Explainable Active Learning (XAL):
Toward AI Explanations as Interfaces for Machine Teachers

Ghai et al. (2020)

Presented by *Lucia Gordon*, *Matthew Nazari*, and *Catherine Yeh*
March 22, 2023

Prediction Task

- Dataset: Adult Income
- Binary prediction: whether the annual income of an individual is > or < $80k
- Linear model: logistic regression with L2 regularization

How much money might a person with the following attributes make in a year?

Attributes	Values
Age	39
Workclass	Private
Years of Education	13
Marital status	Never married
Occupation	Executive & managerial
Race	White
Gender	Male
Hours per week	50

Customer profile

Active Learning

- Current ML development processes are bottlenecked by
  - resources and expertise
  - how, where, and when we can leverage human knowledge

Active Learning in Prediction Task

- Participants judge the income level of data points queried by the model
- Query method: uncertainty sampling
- The annotations could be faulty
- BUT… explanations might be able to help reveal faulty model beliefs

How much money might a person with the following attributes make in a year?

Attributes	Values
Age	39
Workclass	Private
Years of Education	13
Marital status	Never married
Occupation	Executive & managerial
Race	White
Gender	Male
Hours per week	50

Customer profile

Problem Statement

- Active learning interfaces remain minimal and opaque
  - annotator cannot monitor training progress
  - annotator is unaware of the effectiveness of their teaching
- Explanations show potential for better active learning

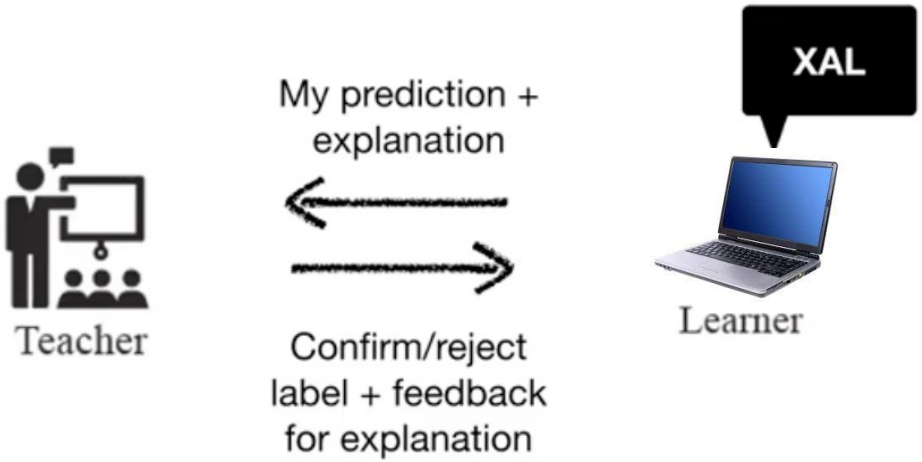
Big Question:

Can explanations improve the active learning process?

Summary of Contributions

1. Provide insights into explainable AI as an interface for active learning
2. Propose a novel active learning paradigm: *explainable active learning* (XAL)
3. Conduct an empirical study using the prediction task we described to investigate the impact of explanations on annotation experience

Explainable Active Learning (XAL)



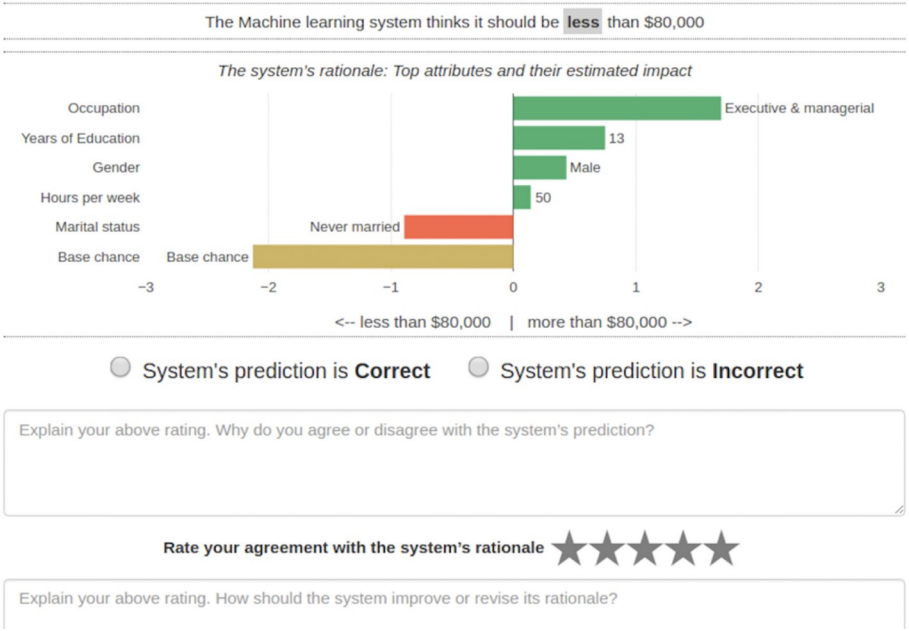
- Mimics how humans teach and learn
- Explanations can make it easier to reject an incorrect prediction
- Explanations can inform the feedback

Presentation of Prediction + Explanation

Prediction

Explanation

Feedback



Design choices

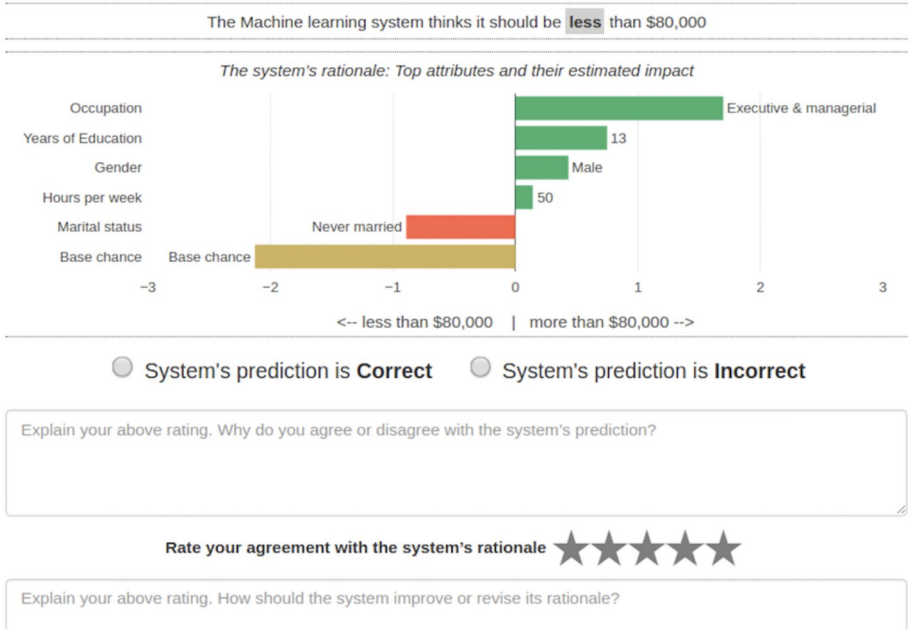
- Explain AL with local feature importance
- Present local feature importance with visualization
- Features were sorted by their importance
- Top 5 most important features were included in the visualization
- Green = positive features
- Red = negative features
- Base chance = model intercept

Presentation of Prediction + Explanation

Prediction

Explanation

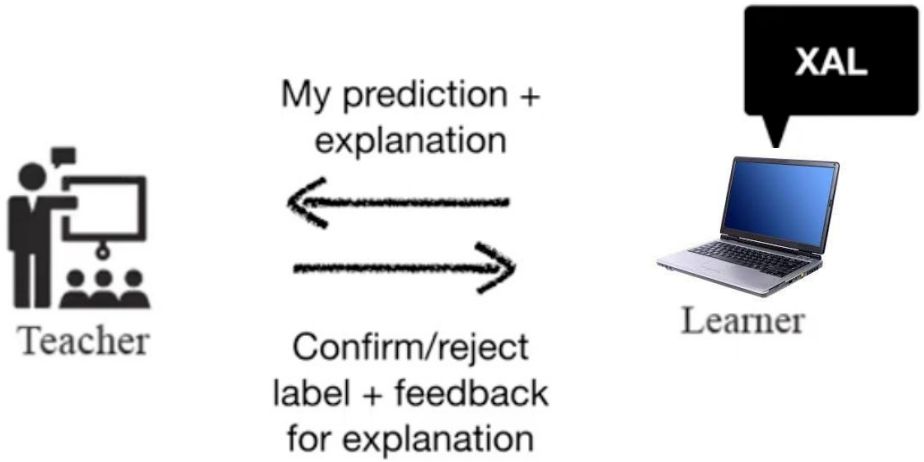
Feedback



How it differs from previous work

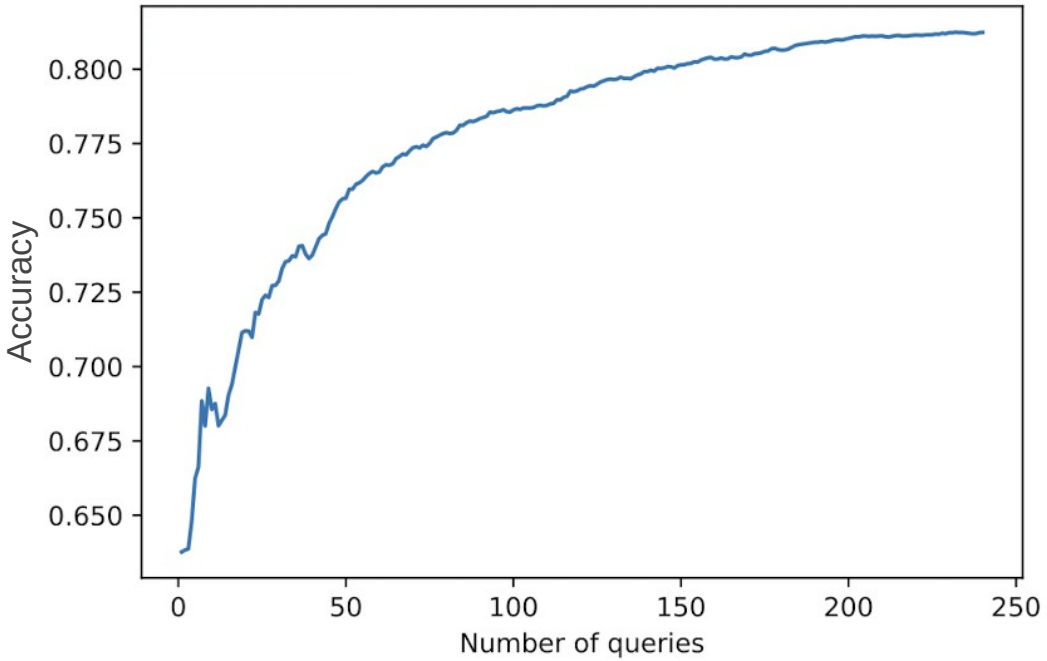
- Presents the model's reasoning rather than requesting global feature weights
- Goes beyond text-based models

Explainable Active Learning (XAL)



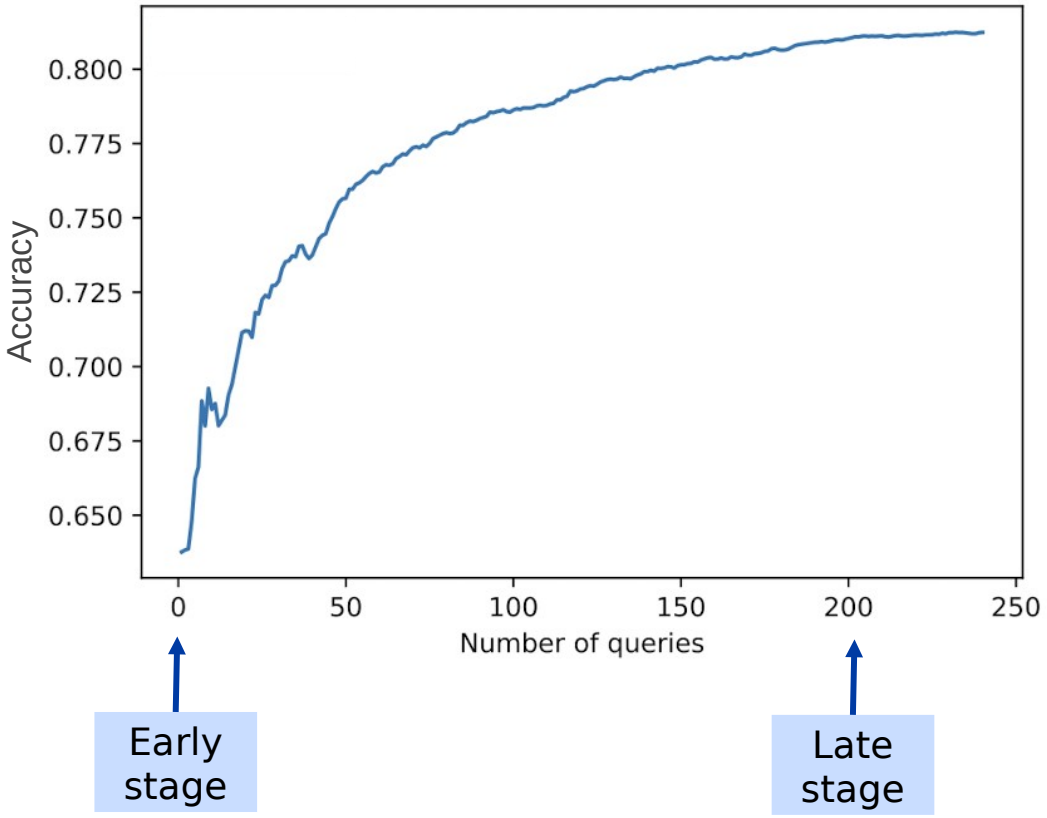
They design a study to explore **how people naturally want to teach a model with explanations** in the form of a local feature importance visualization

Active Learning Simulation



- Wanted to examine the effect of explanations at different stages of AL
- Used a simulation to define “early-stage” and “late-stage” models
- Queried instances were annotated using ground-truth labels in the simulation

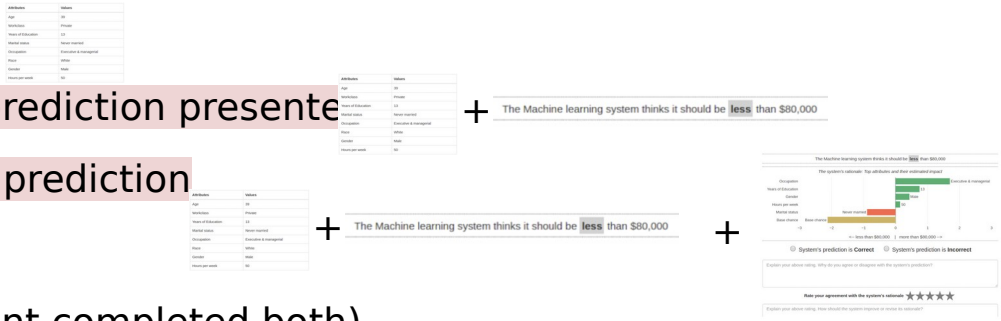
Active Learning Simulation



- Early stage → 0 queries
- Late stage → 200 queries (where accuracy plateaus)

Experimental Design

- Amazon Mechanical Turk: 36 participants
- Three **conditions** (each participant assigned to one)
  a. AL: customer profile presented
  b. CL: customer profile + model prediction presente
  c. XAL: customer profile + model prediction
     + explanation



- Two learning **stages** (each participant completed both)
  a. Early
  b. Late
- 20 annotations per experiment (condition + stage)
- Domain knowledge training (statistics, practice trials, $2 bonus for consistency with ground-truth)

Research Question 1

How do local explanations impact the annotation and training outcomes of AL?

Results: Annotation + Learning Outcomes

Active (AL)
Coactive (CL)
Explainable (XAL)

Table 1. Results of model performance and labels

Stage	Condition	Acc.	Acc. improve	F1	F1 improve	%Agree	Human Acc.
Early	AL	67.0%	13.7%	0.490	0.104	55.0%	66.7%
	CL	64.2%	11.7%	0.484	0.105	58.3%	62.1%
	XAL	64.0%	11.8%	0.475	0.093	62.9%	63.3%
Late	AL	80.4%	0.1%	0.589	0.005	47.9%	54.2%
	CL	80.8%	0.2%	0.587	0.007	55.8%	58.8%
	XAL	80.3%	-0.2%	0.585	-0.001	60.0%	55.0%

Compare to initial model performance (before 20 queries)

RQ1: Overall, human accuracy and % agreement decrease in the later stage, while model accuracy increases.

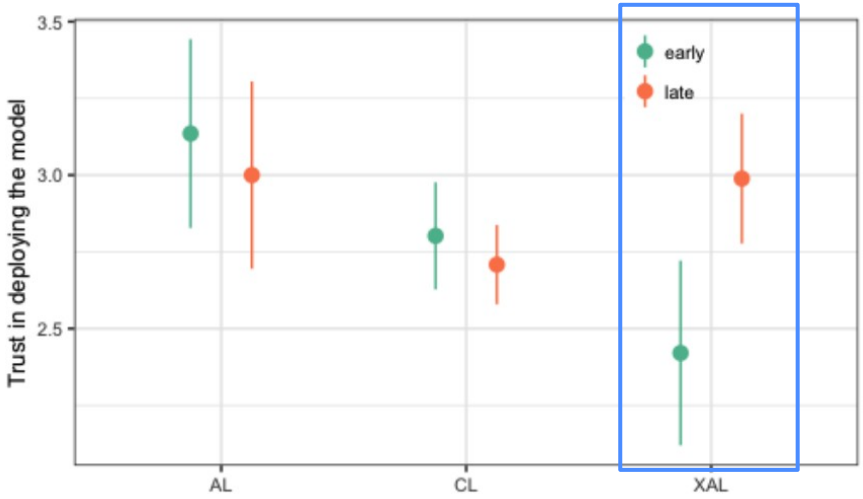
Research Question 2

How do local explanations impact annotator experiences?

Hypotheses

1. Explanations support ***trust calibration***
2. Explanations improve ***annotator satisfaction***
3. Explanations increase perceived ***cognitive workload***

Results: Annotator Experience



H1: Explanations support trust calibration

RQ2: Trust in XAL condition is low early on, but increases with later stage model.

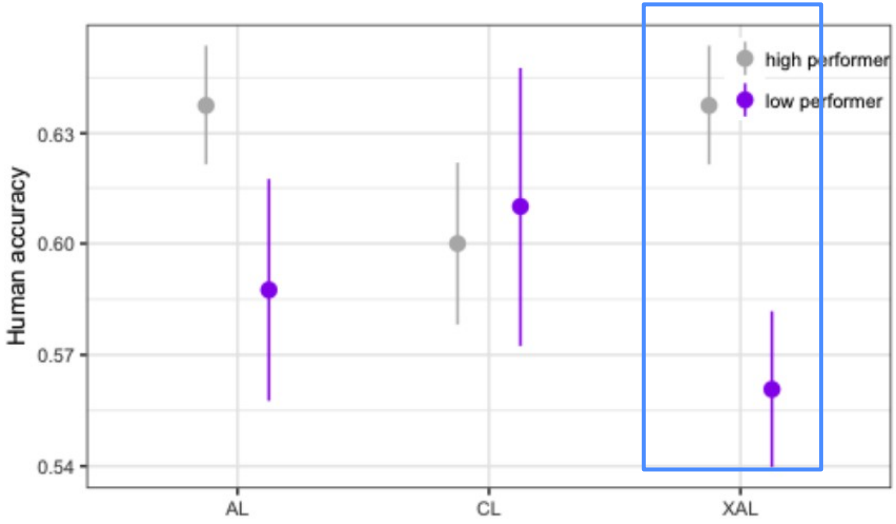
Research Question 3

How do individual factors, specifically ***task knowledge***, ***AI experience***, and ***need for cognition***, impact annotation and annotator experiences with XAL?

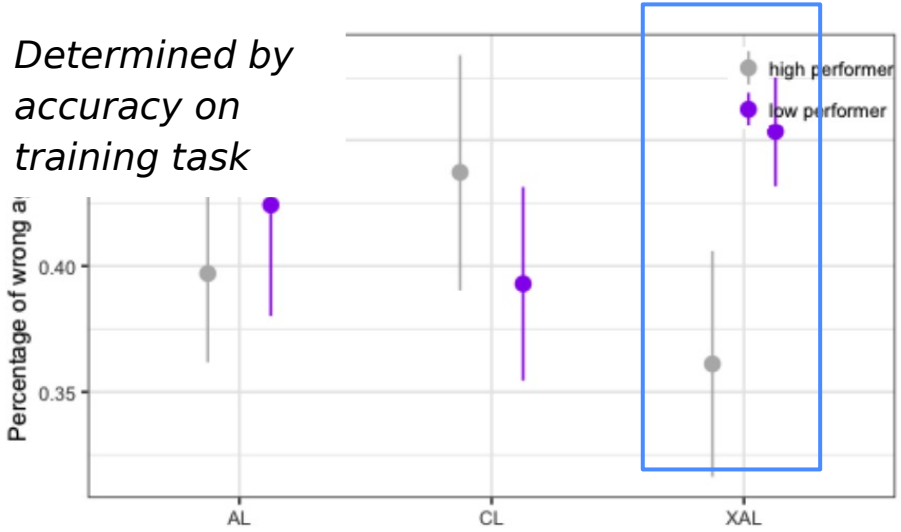
Hypotheses

4. Annotators with lower task knowledge benefit more from XAL
5. Annotators inexperienced with AI benefit more from XAL
6. Annotators with lower need for cognition have a less positive experience with XAL

Results: Annotation + Learning Outcomes



Determined by accuracy on training task



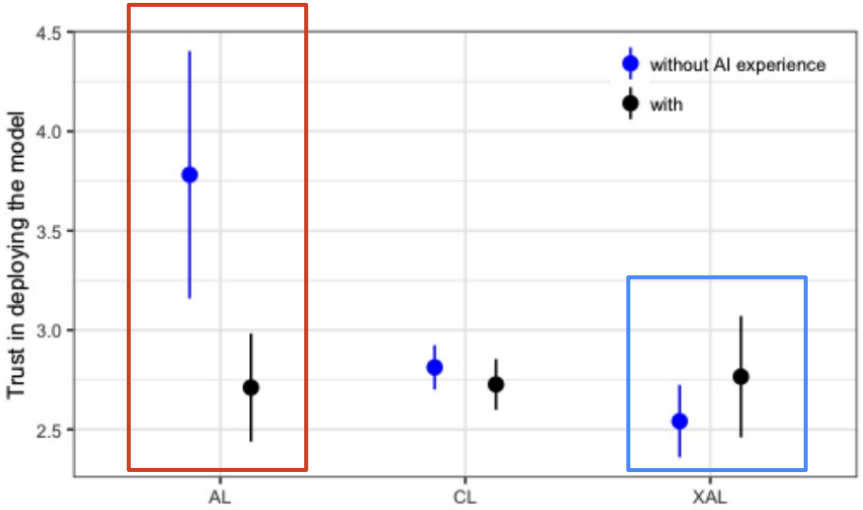
RQ3: Participants with less task knowledge had lower accuracies and higher blind trust in XAL condition.

Results: Annotation + Learning Outcomes

H4: Annotators with lower task knowledge benefit more from XAL

RQ3: Participants with less task knowledge had lower accuracies and higher blind trust in XAL condition.

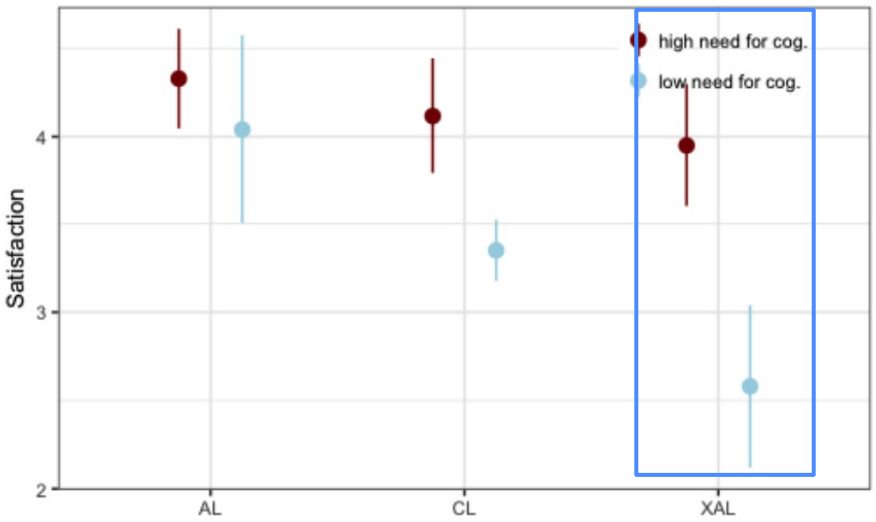
Results: Annotator Experience



H5: Annotators inexperienced with AI benefit more from XAL

RQ3: Explanations helped calibrate trust for those without AI experience.

Results: Annotator Experience

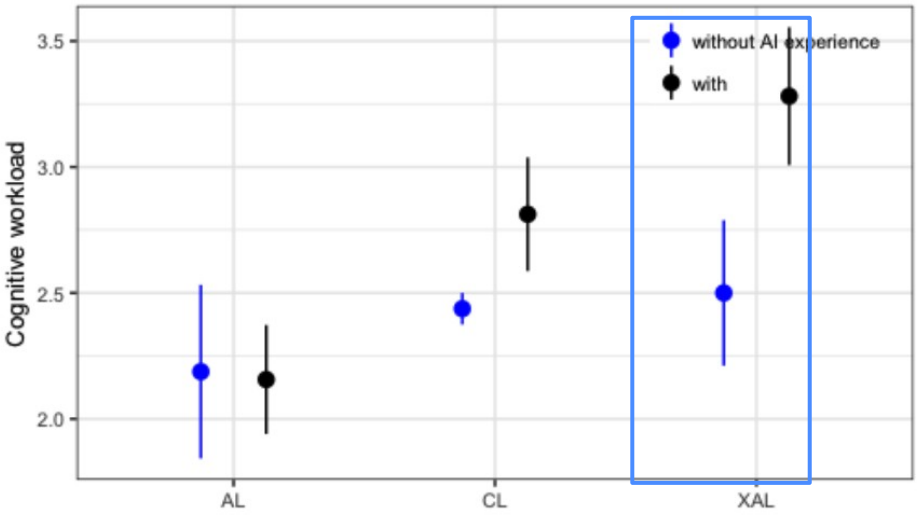


H2: Explanations improve annotator satisfaction

H6: Annotators with lower need for cognition have a less positive experience with XAL

RQ3: Explanations have a significant negative effect on satisfaction for those with low need for cognition.

Results: Annotator Experience



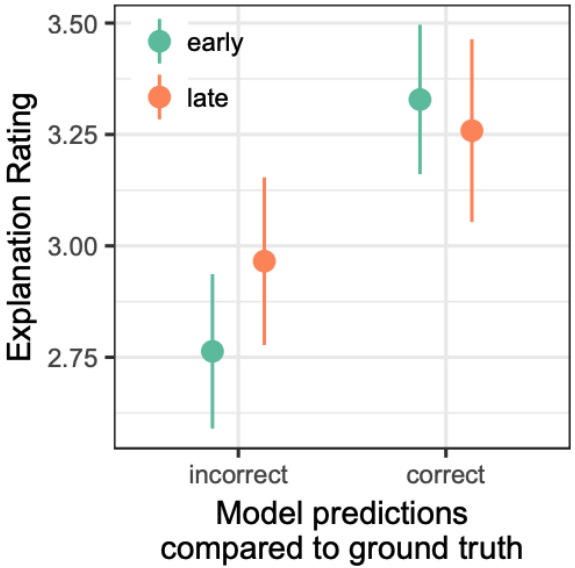
H3: Explanations increase perceived cognitive workload

RQ3: XAL induces significantly higher workload for those with AI experience.

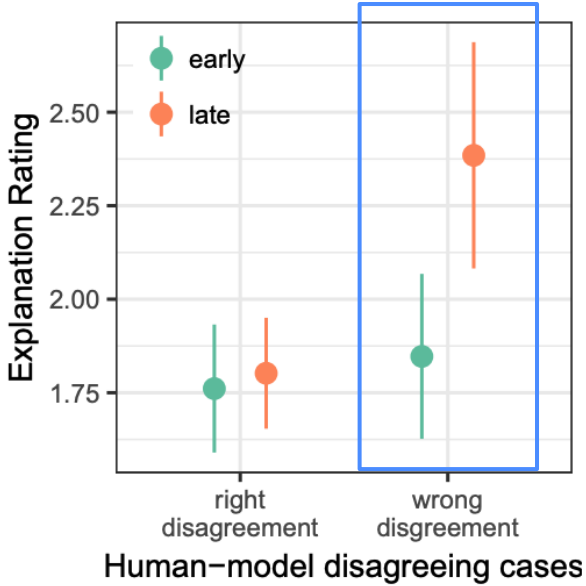
Research Question 4

What kind of feedback do annotators naturally want to provide upon seeing local explanations?

Results: Feedback for Explanation



Explainable (XAL) condition

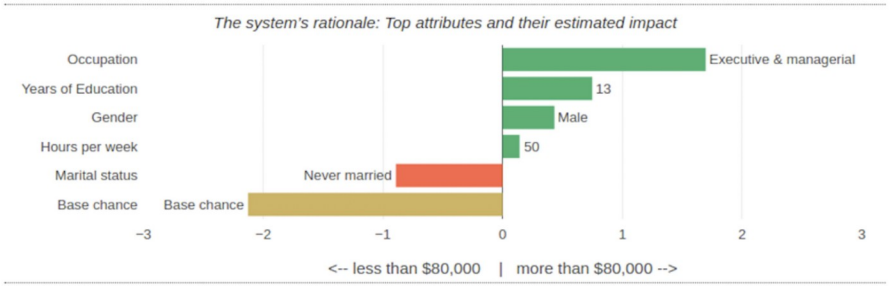


RQ4: Explanation ratings are higher when model is correct and when annotators wrongly disagree with the model in later stage tasks.

Results: Feedback for Explanation

Open Form Feedback

- Tuning weights (N = 81)
- Removing, changing direction of, or adding features (N = 28)
- Ranking or comparing multiple feature weights (N = 12)
- Reasoning about combination and relations of features (N = 10)
- Logic to make decisions based on feature importance (N = 6)
- Changes of explanation (N = 5)



Related Work: Active Learning

- Many AL paradigms have been developed, and the paper lists many
- Very little or no attention has been given to understanding or improving how humans interact with AL algorithms
  - In human-robot interaction, a natural language interface was developed (Cakmak et. al)
- They identify the need to study annotation interactions with real-time AL algorithms

Related Work: Interactive ML

- AL is sometimes considered to be within interactive machine learning (iML)
- iML approaches value transparency over performance
- There have been empirical studies demonstrating iML techniques lower need for data, but little else
- iML approaches are esoteric, but explanations as interfaces could help non-ML experts

Related Work: Explainable AI

- The paper describes explainable AI (XAI) from a broad overview
- Recent XAI studies find users’ understanding of AI systems improved by explanations
- There are still many unknowns to how explanations cold reduce knowledge barriers to train ML models

Future Work

Explanations for Active Learning

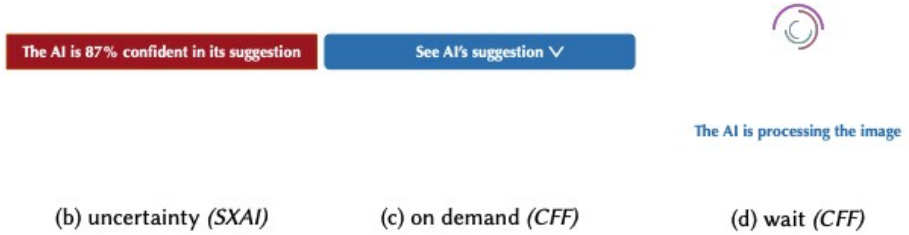
- Alternative designs to mitigate anchoring effect

CSCW '21

ss

To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making

ZANA BUÇINCA, Harvard University, USA
MAJA BARBARA MALAYA, Lodz University of Technology, Poland
KRZYSZTOF Z. GAJOS, Harvard University, USA



(b) uncertainty *(SXAI)* (c) on demand *(CFF)* (d) wait *(CFF)*

Future Work

Learning from explanation based feedback

- Scaffolding the elicitation of high-quality, targeted feedback

CSCW '15

Structuring, Aggregating, and Evaluating Crowdsourced Design Critique

Kurt Luther[1], Jari-Lee Tolentino[2], Wei Wu[3], Amy Pavel[3],
Brian P. Bailey[4], Maneesh Agrawala[3], Björn Hartmann[3], Steven P. Dow[1]

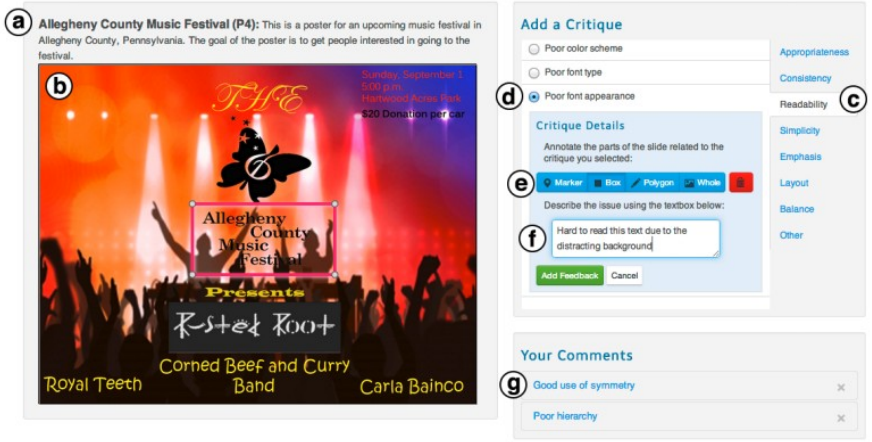
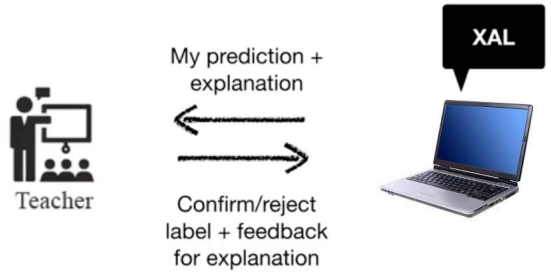


Figure 3. The CrowdCrit critique interface.

Conclusions



Key contributions

- New approach for machine teaching: XAI + AL = X

Limitations

- If domain experts are usually the ones doing the labeling, why not survey them?
- Small number of participants
- Only tried uncertainty sampling (but other strategies may have different impacts on annotator experiences)
- Only tried one explanation method with a highly interpretable model - how would the results change under different explanation methods or more black-box models?
- What if the number of relevant features is larger (such as pixels in an image?)

Discussion Questions

- Has anyone used AL in the past? If so, do you think adding explanations would have been helpful? Why or why not?
- What applications of XAL are you most excited about? Or, any particular scenarios that you think it would work especially well/poorly in?
- Do you think XAL is feasible with less inherently interpretable models or other explanation methods? How do you think the annotations and the annotation experience would change?