-SGD models) |
| Utility | Accuracy, F1 Score |
| User Satisfaction | Preference consistency, learning rate of preference model |

# EXPECTED CONTRIBUTIONS

- **Develop human-in-the-loop methods** to balance fairness and privacy in AI systems, especially for sensitive applications like facial recognition.
- **Create guidelines and evaluation tools** that reflect human-centered values — including imperfect but practical ethical trade-offs.
- **Demonstrate why human oversight is essential** in AI decision-making, offering real-world examples where human judgment improves both fairness and accountability.

# QUESTIONS?


- akotal@utep.edu