# Human Al 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?

For instance, by combining feature visualization (what is a neuron looking for?) with attribution (how does it affect the output?), we can explore how the network decides between labels like Labrador retriever and tiger cat.



Several floppy ear detectors seem to be important when distinguishing dogs, whereas pointy ears are used to classify "tiger cat".

CHANNELS THAT MOST SUPPORT								
$\frac{\text{feature visualization of}}{\text{channel}}$ hover for $\frac{1}{\text{attribution maps}} \rightarrow$								
net evidence	1.63	1.51	1.19		1.32	1.54	1.72	
for "Labrador retriever"	1.22	1.24	1.32		-0.70	-1.24	-0.43	
for "tiger cat"	-0.40	-0.27	0.13		0.62	0.30	1.29	

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

... to identity this bird?



Image from Wikipedia

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

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

Test Scores			Academic		
GRE Verbal:		165	BPA:		3.7
GRE Quant.:		165	Institution Rank:	Rank 1 - 100	-
GRE Writing:		3	Undergraduate Major:	Humanities	•
			Country:	United States	•
Application Materials	6		Additional Attribut	es*	
Statement of Purpose:		4	Additional Attribute 1:		50
Diversity Statement:		- 4	Additional Attribute 2:		80
Letter of Recom. #1:	Strong Letter	*	Additional Attribute 3:	-0	20
Letter of Recom. #2:	Strong Letter	*	*For research purposes,	names of these attributes are	omitted.
Letter of Recom. #3:	Strong Letter	•		Applicar	nt Profil
Admissic	on Result			Algorithm	Decisio
Very likely			Somewhat likely Somewhat likely	Very likely	
rejected			rejected accepted	accepted	

Very likely to be accepted

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



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

• Static vs. dynamic explanations



Test Scores			Academic		
GRE Verbal:		142	GPA:		2.8
GRE Quant .:		140	Institution Rank:	Rank 1 - 100	-
GRE Writing:		3	Undergraduate Major:	Humanities	*
			Country:	Humanities	-
Application Materials Statement of Purpose:		3	Additional Attributes* Additional Attribute 1:	Social Science Engineering Natural Science	50
Letter of Recom #1:	Weak Letter	•	Additional Attribute 3:	Duanteas	80
Letter of Recom. #2: Letter of Recom. #3:	Weak Letter Weak Letter	•	*For research purposes, nam	es of these attributes are on	nitted.
Interactive		Very like	ly to be rejected		Help

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.

**Objective Understanding by Interface Conditions** 



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)



7-point Likert scale

## Trust: would you trust an AI anyway?

What do you think is the right thing to do?



# 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.



#### What do you think is the right thing to do?

#### 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.

\Lambda Origi	nal x	f(x)	G Select
	It is great for kids.	) +	for use cases
🕒 Polyje	uice generates $\hat{x}_i$	$f(\hat{x}_i)$	Training
delete	lt is great <mark>for kids</mark> .	+	Evaluation
negation	It is not great for kids. It is great for <mark>kids</mark> →no one	- +	Explanation D
lexical	It is great for <mark>kids</mark> →adults. It is <mark>great</mark> →scary for kids.	+++++++++++++++++++++++++++++++++++++++	It is great for kids

## 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)

Language models are not yet good enough to be reliable thinking partners. Their frequent hallucinations make it difficult to know if their factual claims are valid. They are unreliable narrators until proven otherwise. If you ask them for references, they'll happily generate very real sounding journal names, author names, and URLs. None of which exist.

Until we drastically improve their ability to respond with accurate factual information and real

#### 

•

# 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"

Language models are not yet good enough to be reliable thinking partners. Their frequent hallucinations make it difficult to know if their factual claims are valid. They are unreliable narrators until proven otherwise. If you ask them for references, they'll happily generate very real sounding journal names, author names, and URLs. None of which exist.

Until we drastically improve their ability to respond with accurate factual information and real

#### 

•