

# What's Opaque to Whom - and Why?

---

Anders Søgaard

coASTal



‘people working at the  
cutting edge of AI [...] consider us, at best,  
lunchtime  
entertainment’

*Cappelen & Dever (2022)*



# DNNs and Opacity

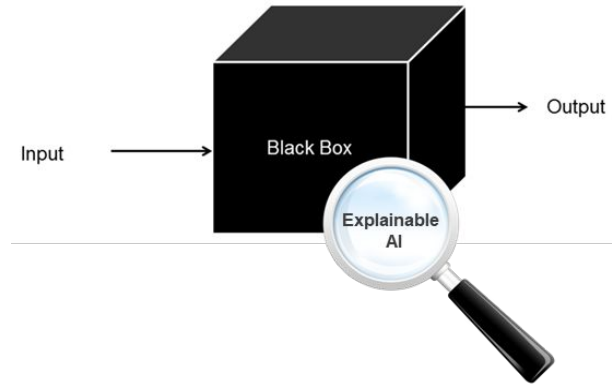
---

### Standard view

DNNs are black boxes in need of explanations.

### Alternative view

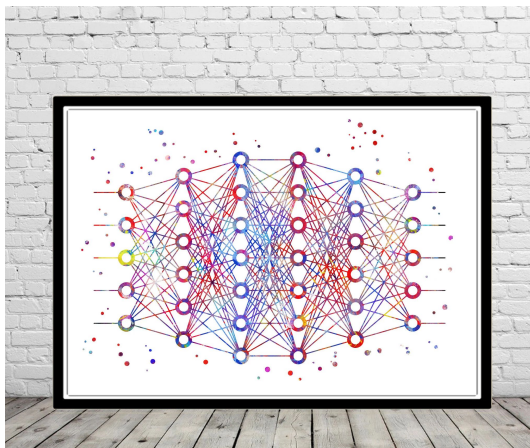
DNNs are *white* boxes *and* explanations.



# DNNs = White-Box Explanations?

## DNNs are white boxes

There is nothing secret about the inner workings of DNNs. You can print everything out. They are, as such, white boxes. Just very big ones.

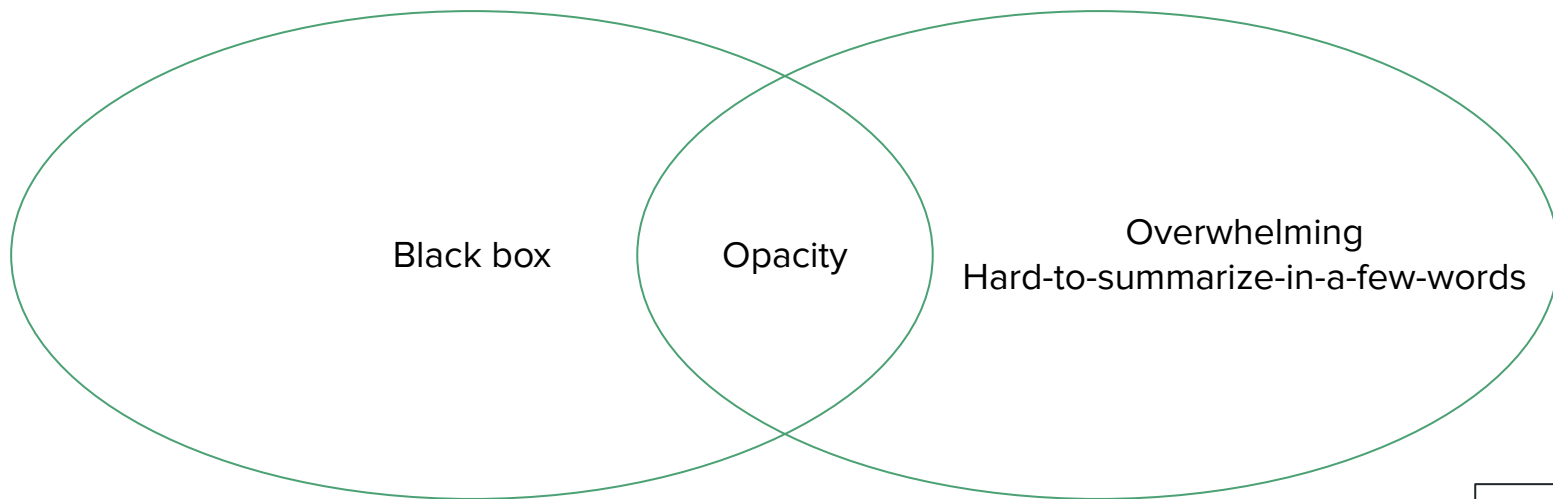


## DNNs are explanations

“who can explain a phenomenon understands it, in the sense that [they] can predict it” [...] “the very essence of explanation is generalization”

*Kenneth Craik*

DNNs provide us with prediction and generalization. A DNN explains Y in terms of X. Its explanation may be **false**, i.e., fail to generalize properly, or **overwhelming**, rendering it completely useless for some purposes, including teaching.



**Terminology**



**What I'd say**

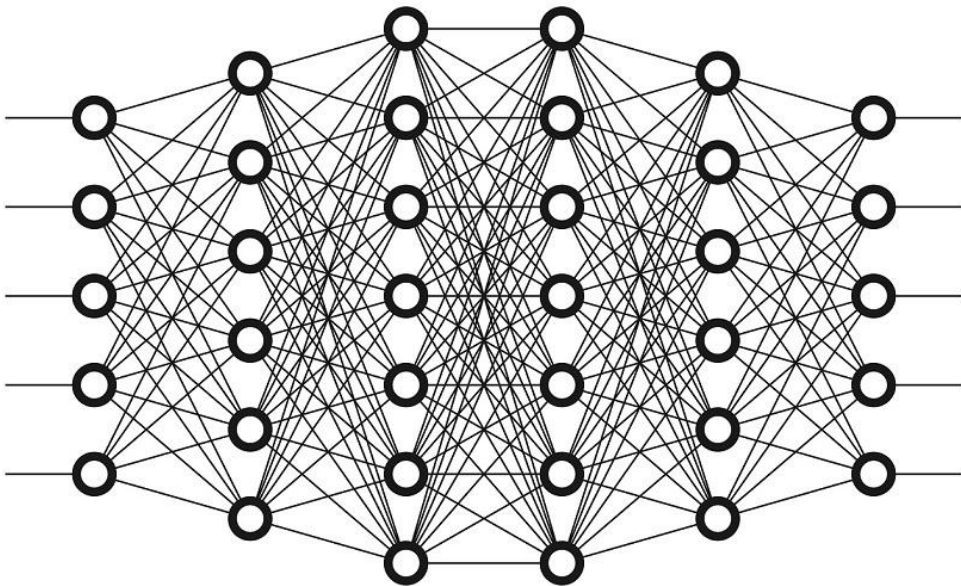
### **opacity (n.)**

1550s, "darkness of meaning, obscurity," from French *opacité*, from Latin *opacitatem* (nominative *opacitas*) "shade, shadiness," from *opacus* "shaded, dark, opaque," a word of unknown origin. The literal sense "condition of being impervious to light; quality of a body which renders it impervious to rays of light" in English is recorded from 1630s.

# Two Kinds of Opacity

1. Why do neural networks predict what they do, given the input data?
2. How do neural networks settle on the functions they settle on, given the training data?

-  Inference-opacity
-  Training-opacity



Contenders	Citations
Number of parameters	"it is generally difficult to interpret or explain how or why a DL algorithm arrives at a particular decision, given that they are built on numerous hidden layers and millions of neurons, which makes them opaque."
Nonlinearities	DNNs exhibit "non-linear structure which makes them opaque."
Continuity	"discrete representations have the advantage of being readily interpretable"
Lack of grounding	DNNs do not "work in accordance with the ways humans themselves assign meaning to the reality that surrounds them."
Lost training history	"learning algorithms are even more opaque because they do not rely on pre-specified instructions, but on evolving weights and networks of connections that get refined with each additional data point"

Inference-opacity

Training-opacity

Size	Nonlinearities	Continuity	Instrumentality	Incrementality
<div></div> <div></div>				




# Summary

DNNs are white-box explanations, but very large. This makes XAI a summarization problem, i.e., an abtractiveness-faithfulness trade-off. Or, equivalently, a bias-variance trade-off.

$$\arg \min_{\theta^a} \sum_{i \leq n} \ell(y_i^a, y_i^o) + \lambda \|\theta^a\|^0$$

... which, in my view, suggests a continuous, instrumentalist XAI, providing explanations at different granularities.

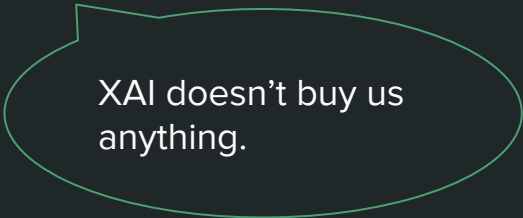




XAI gives us unique opportunities!

# Why XAI for Science?

---



XAI doesn't buy us anything.

# Updating our language

## From recent article:

‘Models developed using machine learning are increasingly prevalent in scientific research. At the same time, these models are notoriously opaque. Explainable AI aims to mitigate the impact of opacity by rendering opaque models transparent. More than being just the solution to a problem, however, Explainable AI can also play an invaluable role in scientific exploration.’

## Our version:

‘Models developed using machine learning **provide potential explanations in** scientific research. ~~At the same time, these models are notoriously opaque.~~ Explainable AI aims to **summarize these often overwhelmingly complex explanations.** More than being just the solution to a problem, however, Explainable AI can also play an invaluable role in scientific exploration.’



Two motivations

## XAI in Science

- a) Extracting hypotheses
- b) Verifying our models are  
*right for the right reasons*

# Extracting Hypotheses

## Saliency Maps (inference-opacity)

Explanations are distributions over input features, e.g., reflecting a local, linear approximation.



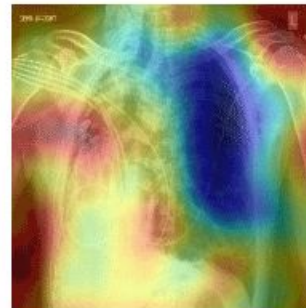
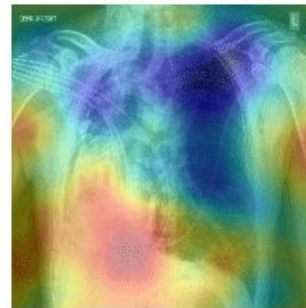
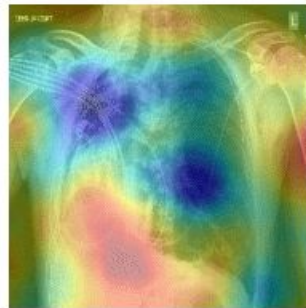
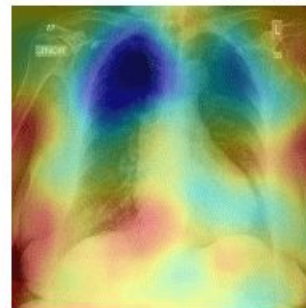
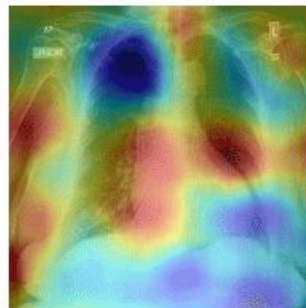
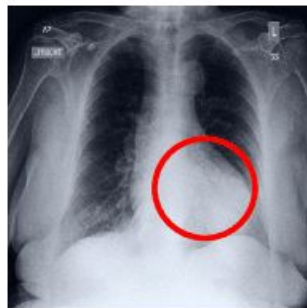
Low recall.

## Influence Functions (training-opacity)

Influence estimates are distributions over training data.



Low precision. Sensitive to initialization.



# Inherent Limitations

We still have low recall, because:

1. Post-hoc XAIs cannot explain in terms of configurations of features.
2. Post-hoc XAIs cannot explain in terms of accumulated statistics.
3. Post-hoc XAIs cannot explain in terms of absence of input features.



**Scenarios:** a) Corners. b) Average pixel value. c) No star-shaped objects. d) Combination of a)-c).



# Right for the Right Reasons

We need to verify that our DNNs or DNN+XAIs did not learn from spurious correlations.

**If** we verify that our XAI methods work, **and** show our models (robustly) rely on robust rationales, our models are (in some sense) trustworthy.

This is Klaus' Clever Hans.



But there's a data bottleneck challenge here!





# Right for the Right Reasons

The verification is typically against human rationale annotations. We recently collected multilingual eye-tracking data across demographics from webcams ([WebQamGaze](#)) to see if this provides cheap rationale annotations. So far, gaze data seems to induce similar method and model rankings as rationale annotations.



## Movie Reviews

In this movie, ... Plots to take over the world. The acting is great! The soundtrack is run-of-the-mill, but the action more than makes up for it

(a) Positive (b) Negative

## e-SNLI

H A man in an orange vest leans over a pickup truck  
P A man is touching a truck

(a) Entailment (b) Contradiction (c) Neutral

## Commonsense Explanations (CoS-E)

Where do you find the most amount of leafs?

(a) Compost pile (b) Flowers (c) Forest (d) Field (e) Ground

## Evidence Inference

**Article** Patients for this trial were recruited ... Compared with 0.9% saline, 120 mg of inhaled nebulized furosemide had no effect on breathlessness during exercise.

**Prompt** With respect to *breathlessness*, what is the reported difference between patients receiving *placebo* and those receiving *furosemide*?

(a) Sig. decreased (b) No sig. difference (c) Sig. increased



# Debates in Philosophy of Science

**Carlos Zednik:** DNN+XAIs “possess unique epistemic qualities.”



**Cynthia Rudin:** “Explanations must be wrong. They cannot have perfect fidelity with respect to the original model.”

**Solution:** Both DNNs and DNN+XAIs can be (more or less wrong) explanations.

1. XAI can help us identify relevant hypotheses; but so could **feature selection**.
2. XAI can estimate robustness if **ground truth** rationales are available; but so could more data.
3. So, DNN+XAI does **not** have unique epistemic qualities, but still useful.

**Preview:** Ground truth in XAI evaluation will be as problematic as for DNN evaluation.

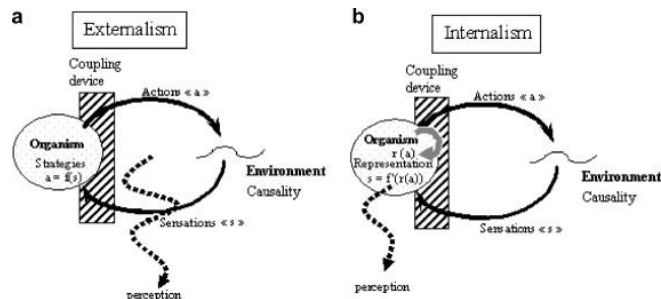
# Debates in Philosophy of Science

**Internalism:** XAI through analysis of internal representations. Examples: Vanilla gradients, LRP, similarity.

**Externalism:** XAI through external, causal chains. Two kinds: conservative (only explain in terms of past interactions) and progressive (also explain in terms of future interactions). Examples: influence functions (conservative), uptraining (progressive), LIME (progressive).

## Solution:

- a) Conservative externalism is not gonna cut it, because models are not determined by their training data.
- b) Internalism and progressive externalism often form a continuum, because you can probe the relevant representational differences.

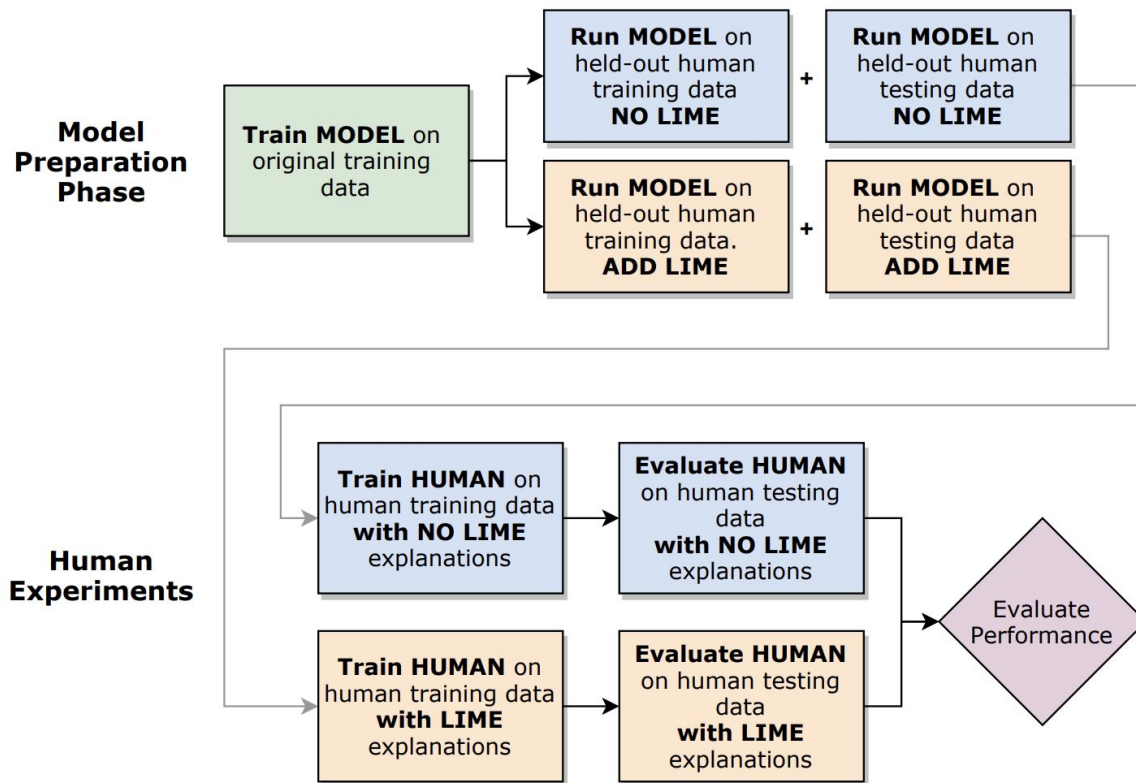


# Ways of Evaluating XAIs

---

# Evaluation Strategies

1. Heuristic Evaluation (input reduction, hot flip, etc.)
2. Human Annotation (e.g., ERASER)
3. Human-in-the-Loop (forward prediction, **Reverse Turing Test**)
4. Real-Life Experiments (QuizBowl, EkstraBladet)
5. Inherent Limitations



# Evaluation Strategies

1. Heuristic Evaluation (input reduction, hot flip, etc.)
2. Human Annotation (e.g., ERASER)
3. Human-in-the-Loop (forward prediction, Reverse Turing Test)
4. Real-Life Experiments (QuizBowl, EkstraBladet)
5. Inherent Limitations

Do you think this comes from a reliable or an unreliable source?

The model has predicted unreliable with 99.96% confidence

Legend of model decisions: ■ Unreliable □ Neutral ■ Reliable

**Claim:** Travis Scott is starting to doubt if he 's Stormi 's Dad Amid Tim Chung Rumors

**Article:** Earlier this month , rumors started swirling that Kylie Jenner 's daughter Stormi Webster might not be her boyfriend Travis Scott 's child because of the shocking resemblance that Stormi has to Kylie 's hot AF bodyguard Tim Chung . Now that baby Stormi is four months old , she 's starting to show more defined features in her face and sources exclusively revealed to In Touch that there is a part of Travis that is starting to doubt Stormi 's paternity and whether or not he is the little girl 's father . “ Travis is starting to get a little worried and questioning Kylie about this whole bodyguard situation . Not to diss or say Kylie 's a liar but he doesn 't watch her every move . He 's not with her 24 / 7 and there were times they were a part from each other nine months ago . He loves Stormi and truly believes that 's his daughter but can 't help but notice that she doesn 't look like him . In the back of his mind he wonders if Kylie strayed . If that happened and Stormi 's not his , that would be the most devastating news of his life . He flat out wants to talk to Kylie and Tim together , to once and for all get to the bottom of this . ” Ever since Kylie started sharing more and more photos of baby Stormi , fans started to realize that Ky 's daughter kind of looks like she has some of Tim 's features and they started to point it out in the comments . And given the whirlwind nature of Kylie 's relationship with Travis — they started dating just one month after her split from her ex - boyfriend Tyga — and the fact that Travis was busy touring during the beginning of their relationship , it seems like there could have very well been a few instances where Kylie could have been alone with her security guard . According to sources , Tim even bragged about how much alone time he has been able to spend with Kylie . “ He 's telling his friends that he 's been alone with Kylie tons of times in her house but when they ask if [ they 've been intimate ] , he simply smiles , ” and an insider revealed to In Touch .

reliable

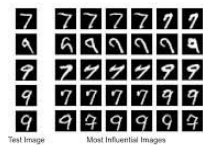
unreliable

# Human Evaluation for Influence?

Human evaluation of feature attribution is common, through annotation or forward prediction, but what about human evaluation of training data influence? Seems infeasible, but consider simple cases:

$f(3)=3$ ,  $f(4)=5$ ,  $f(5)=6$ ,  $f(6)=?$

**Protocol:** Ask participants to predict  $f(6)$  given the above sequence. Subsequently, ask them what training data influenced their decision.



Test Image

Most Influential Images



Test Image

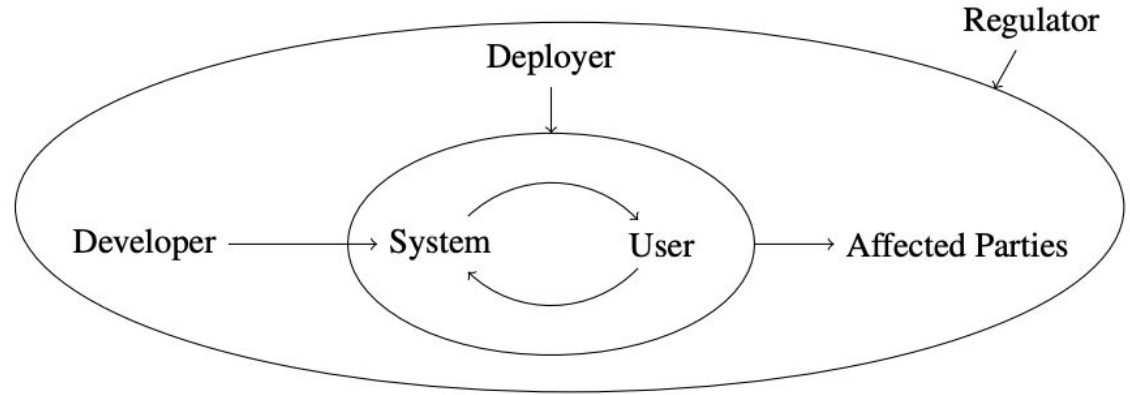
Most Influential Images

# Whose Explanations?

---

# People and Explanations

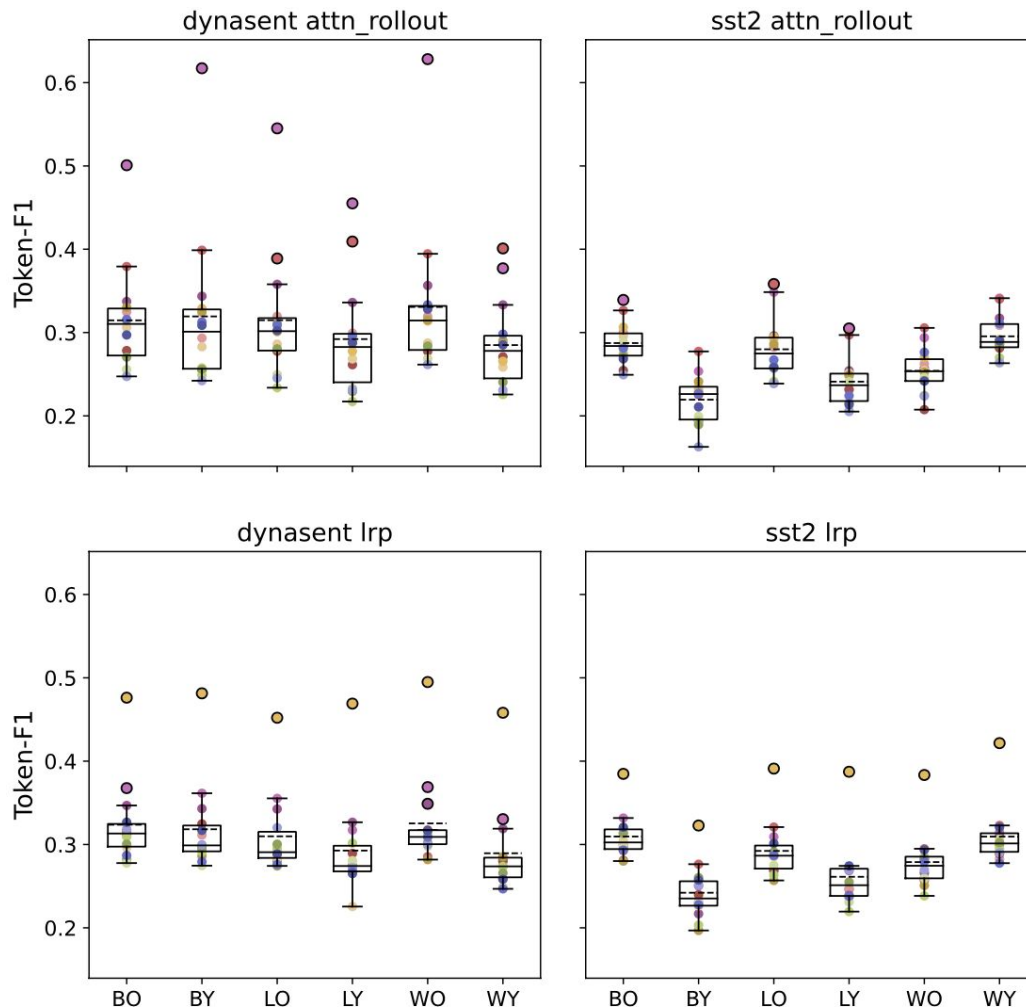
- a) Different stakeholders to XAI.
- b) E.g.: Influence functions are useful for developers and regulators, but probably not for other stakeholders.
- c) Scientists may be willing to consider more uncertain associations than affected parties.
- d) But when evaluating whether models are right for the right reasons, is that subjective or objective?





# Whose Explanations?


We collected rationale annotations across six demographics - the cross-product of {young, old} and {Black, Latin, White}. We see statistically significant differences across two sentiment analysis tasks and a question answering task.



# Whose Explanations?

This view is under-explored. E.g., in ACL 2021, none of 18 XAI papers looked at fairness or bias. Related findings:

- a) Multilingual models are not equally right for the right reasons across languages ([BlackBoxNLP 2022](#)).
- b) Interpretability and fairness are often at odds; interpretability and privacy too (submitted to [AISTATS 2023](#)).

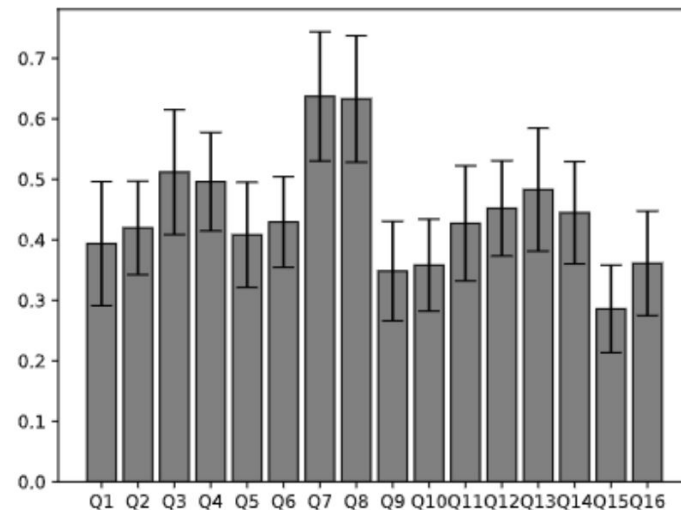


Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
ACL 2021 oral papers	461	69.4%	38.8%	13.9%	6.3%	17.8%	11.7%	6.1%
MT and Multilinguality	58	0.0%	15.5%	56.9%	5.2%	19.0%	6.9%	13.8%
Interpretability and Analysis	18	88.9%	27.8%	5.6%	0.0%	5.6%	66.7%	5.6%
Ethics in NLP	6	83.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Dialog and Interactive Systems	42	90.5%	21.4%	0.0%	9.5%	23.8%	2.4%	2.4%
Machine Learning for NLP	42	66.7%	40.5%	19.0%	4.8%	50.0%	4.8%	9.5%
Information Extraction	36	80.6%	91.7%	8.3%	0.0%	25.0%	5.6%	8.3%
Resources and Evaluation	35	77.1%	42.9%	5.7%	8.6%	5.7%	14.3%	5.7%
NLP Applications	30	73.3%	43.3%	0.0%	10.0%	20.0%	10.0%	0.0%

# Thought Experiments

Imagine you go to your personal doctor. Your doctor says they have been screening your personal health records, feeding them to an advanced predictive model, which predicts you to have *Snepsosis* - a disease for which a treatment with only moderate side-effects exists. How likely would you be to accept the treatment, based on this prediction?

Imagine you go to an interview for a job in a large, international company. The CEO tells you that they have been screening your application with an advanced predictive model. She tells you that while you were found to be the most qualified candidate for the job in question, the model found you to be even more qualified for another job. The job has the same status as the one you applied for, but the domain is a bit different. How likely would you be to accept the proposed job, based on this prediction?



**Average distrust.** Odd: Own trust. Even: Perceived trust.  
Q1, Q2, Q9, Q10: DNNs. Others: DNN+XAIs. **Findings:** XAI lowers trust. Perceived trust higher.

?

coAStal

