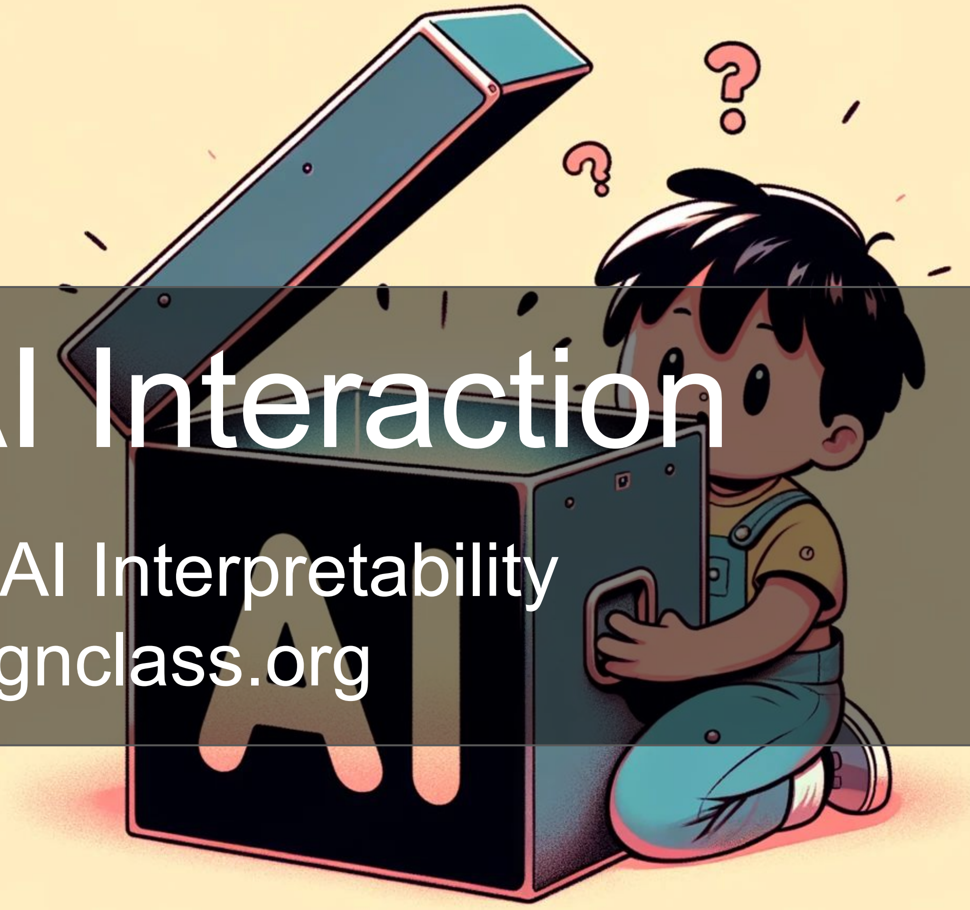
Human AI Interaction

Lecture 16: AI Interpretability

aidesignclass.org



Learning goals today

- What is interpretability and when is it helpful?
- If an algorithm can be interpreted, does it improve:
  - Decision-making?
  - Trust?
- Counterfactuals: what would make this not-true?
  - Does this help decision-making?
- What should you do as a designer?
- Project info
- Thursday: Guest lecture with Dr Ding Wang, on “responsible data”
- Reminder: if you haven’t asked for APIkeys, you should!

Bad news: there is no consensus definition of interpretability

But the general hypothesis is: “If you can follow the reasoning of an AI system, then you can know if its answers are correct”

E.g. Decision-tree “AI” which tells you whether you should walk to school:

```
IF weather is bad, THEN don't walk (take the bus)
```

```
IF weather is good && School is close, THEN walk. ELSE,
don't walk
```

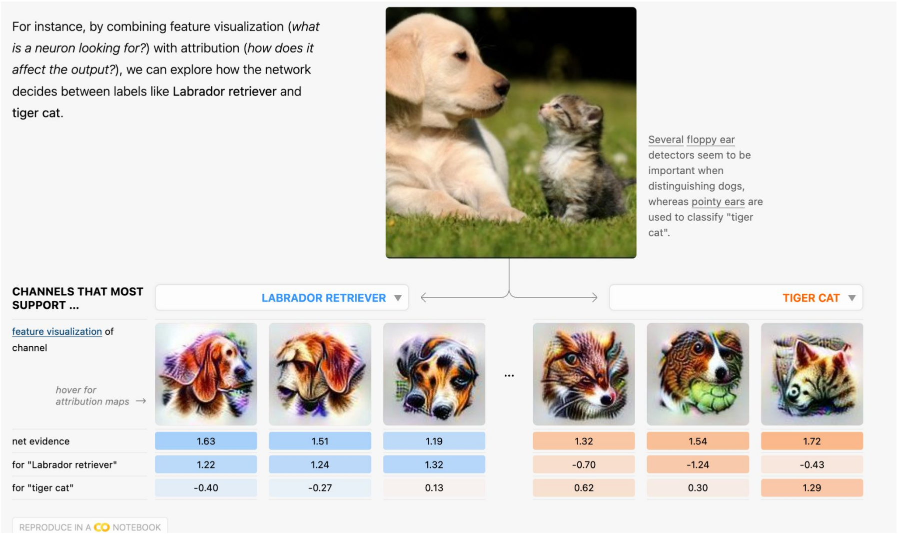
Other examples of interpretability

[Attribution and attention]

This page does three things: visualize features, show where they are detected, and show net evidence for the feature

Your task#1:

- Play with a few examples on this page
- Do you find these neural networks more understandable?



Task #2: can you predict what features will be used?

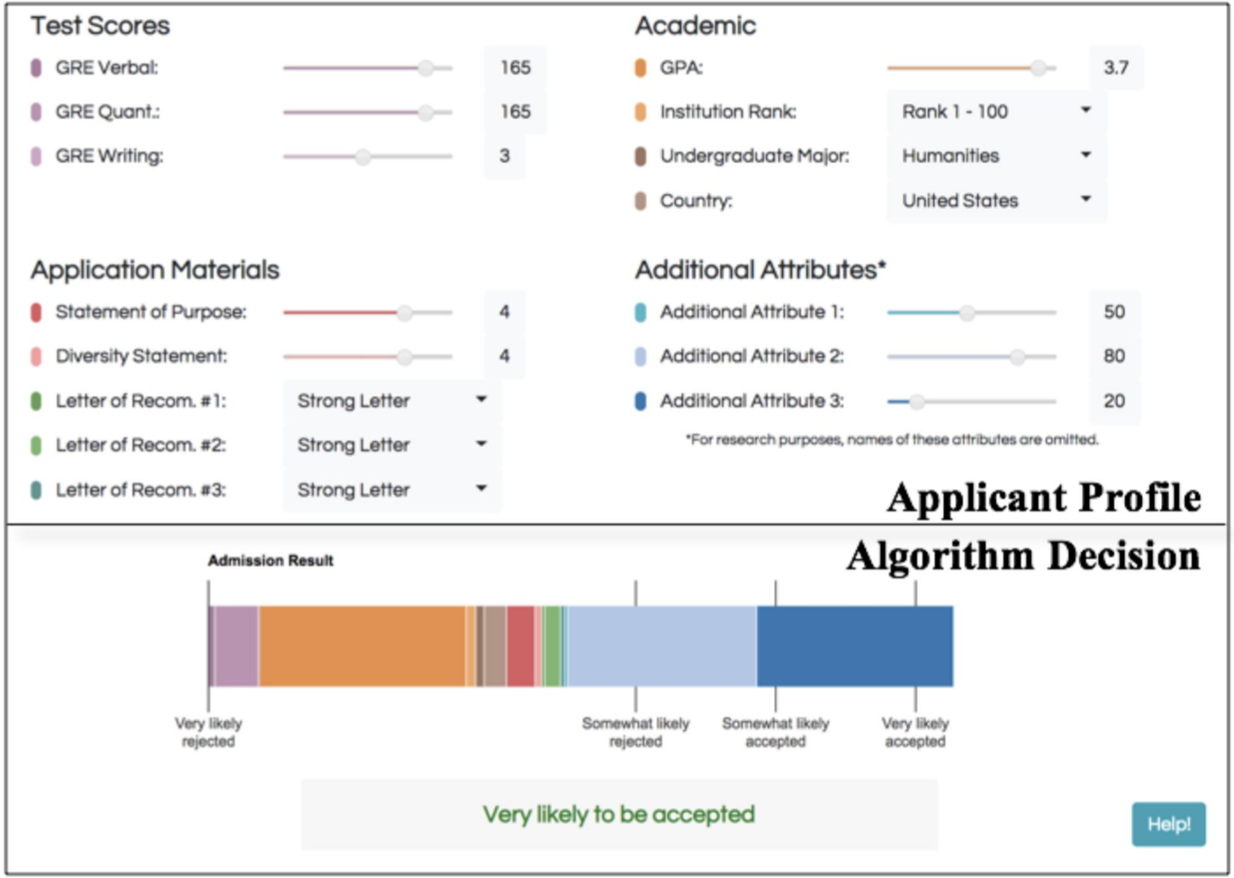
… to identity this bird?

Image from [Wikipedia]

Interpretability by breaking open the black box

Imagine you had a system that determined if a student was admitted into grad school

(Images on the next 5 slides from [this paper](#) by Hao-Fei Cheng 1, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, Haiyi Zhu)



Interpretability by breaking open the black box

Imagine you had a system that determined if a student was admitted into grad school

- Many ways to do interpretability: let's first see black vs. white box

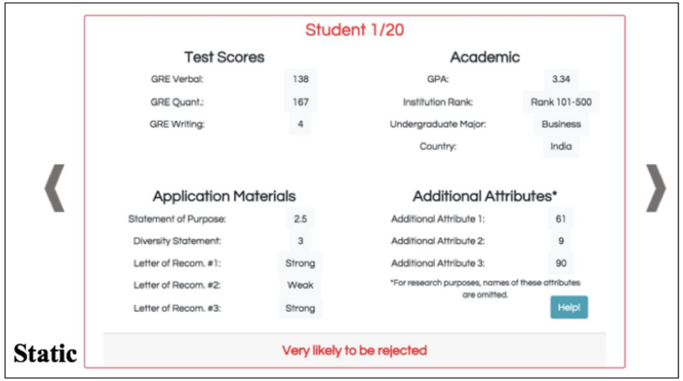
Black-box

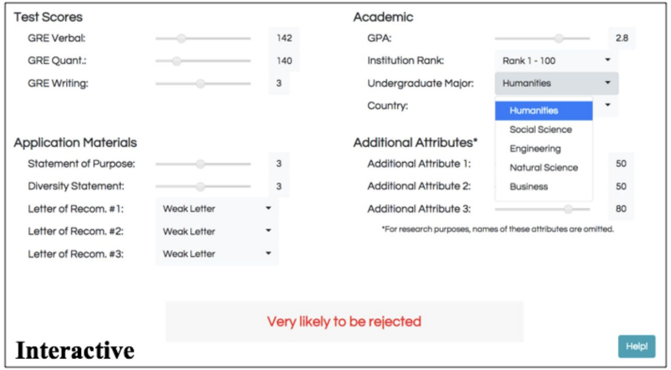
White-box

Interpretability by breaking open the black box

Imagine you had a system that determined if a student was admitted into grad school

- Static vs. dynamic explanations





Interpretability by breaking open the black box

Imagine you had a system that determined if a student was admitted into grad school

Predict results!

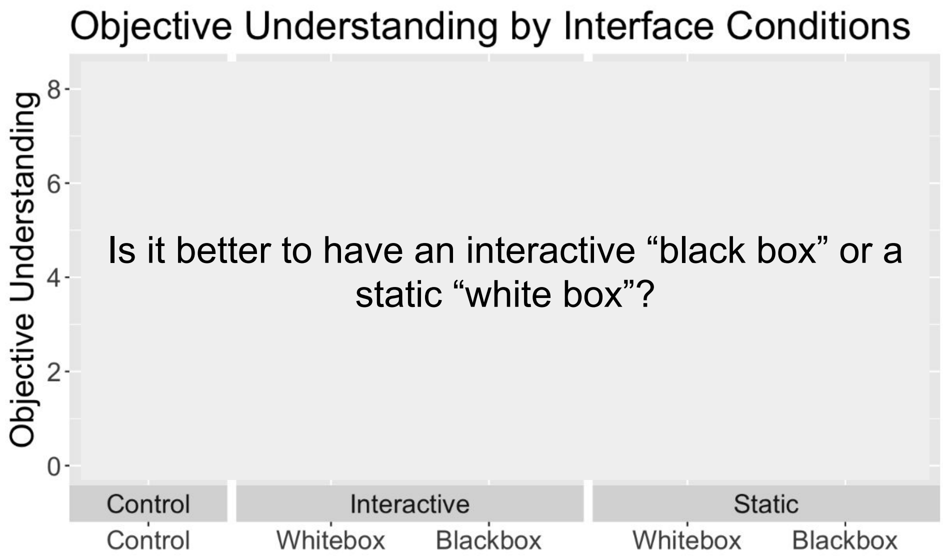


Figure 2: Participants’ objective understanding of the algorithms by interface conditions. Error bars represent 95% confidence intervals.

Interpretability: design implications

- Interactive “whitebox” models are most understandable
- When you can’t open the blackbox (i.e. reveal how it works), interactivity has nearly the same benefit.

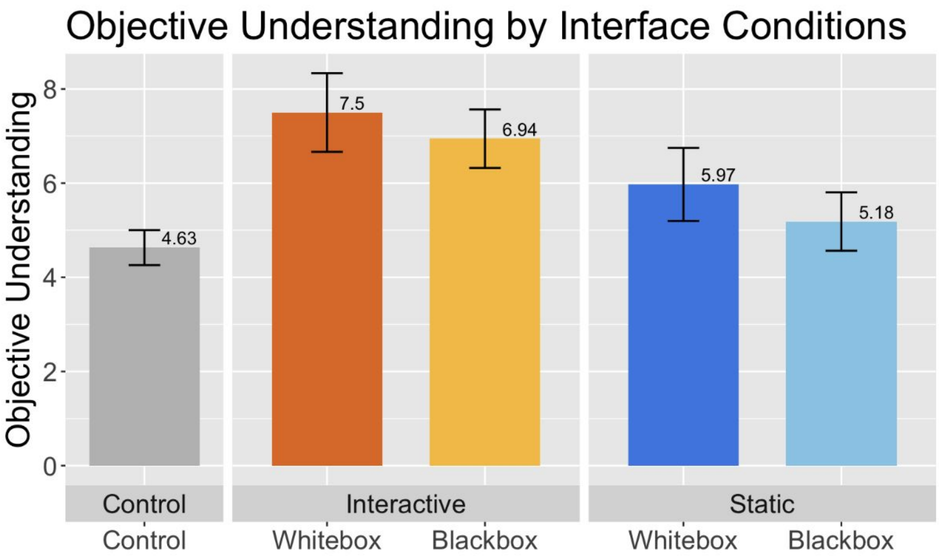
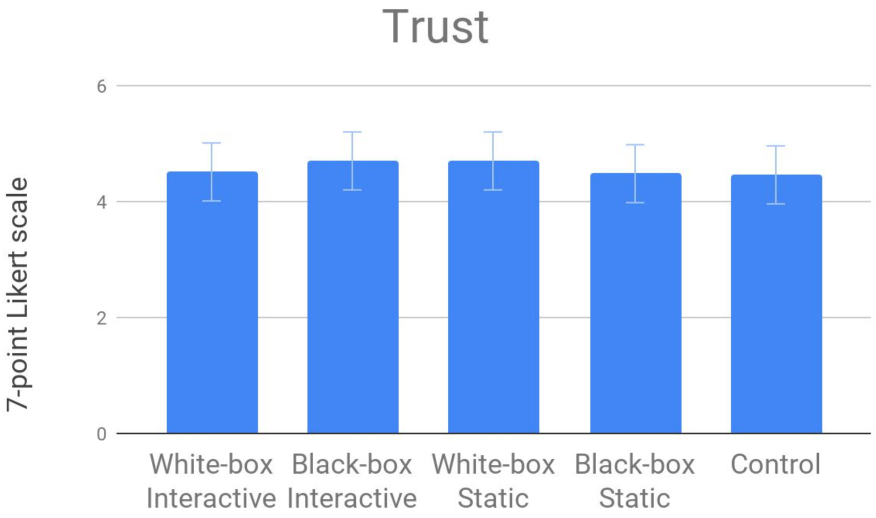


Figure 2: Participants’ objective understanding of the algorithms by interface conditions. Error bars represent 95% confidence intervals.

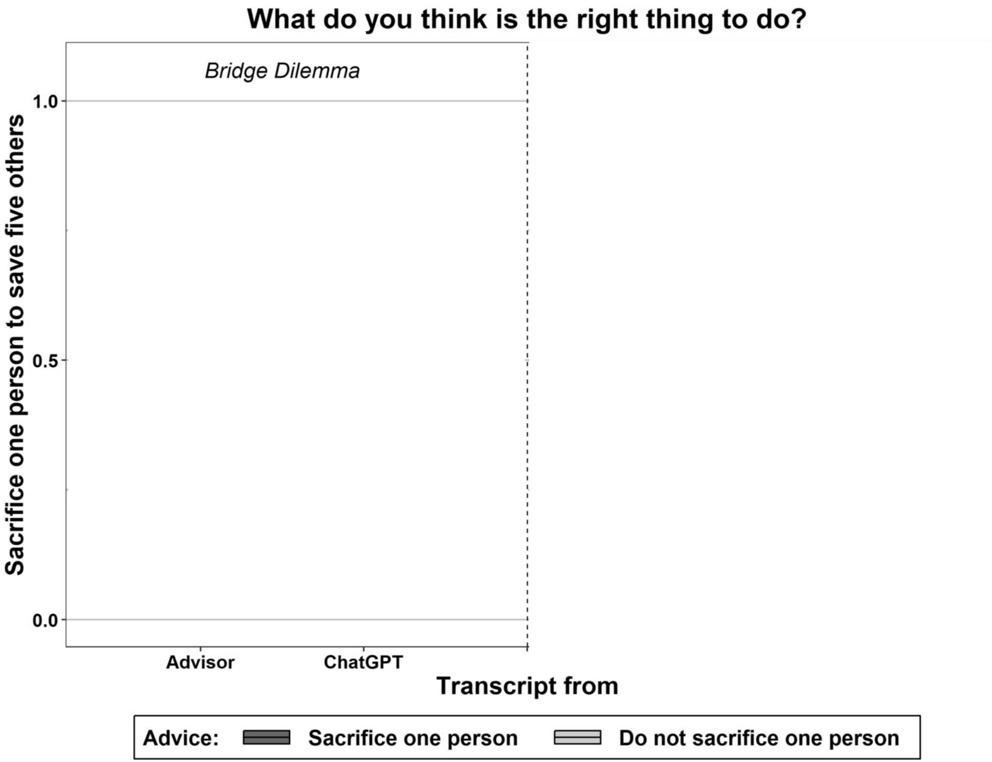
Trust

Even when participants (don't) understand the algorithm, they may still trust it

(chart from Haiyi Zhu)



Trust: would you trust an AI anyway?

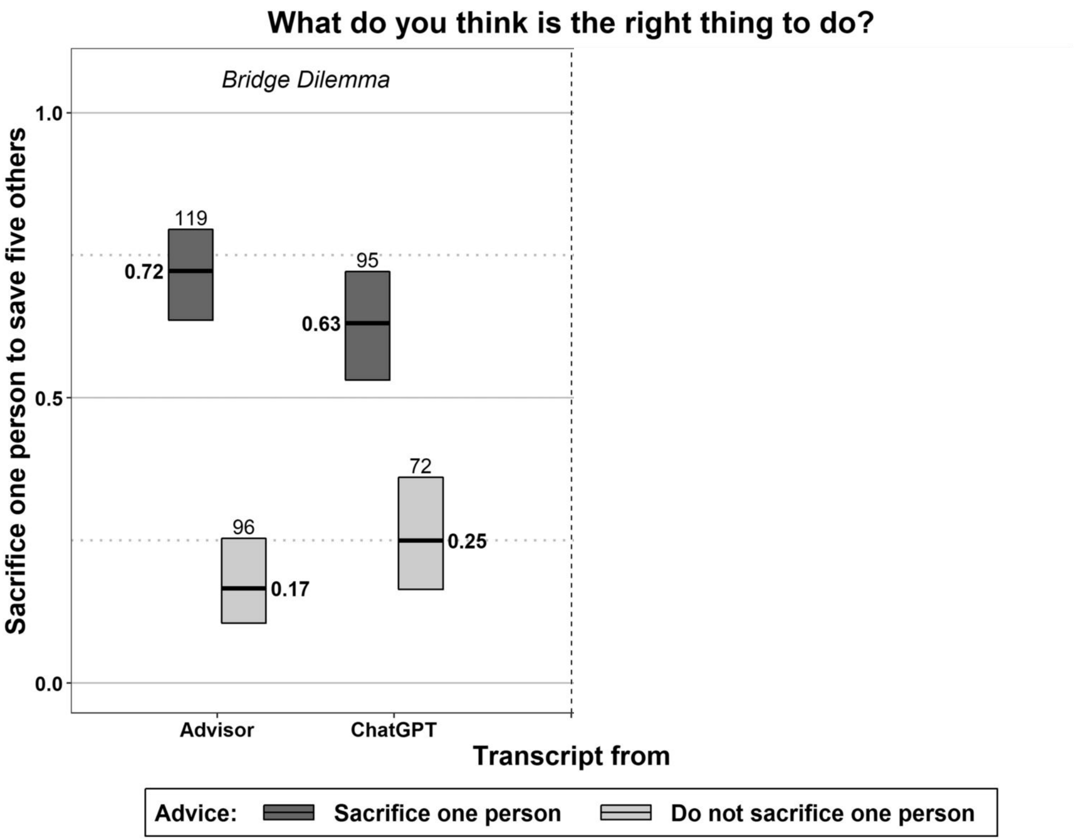


Trust: it gets worse

People take advice on ethical issues from AI, even when the AI is inconsistent!

- Merely telling people "hey this comes from a probabilistic AI system" isn't enough to discount its dubious advice.

Chart from this paper.



Counterfactuals

Counterfactual: “That which is not the case”

Images from [Polyjuice]

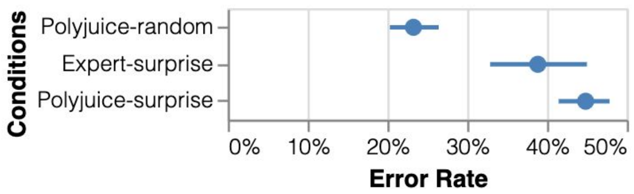


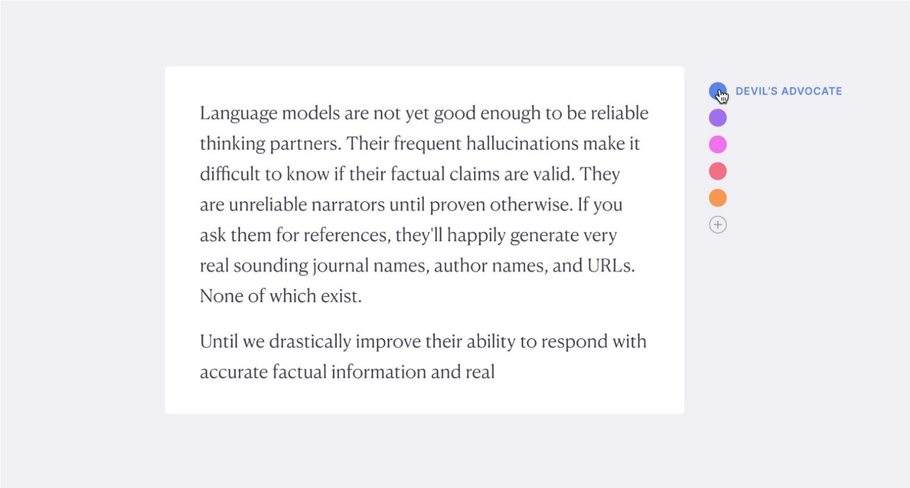
Figure 4: Simulation error rates per condition (higher the better). POLYJUICE-*surprise* has the highest error rate, indicating these counterfactuals would add the most information to users if displayed.

Project info

Teams up to three

Default project: “create Daemons people can trust”

- Starter code is provided
- Must implement prompts and ask at least 5 users questions around trust (remember: benevolence, ability, integrity)



Project info

Teams up to three

Alternative project

- You pick what you want to do
- Allowed to reuse a project you are working on
- Requirements: must involve some implementation, some measurement of a concept of interest
- **MUST GET PRIOR APPROVAL**
- Rewarded for “principled risk taking”