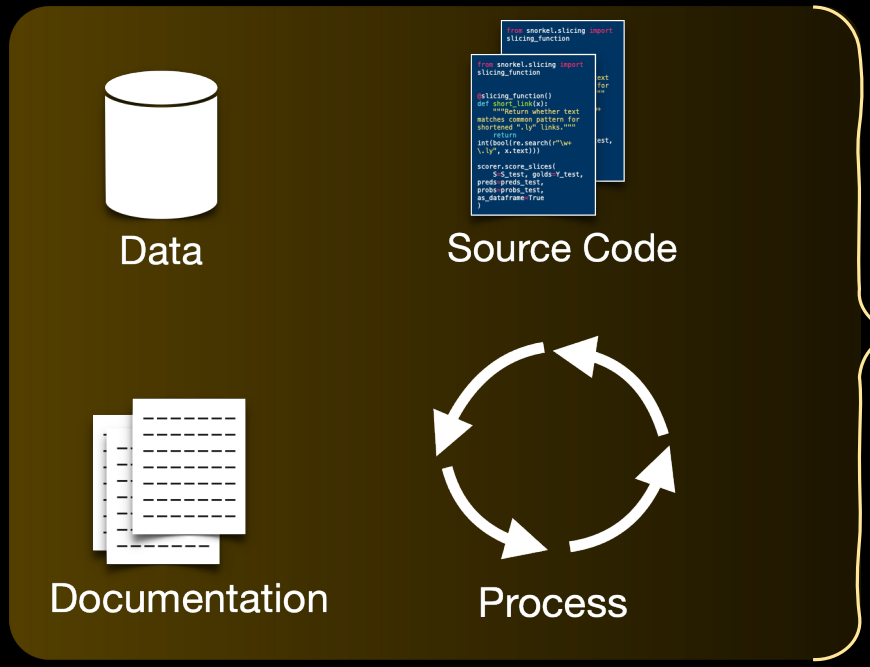
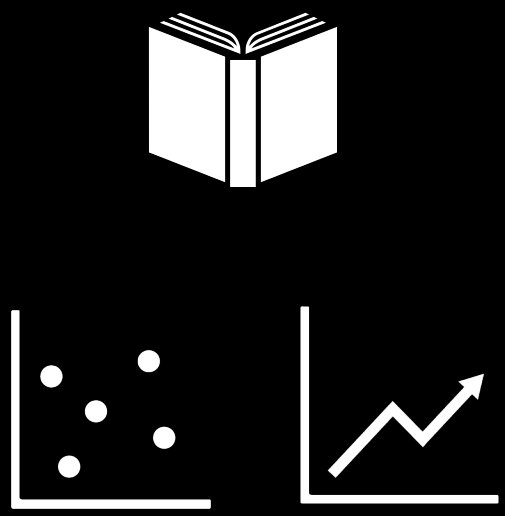
Transparency and Explainability

Jin L.C. Guo

SOCS, McGill University

Transparency: AI systems should be designed and implemented in such a way that oversight of their operations are possible.





Explainability: translate the technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation.

Activity

- Consider the system you chose for your assignments:
  - When and where do the stakeholders of the system have the need to understand the AI? Why? (consider only one type of stakeholder in this activity)
  - How the need can be addressed?
  - List three concrete questions the stakeholder might ask?
  - What should the explanation look like to the questions you gathered?
  - Put your questions into appropriate categories on Miro board (XAI Questions)

Is the question from

Data Scientists

ML Engineers

End Users

Product Managers

Auditors

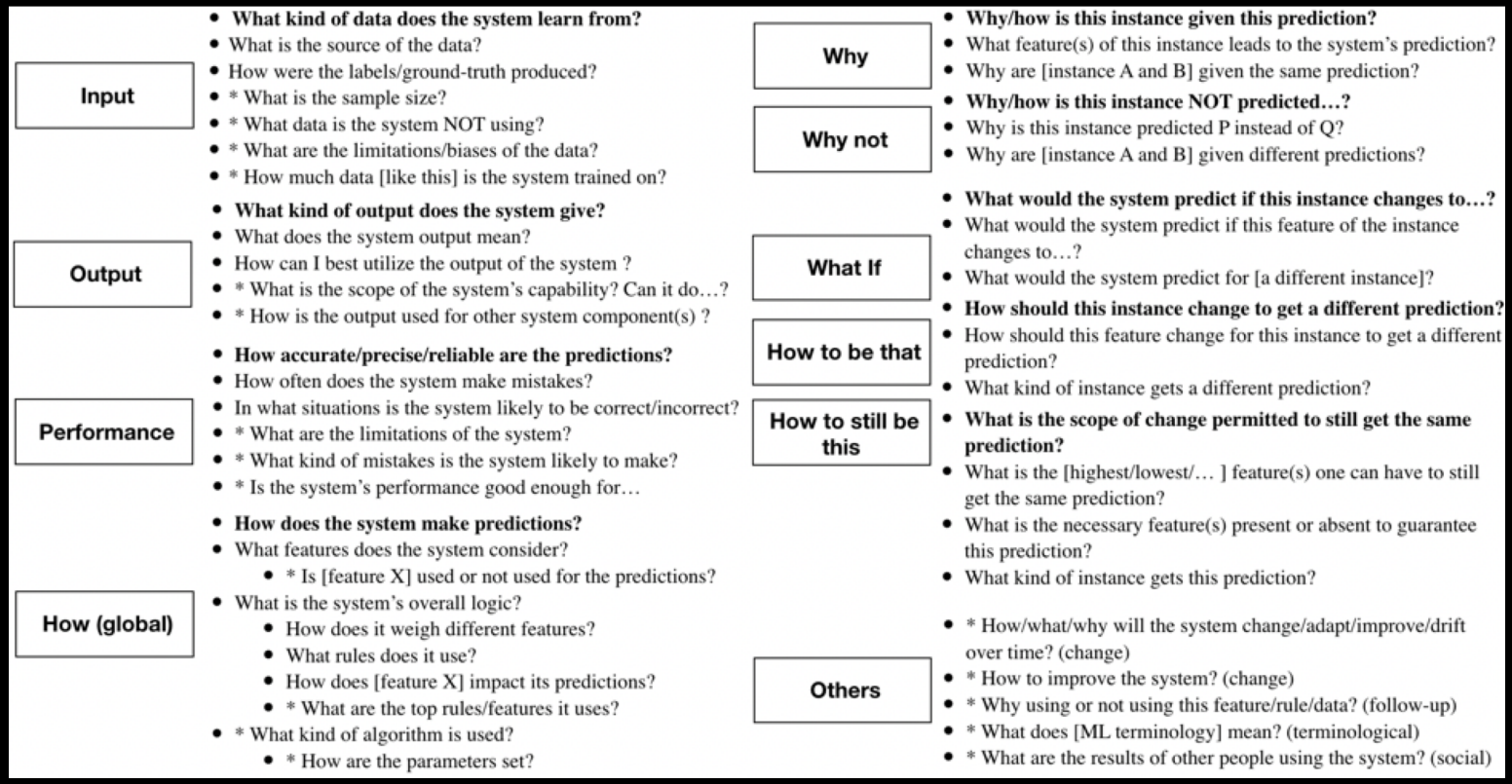
Is the question asking

How

Why, why not

What if

How to be that / How to still be that



Liao, Q. Vera, Daniel Gruen, and Sarah Miller. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-15. 2020.

What makes a good explanation?

Human Explanation

- Contrastive
- Selective
- Interactive

'Why P not Q?'

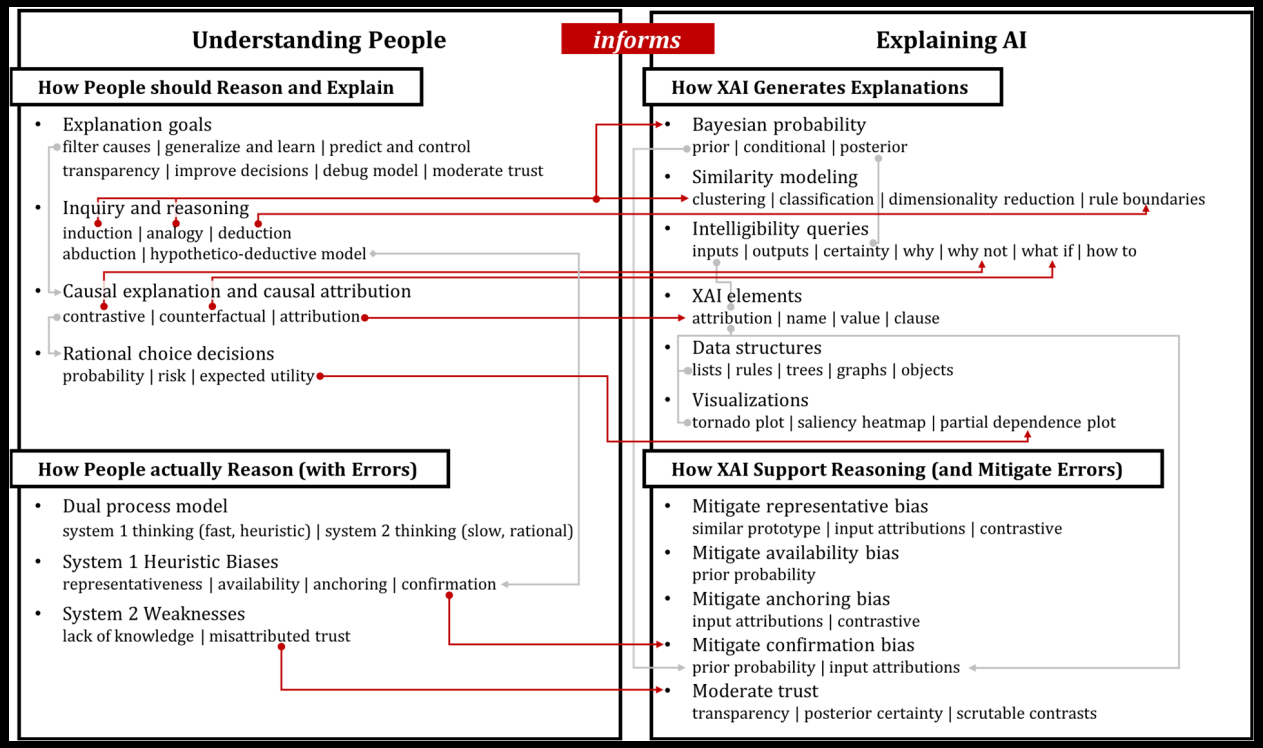
"Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation"

"Explanations are social, insofar as they involve an interaction between one or more explainers and explainees"

To what extent are they applicable to the questions you listed in the activity? Why?

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279-288. 2019.

Explanation Human Reasoning



Wang, Danding, Qian Yang, Ashraf Abdul, and Brian Y. Lim. "Designing theory-driven user-centric explainable AI." In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1-15. 2019.

Explanation for Human Reasoning

Heuristic Bias	Description	Strategies to overcome systematic errors [64]	XAI Strategies for Medical Decisions
Represent-ativeness	Judging likelihood of an event 'A' belonging to a condition due to similarities between the two, but not judging whether A belongs to some other process that could be more similar.	**Compare** disease with **prototypes** of the condition; be suspicious when there is **no good match.**	- Identify **prototypes** of patient instances for each diagnosis - Show similarity between current patient and prototype(s) via **similarity distance**. - Highlight similarity and **contrast** differences in terms of data feature value or **attributions**.
Availability	Bias in perceiving that memorable, unusual or adverse events are more likely (frequent) than they truly are.	Seek **base rate** of a diagnosis.	- Show **prior probability** (equivalent to SHAP bias) of diagnoses (in dataset).
Anchoring	Skewed perception of a value due to a supplied numerical value (anchor).	Avoid confirmation and early closure; make use of **lab tests to "prove" other** leading diagnoses. "Crystal ball" exercise (**"premortem"** prospective hindsight [51]).	- Show input **attributions** for multiple outcomes to allow *contrastive reasoning*. - Facilitate **counterfactual** to test *How To* reduce the probability of primary diagnosis with Rules (e.g., aLIME, LORE). - Facilitate **sensitivity analysis** with *What If* explanations to test stability of primary hypothesis.
Confirmation	Collecting redundant information to confirm an existing hypothesis, instead of finding evidence of competing possibilities.	- Use **hypothetical-deductive** method to assess value and role of contemplated tests. - Try to **disprove your diagnosis**, consider conditions of **higher prevalence**.	- Show Findings (**input attribution**) first, instead of Hypotheses (**output posterior probability**). ***Insight:*** *this is opposite to typical Machine Learning apps to show output uncertainty first.* - Show **prior probability** (equivalent to SHAP bias) of diagnoses (in dataset).

Wang, Danding, Qian Yang, Ashraf Abdul, and Brian Y. Lim. "Designing theory-driven user-centric explainable AI." In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1-15. 2019.

Explanation for Human Reasoning

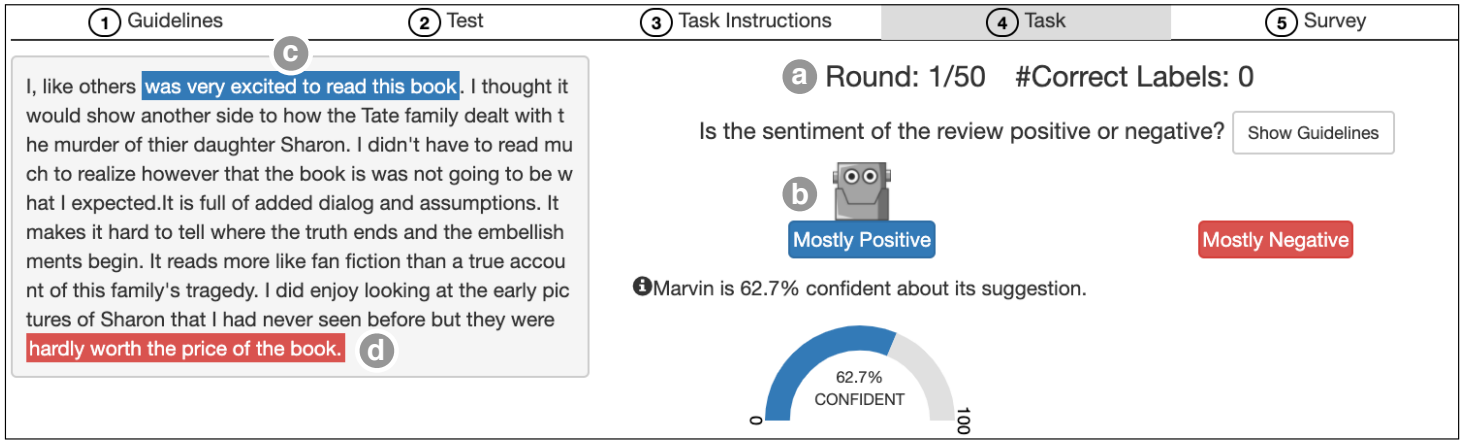


Figure 2: A screenshot of the Team (Adaptive, Expert) condition for the *Amzbook* reviews dataset. Participants read the review (left pane) and used the buttons (right pane) to decide if the review was mostly *positive* or *negative*. The right pane also shows progress and accuracy (a). To make a recommendation, the AI (called "Marvin") hovers above a button (b) and displays the confidence score under the button. In this case, the AI incorrectly recommended that this review was positive, with confidence 62.7%. As part of the explanation, the AI highlighted the most positive sentence (c) in the same color as the *positive* button. Because confidence was low, the AI also highlights the most negative sentence (d) to provide a counter-argument.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. "Does the whole exceed its parts? the effect of ai explanations on complementary team performance." CHI 2021

Explanation for Human Reasoning

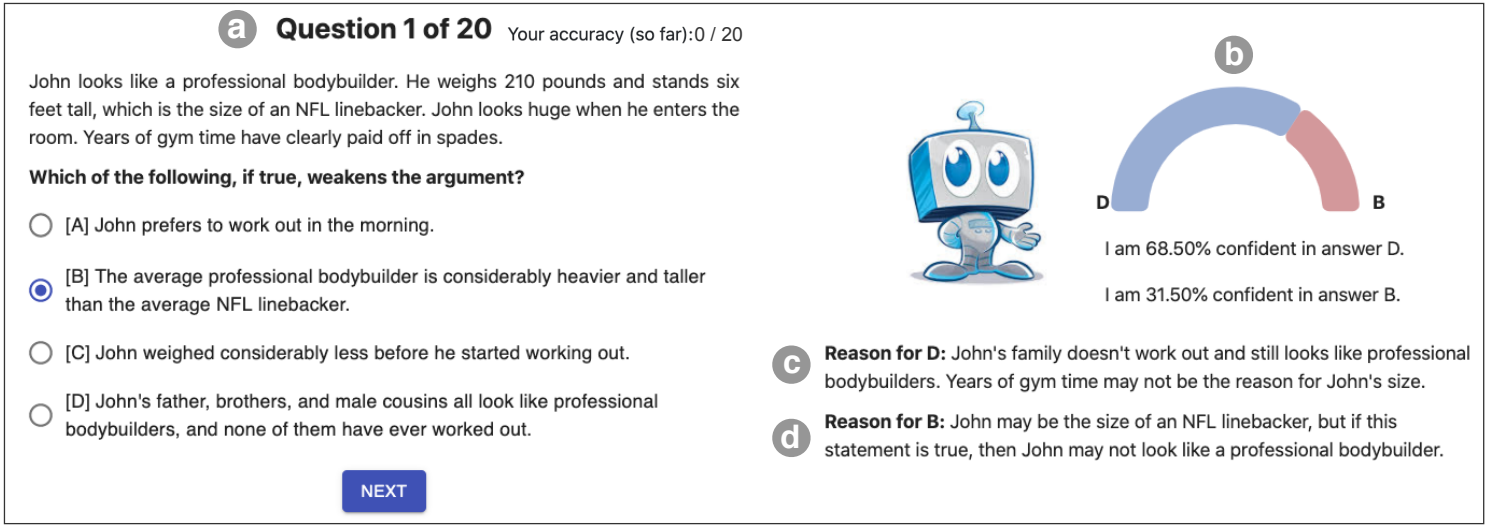
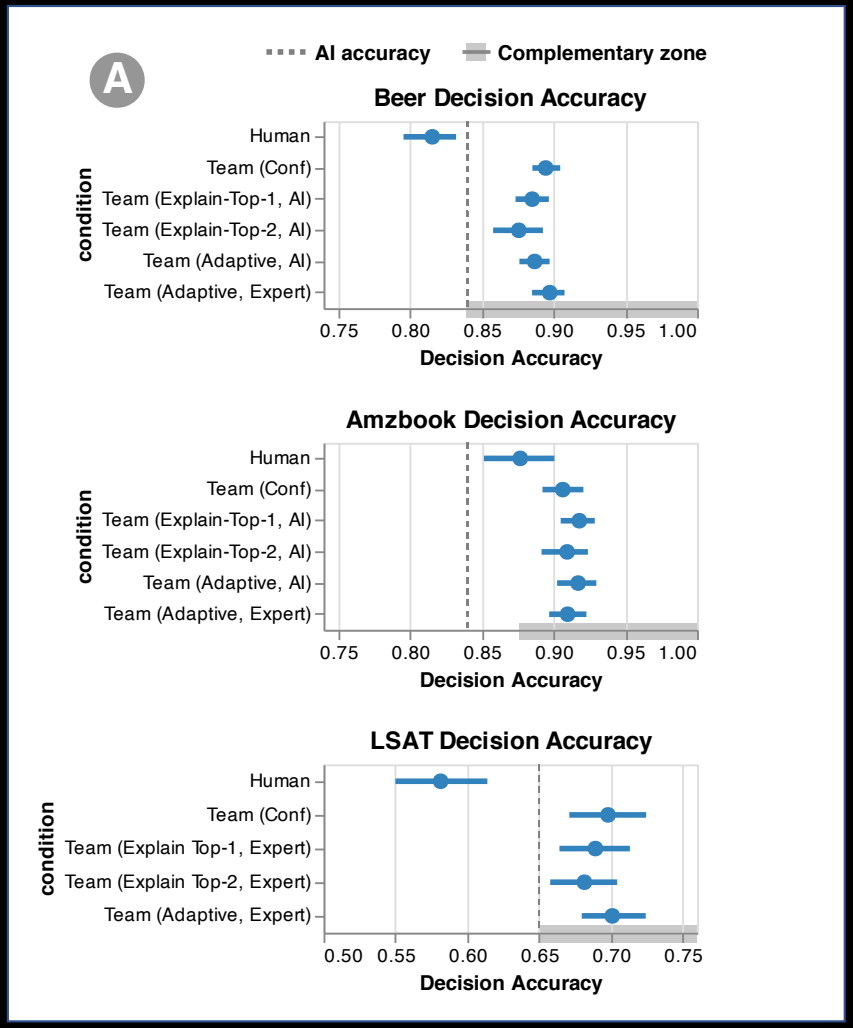
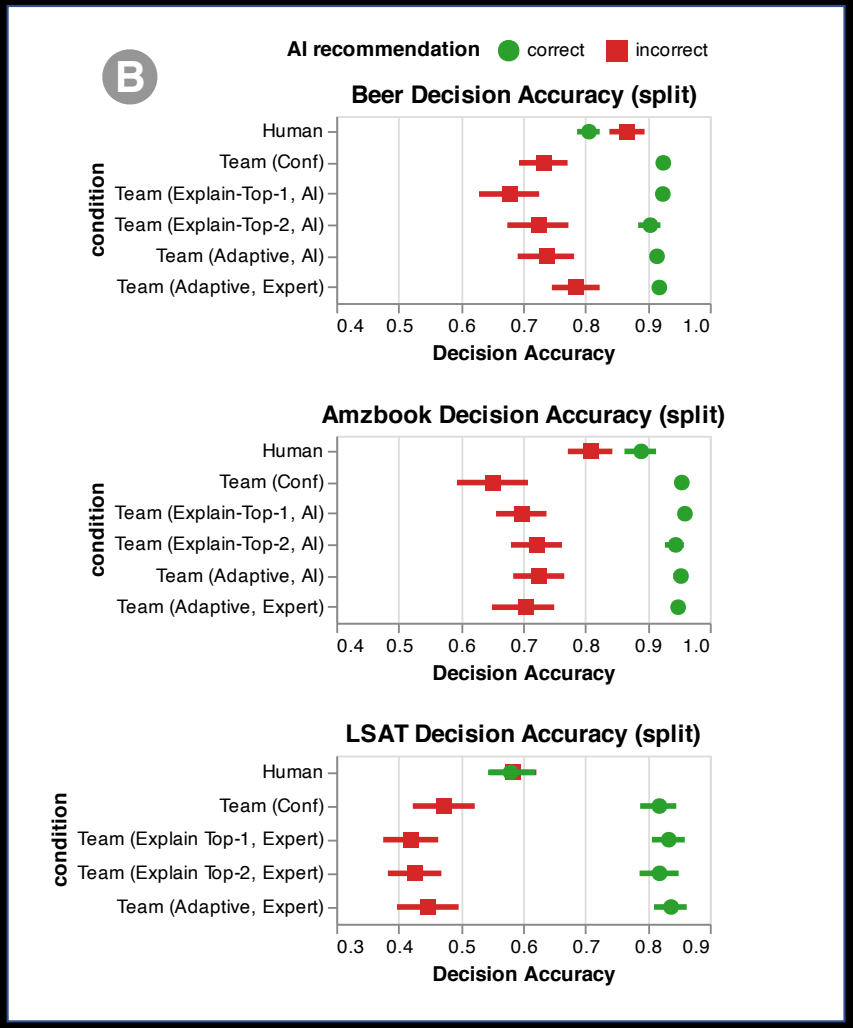


Figure 3: A screenshot of Team (Adaptive, Expert) for *LSAT*. Similar to Figure 2, the interface contained a progress indicator (a), AI recommendation (b), and explanations for the top-2 predictions (c and d). To discourage participants from blindly following the AI, all AI information is displayed on the right. In (b), the confidence score is scaled so those for top-2 classes sum to 100%.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. "Does the whole exceed its parts? the effect of ai explanations on complementary team performance." CHI 2021


A
AI accuracy
Complementary zone
Beer Decision Accuracy
condition
Human
Team (Conf)
Team (Explain-Top-1, AI)
Team (Explain-Top-2, AI)
Team (Adaptive, AI)
Team (Adaptive, Expert)
0.75 0.80 0.85 0.90 0.95 1.00
Decision Accuracy
Amzbook Decision Accuracy
condition
Human
Team (Conf)
Team (Explain-Top-1, AI)
Team (Explain-Top-2, AI)
Team (Adaptive, AI)
Team (Adaptive, Expert)
0.75 0.80 0.85 0.90 0.95 1.00
Decision Accuracy
LSAT Decision Accuracy
condition
Human
Team (Conf)
Team (Explain Top-1, Expert)
Team (Explain Top-2, Expert)
Team (Adaptive, Expert)
0.50 0.55 0.60 0.65 0.70 0.75
Decision Accuracy


B
AI recommendation
correct
incorrect
Beer Decision Accuracy (split)
condition
Human
Team (Conf)
Team (Explain-Top-1, AI)
Team (Explain-Top-2, AI)
Team (Adaptive, AI)
Team (Adaptive, Expert)
0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Accuracy
Amzbook Decision Accuracy (split)
condition
Human
Team (Conf)
Team (Explain-Top-1, AI)
Team (Explain-Top-2, AI)
Team (Adaptive, AI)
Team (Adaptive, Expert)
0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Accuracy
LSAT Decision Accuracy (split)
condition
Human
Team (Conf)
Team (Explain Top-1, Expert)
Team (Explain Top-2, Expert)
Team (Adaptive, Expert)
0.3 0.4 0.5 0.6 0.7 0.8 0.9
Decision Accuracy

Takeaways

- Informative instead of convincing
- Move beyond confidence score
- Explanation is part of the design

Next

Fairness