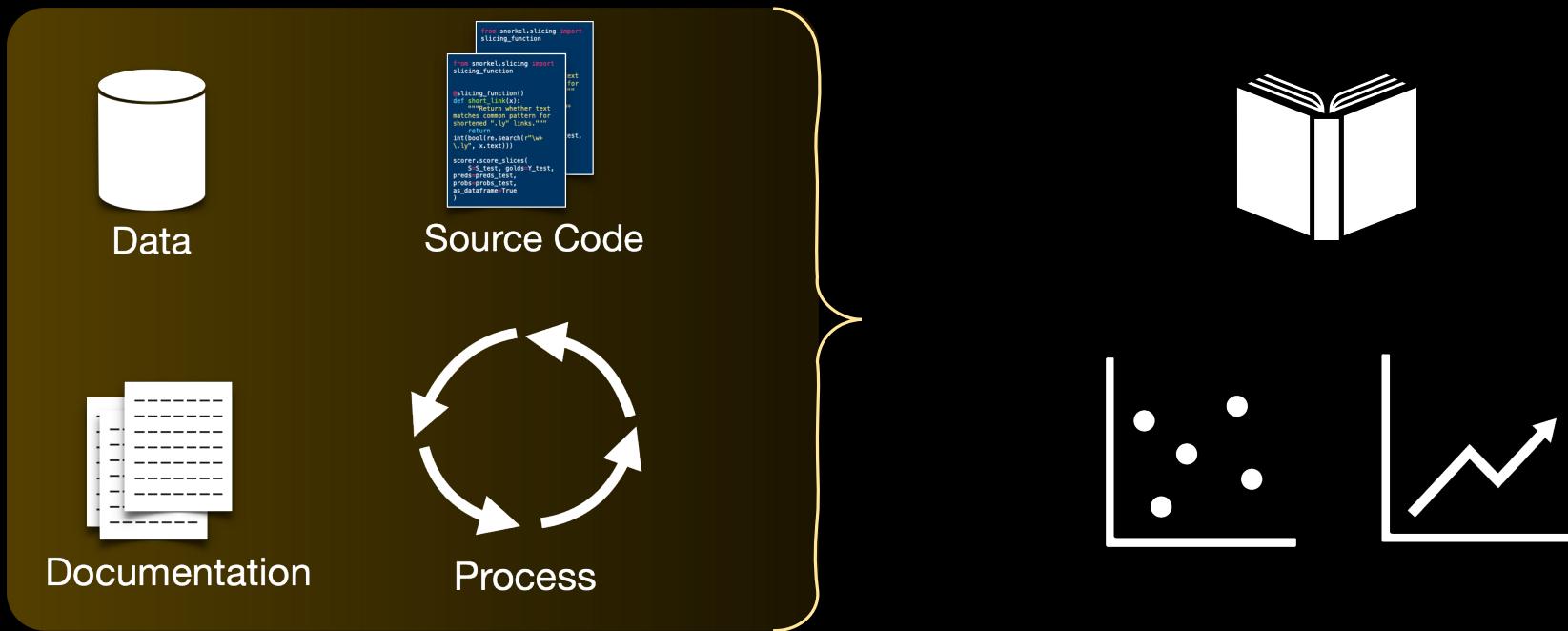




Transparency and Explainability

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

Input	<ul style="list-style-type: none"> • What kind of data does the system learn from? • What is the source of the data? • How were the labels/ground-truth produced? • * What is the sample size? • * What data is the system NOT using? • * What are the limitations/biases of the data? • * How much data [like this] is the system trained on? 	Why	<ul style="list-style-type: none"> • Why/how is this instance given this prediction? • What feature(s) of this instance leads to the system's prediction? • Why are [instance A and B] given the same prediction?
Output	<ul style="list-style-type: none"> • What kind of output does the system give? • What does the system output mean? • How can I best utilize the output of the system ? • * What is the scope of the system's capability? Can it do...? • * How is the output used for other system component(s) ? 	Why not	<ul style="list-style-type: none"> • Why/how is this instance NOT predicted...? • Why is this instance predicted P instead of Q? • Why are [instance A and B] given different predictions?
Performance	<ul style="list-style-type: none"> • How accurate/precise/reliable are the predictions? • How often does the system make mistakes? • In what situations is the system likely to be correct/incorrect? • * What are the limitations of the system? • * What kind of mistakes is the system likely to make? • * Is the system's performance good enough for... 	What If	<ul style="list-style-type: none"> • What would the system predict if this instance changes to...? • What would the system predict if this feature of the instance changes to...? • What would the system predict for [a different instance]?
How (global)	<ul style="list-style-type: none"> • How does the system make predictions? • What features does the system consider? <ul style="list-style-type: none"> • * Is [feature X] used or not used for the predictions? • What is the system's overall logic? <ul style="list-style-type: none"> • How does it weigh different features? • What rules does it use? • How does [feature X] impact its predictions? • * What are the top rules/features it uses? • * What kind of algorithm is used? <ul style="list-style-type: none"> • * How are the parameters set? 	How to be that	<ul style="list-style-type: none"> • How should this instance change to get a different prediction? • How should this feature change for this instance to get a different prediction? • What kind of instance gets a different prediction?
		How to still be this	<ul style="list-style-type: none"> • What is the scope of change permitted to still get the same prediction? • What is the [highest/lowest/...] feature(s) one can have to still get the same prediction? • What is the necessary feature(s) present or absent to guarantee this prediction? • What kind of instance gets this prediction?
		Others	<ul style="list-style-type: none"> • * How/what/why will the system change/adapt/improve/drift over time? (change) • * How to improve the system? (change) • * Why using or not using this feature/rule/data? (follow-up) • * What does [ML terminology] mean? (terminological) • * What are the results of other people using the system? (social)

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?'

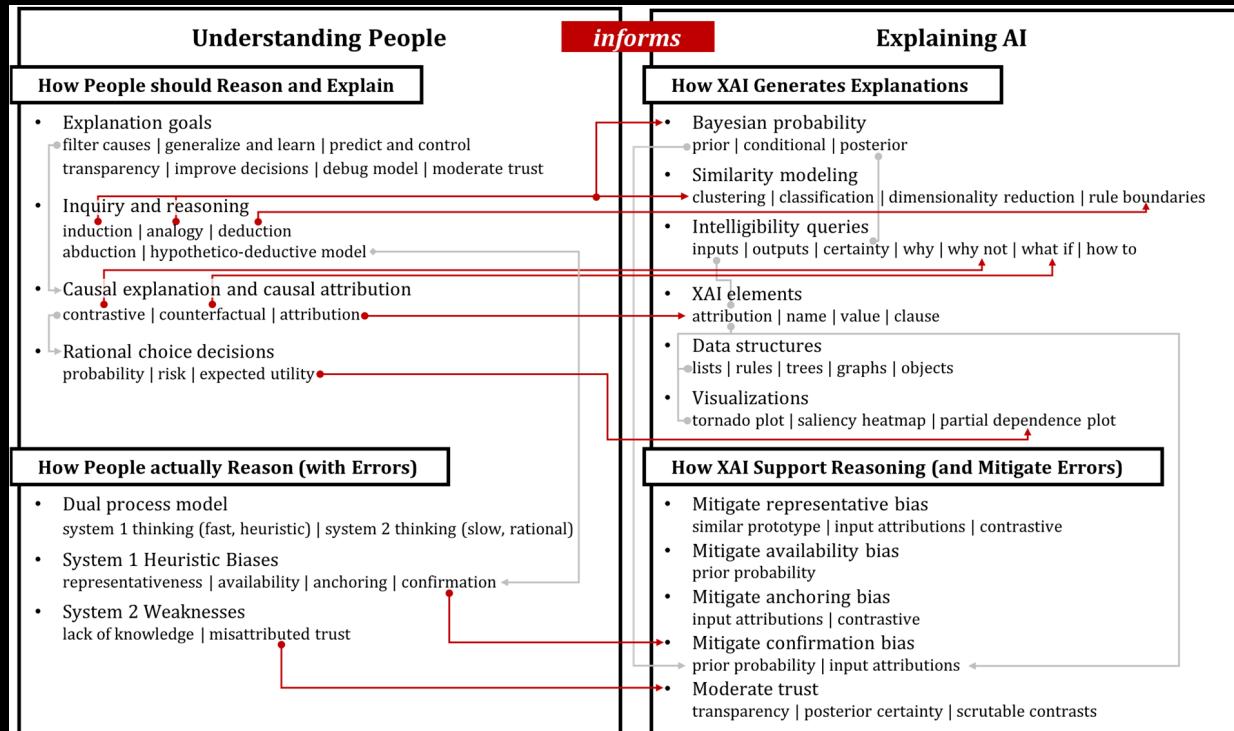
"Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation"

"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



Explanation for Human Reasoning

Heuristic Bias	Description	Strategies to overcome systematic errors [64]	XAI Strategies for Medical Decisions
Representativeness	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 .	<ul style="list-style-type: none"> - 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.	<ul style="list-style-type: none"> - 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]).	<ul style="list-style-type: none"> - Show input attributions for multiple outcomes to allow <i>contrastive reasoning</i>. - Facilitate counterfactual to test <i>How To</i> reduce the probability of primary diagnosis with Rules (e.g., aLIME, LORE). - Facilitate sensitivity analysis with <i>What If</i> explanations to test stability of primary hypothesis.
Confirmation	Collecting redundant information to confirm an existing hypothesis, instead of finding evidence of competing possibilities.	<ul style="list-style-type: none"> - Use hypothetical-deductive method to assess value and role of contemplated tests. - Try to disprove your diagnosis, consider conditions of higher prevalence. 	<ul style="list-style-type: none"> - Show Findings (input attribution) first, instead of Hypotheses (output posterior probability). <i>Insight: this is opposite to typical Machine Learning apps to show output uncertainty first.</i> - 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

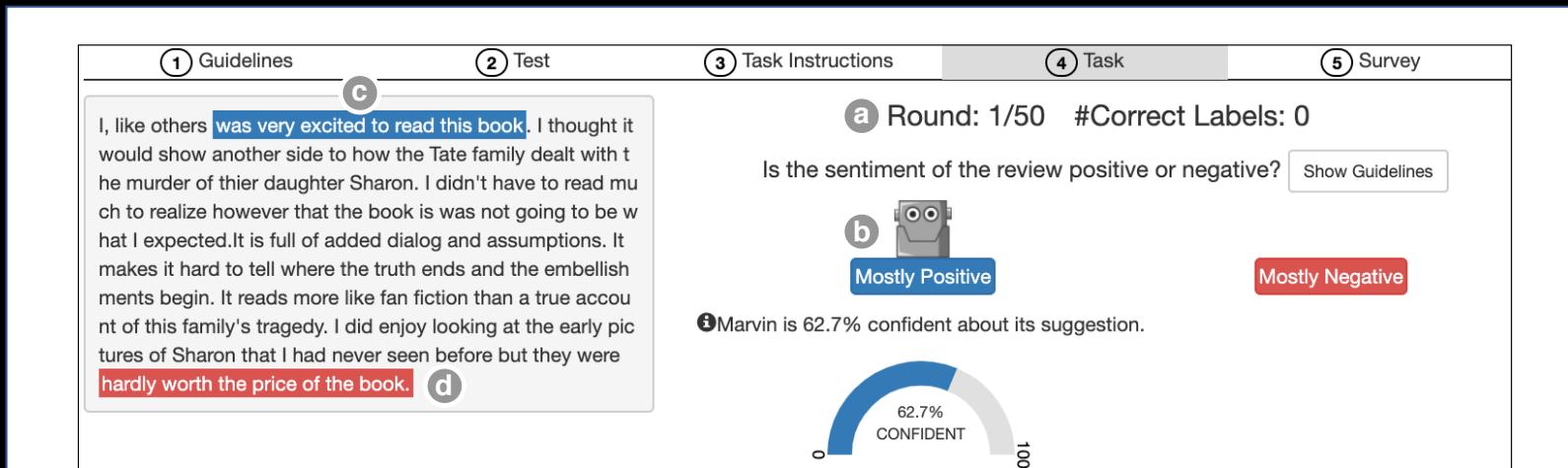


Figure 2: A screenshot of the Team (Adaptive, Expert) condition for the *Amzbook* reviews dataset. Participants read the review (left pane) and used the buttons (right pane) to decide if the review was mostly *positive* or *negative*. The right pane also shows progress and accuracy (a). To make a recommendation, 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

a Question 1 of 20 Your accuracy (so far): 0 / 20

John looks like a professional bodybuilder. He weighs 210 pounds and stands six feet tall, which is the size of an NFL linebacker. John looks huge when he enters the room. Years of gym time have clearly paid off in spades.

Which of the following, if true, weakens the argument?

- [A] John prefers to work out in the morning.
- [B] The average professional bodybuilder is considerably heavier and taller than the average NFL linebacker.
- [C] John weighed considerably less before he started working out.
- [D] John's father, brothers, and male cousins all look like professional bodybuilders, and none of them have ever worked out.

b



A progress indicator showing a semi-circle with a blue arc from the left and a red arc from the right, meeting at the center. The blue arc is larger, representing 68.50% confidence in answer D. The red arc represents 31.50% confidence in answer B.

I am 68.50% confident in answer D.
I am 31.50% confident in answer B.

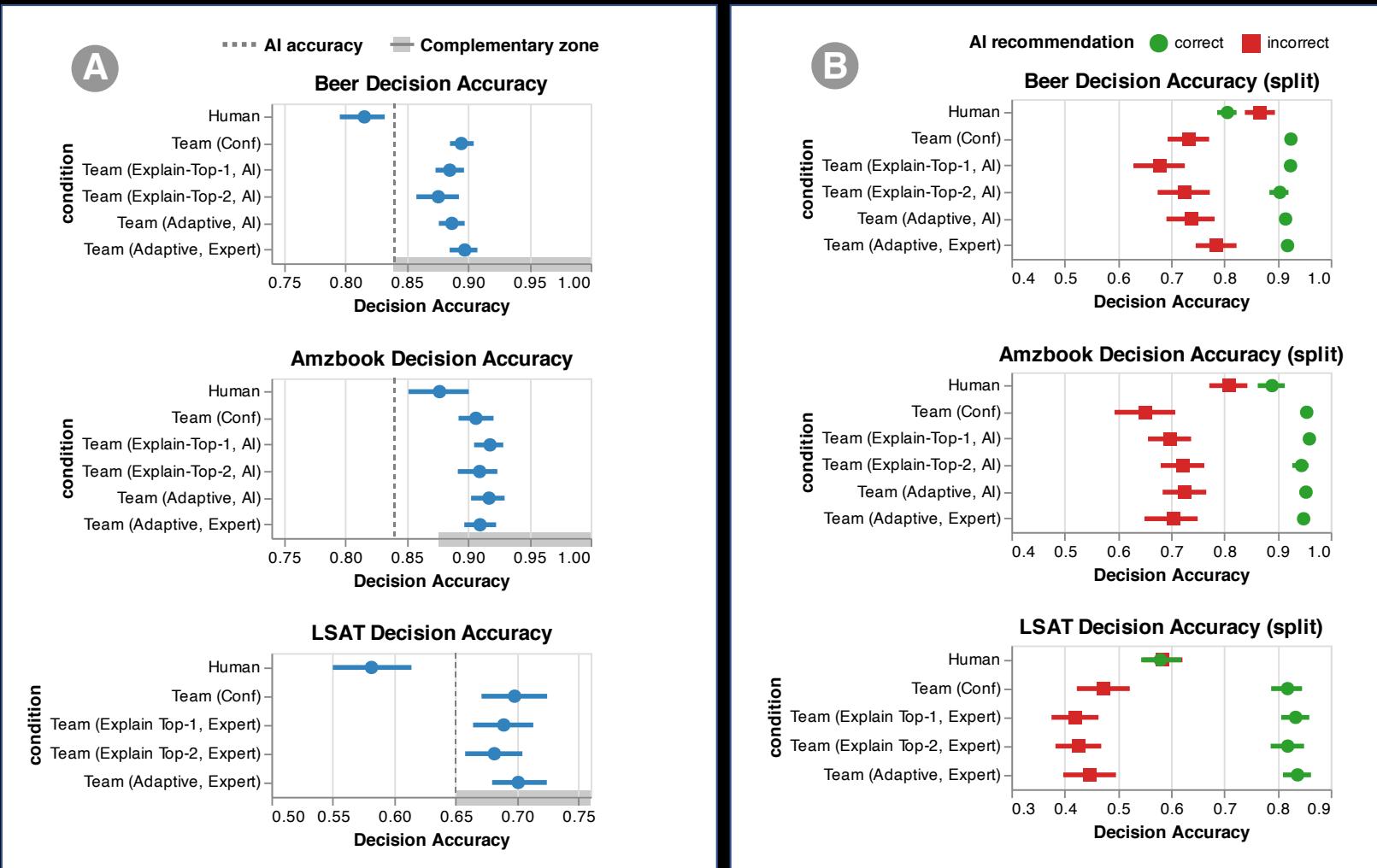
c **Reason for D:** John's family doesn't work out and still looks like professional bodybuilders. Years of gym time may not be the reason for John's size.

d **Reason for B:** John may be the size of an NFL linebacker, but if this statement is true, then John may not look like a professional bodybuilder.

NEXT

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



Takeaways

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

Next

Fairness