Can AI explanations skew our causal intuitions about the world? If so, can we correct for that?

Marko Tešić
Leverhulme Centre for the Future of Intelligence
University of Cambridge
marko.tesic375@gmail.com

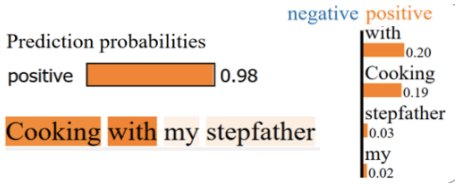
Causal XAI workshop
October 26, 2023

Role of explanation in AI

- Introduce transparency in (black-box) AI's decisions/recommendation
- Increase human understanding of AI's decisions/recommendations
- Promote trust in a (highly accurate) AI system
- Help calibrate trust in an AI system

Explanation in AI

- Top-down: start with a definition of an explanation; develop methods to output explanations



Feature contribution

Saliency maps

Explanation by example

For this defendant, our model would have made the opposite prediction (i.e., predict this defendant "will not reoffend") in the each of following cases:

- **Race**: If the defendant's Race had been **Hispanic** instead of White
- **Gender**: If the defendant's Gender had been **female** instead of male
- **Age**: If the defendant's Age had been **29** instead of 26
- **Prior Count**: If the defendant's Prior Count had been **1** instead of 2
- **Charge Name**: If the defendant's Charge Name had been **Driving with a Suspended License** instead of Grand Theft
- **Charge Degree**: If the defendant's Charge Degree had been **misdemeanor** instead of felony

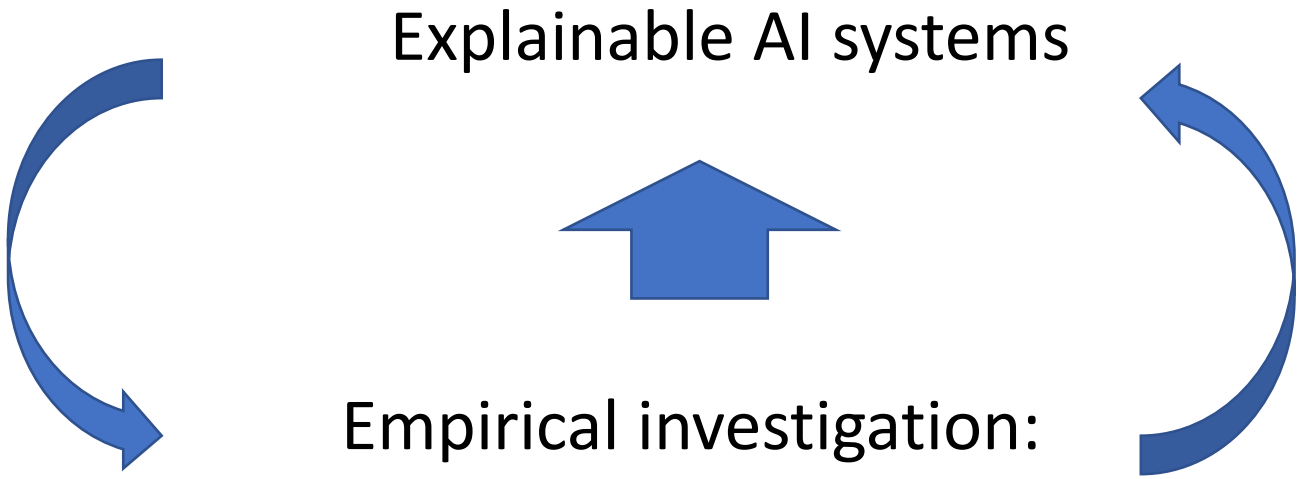
Counterfactuals

Empirical investigation:

which explanations are most suitable, which promote trust and when, etc.

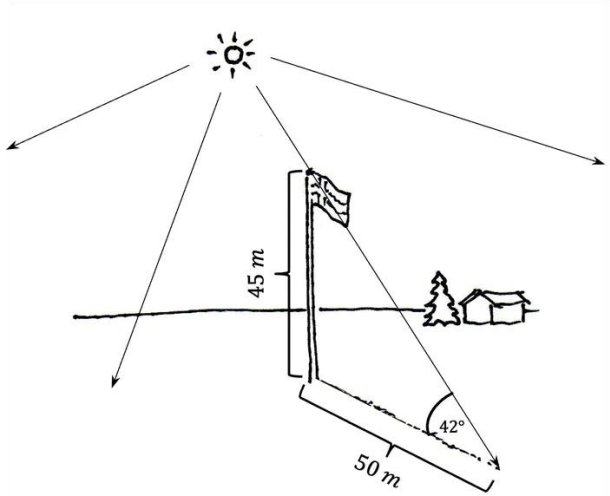
Try the other direction as well

- Bottom-up: start with an empirical investigation, which could then inform development of explainable AI systems



What types of explanations people provide?
What are some of the consequences of providing explanations?
Do explanations affect trust and when? etc.

Explanation in cognitive science, psychology, and philosophy that could inform XAI research



- Causal and asymmetric
- **Contrastive/counterfactuals**
  - Why this outcome instead of that?
- Selective (limit to how much info to include in an explanation)
- Social dimension of explanations (communicative acts) → trust
- Explanatory virtues (markers of explanatory goodness): simplicity, coherence, explanatory power, unification, etc.

Some research questions for the bottom-up approach

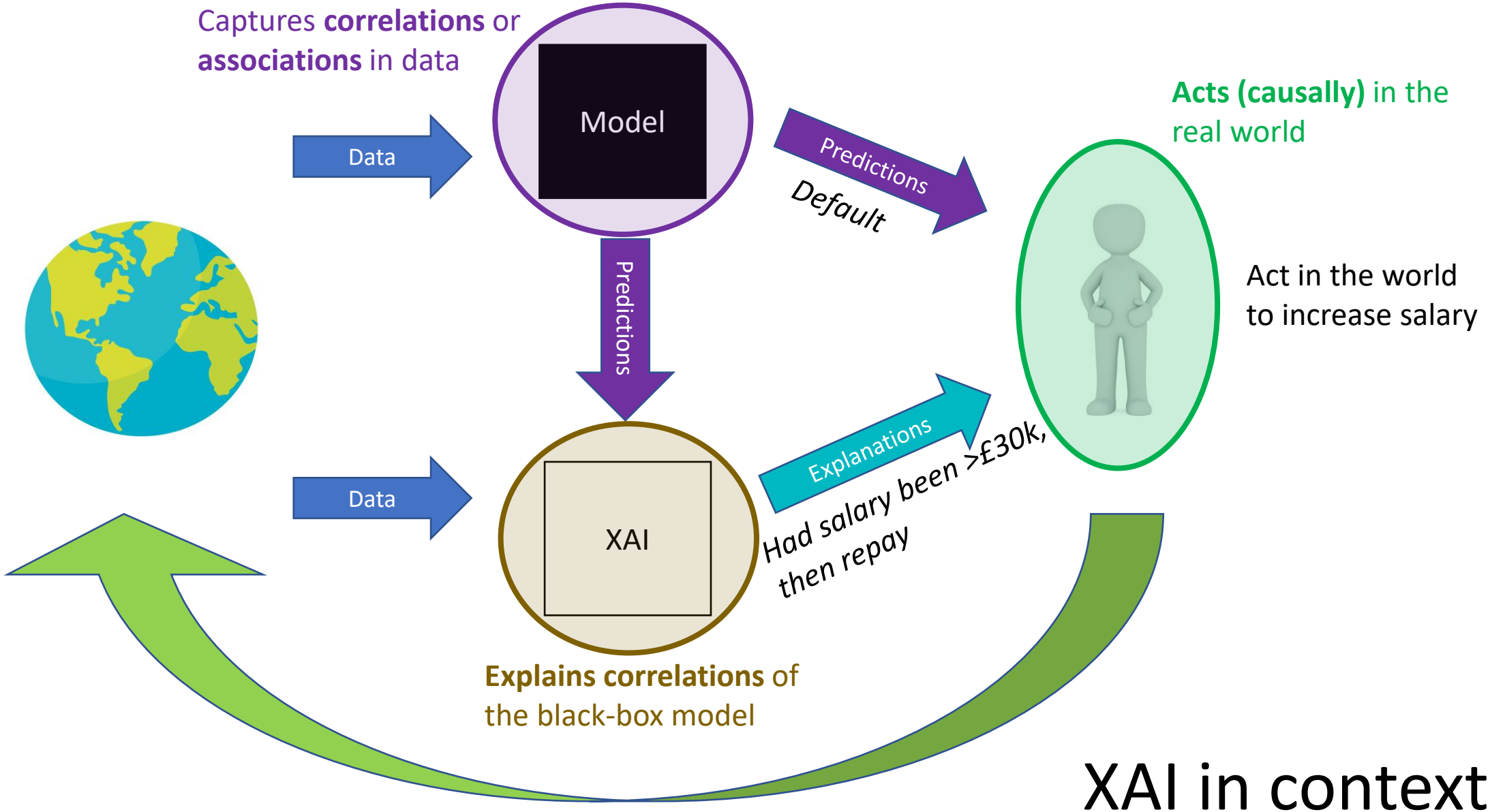
- What kinds of explanations people provide?

- Consequences of explanations on our beliefs in AI context?

- Does our perception of AI systems' reliability affect how we perceive explanations of AI decisions? Do they modify consequences of AI explanations?

Particularly important when AI explanations are inspired by work in psychology and cognitive science, and designers hope that they will yield the same positive effects as in non-AI contexts, without considering any potential side effects

Step back



XAI in context

Causal explanations and correlational AI systems

- People seem to provide and research in psychology focused on **causal** explanations of (often real-world) situations

Causal explanations and correlational AI systems

- AI systems capture **correlations**

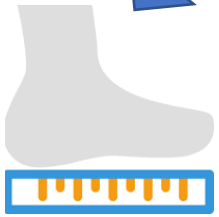
Published: 09 March 2000

Vision

Myopia and ambient night-time lighting

Karla Zadnik, Lisa A. Jones, Brett C. Irvin, Robert N. Kleinstein, Ruth E. Manny, Julie A. Shin & Donald O. Mutti

Nature **404**, 143–144 (2000) | Cite this article



Published: 13 May 1999

Myopia and ambient lighting at night

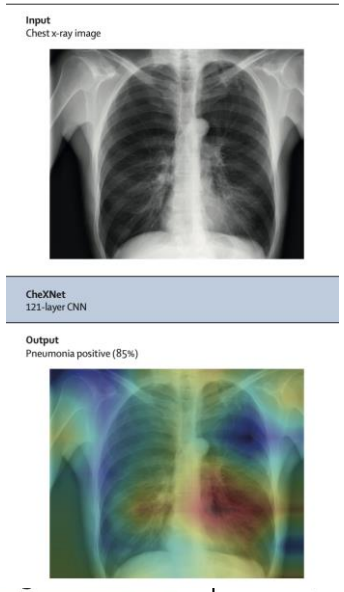
Graham E. Quinn, Chai H. Shin, Maureen G. Maguire & Richard A. Stone

Nature **399**, 113–114 (1999) | Cite this article

Potential consequences of directly applying insights from psychology to XAI

- Making AI explanations more human-like may lead them to think that they explain a certain **causal connection**
- Human tendencies amplified by making explanations more human-like esp. when there's room for interpretation
- Particularly prone to happen when using AI models to **teach** people about the real world
- Using desirable features such as selectivity may further reinforce causal intuitions

- Counterfactual (CF) explanations: we use them to communicate causality
- XAI uses CF explanations to explain decisions of AI systems



Counterfactual explanations in AI context

- AI **prediction**: Tom's salary is lower than £30K
- Why did the AI system predict that?
- Counterfactual **explanation**: If Tom had more than 500 LinkedIn connections, the AI system would have predicted ≥ £30k.
- Counterfactuals suggest **recourse/action**: maybe Tom should up his LinkedIn game
- Does the number of LinkedIn connections *affect* salary? Perhaps, but the AI system does not know that.
- Should Tom *increase* the number of his LinkedIn connections (i.e. act in the real-world) to increase his salary? Perhaps, but the explanation of AI prediction is not evidence enough to support that.

Research questions I

- Do our causal intuitions about factors the AI uses to make predictions become ***unjustifiably stronger*** if we are presented with counterfactual explanations of these predictions?

- If they do, is there anything we can do to correct for that?

Experiment 1

- How a range of factors influence salary?
- Three groups: Control, AI Prediction, AI explanation
- Total number of participants: 93
- Dependent variables:
  - Expectation, Confidence, and Action

Reminder: The AI system predicts that Tom's yearly salary is ***lower than £30k (< £30k)***. [AI Prediction and AI Explanation groups]

Factor: **Education level** [all three groups]

Explanation: If Tom had **had an advanced degree (e.g. masters)**, the AI system would have predicted that his **salary** was **higher than/equal to £30k (≥ £30k)**. [only the AI Explanation group]

Q. Would you expect that employees who **have an advanced degree (e.g. masters)** also have a **higher salary**? [Expectation question, same for all three groups]

Please rate your answer from 0 (No, not at all) to 100 (Yes, absolutely).

Q. How confident are you in your response? [Confidence question, same for all three groups]

Q. Assuming Tom has the resources (time, money, etc.), would you recommend he **starts an advanced degree (e.g. masters)** with the hope of increasing his **salary**? [Action question, same for all three groups]

Please rate your answer from 0 (not at all) to 100 (totally).

Your good friend Tom is looking to increase his **salary**. He's asked you for advice on how to best achieve that. [all three groups]

There are a range of factors that are related to a higher salary. You will now consider some of these factors. [only the Control group]

In your search for ways to help your friend you have found an **AI system** that can predict whether people's yearly **salaries** are *higher than/equal to £30k (≥ £30k)* or *lower than £30k (< £30k)*. [AI Prediction and AI Explanation groups]

The AI system uses a number of **factors** to make the prediction. You do not know, however, how much each factor is important for the AI system when it is making its predictions. [only the AI Prediction group]

The AI system uses a number of **factors** to make the prediction. The AI system also has an option to provide you with **explanations** regarding its predictions. [only the AI Explanation group]

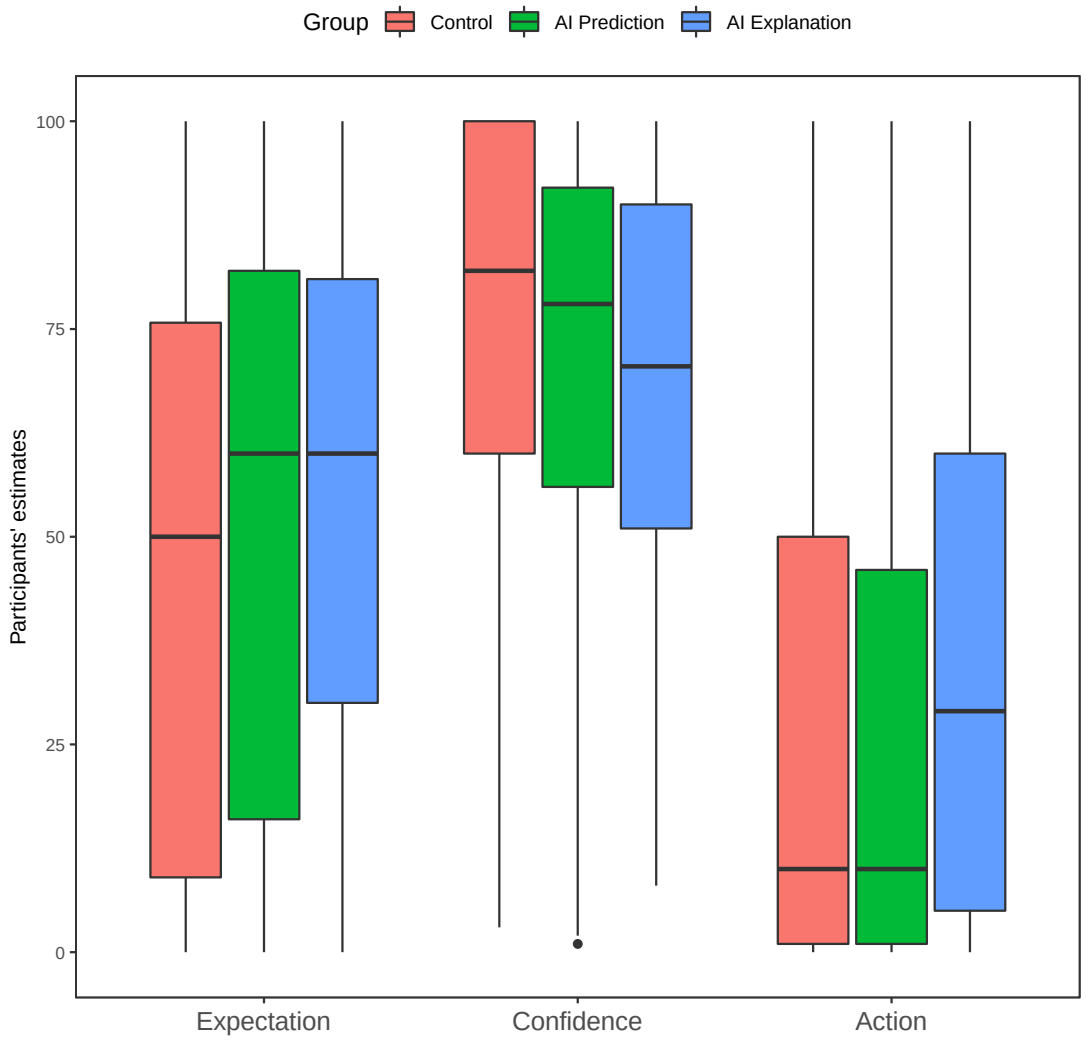
[NEXT PAGE]

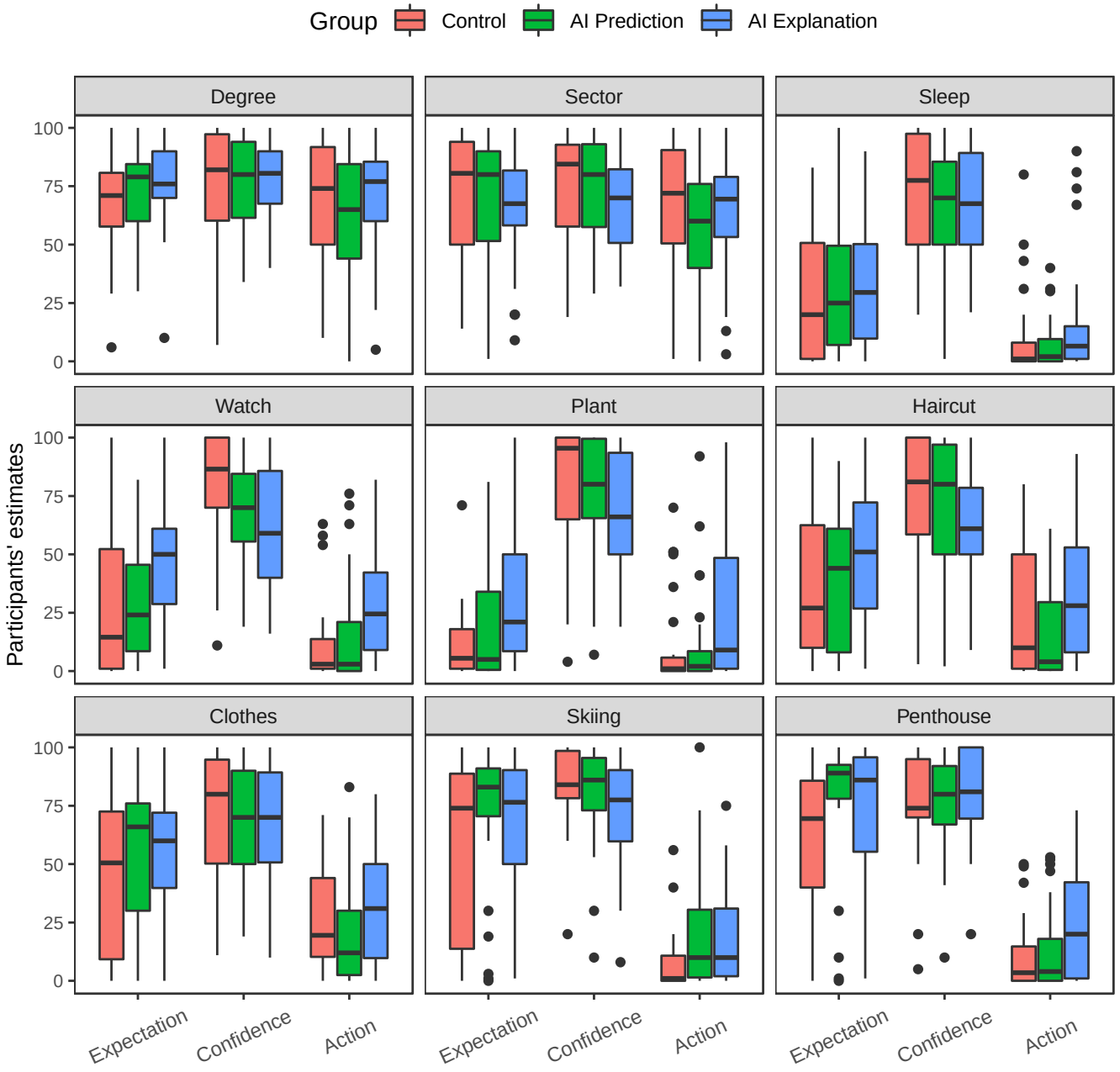
You input Tom's details for all factors into the AI system and it predicts that his yearly salary is ***lower than £30k (< £30k)***. [AI Prediction and AI Explanation groups]

The AI system now provides you with explanations with respect to each factor as to why it predicts that Tom's salary is lower than £30k (< £30k). [only the AI Explanation group]

You will now be asked questions related to the factors that the AI system used to make the prediction. [AI Prediction and AI Explanation groups]

Results


Group
Control
AI Prediction
AI Explanation
Participants' estimates
100
75
50
25
0
Expectation
Confidence
Action


Group
Control
AI Prediction
AI Explanation
Degree
Sector
Sleep
Watch
Plant
Haircut
Clothes
Skiing
Penthouse
Participants' estimates
100
75
50
25
0
Expectation
Confidence
Action

Experiment 2

- Same materials as in Experiment 1
- Same dependent variables
- 6 groups:
  - 3 the same as in Experiment 1
  - Another 3 were additionally told about an important note
- Total number of participants: 271

Important note

Correlation does not imply causation. Even though some factors may be highly **correlated** with higher salary that ***does not*** mean that they are **causing** higher salary.

In the AI Prediction group the note read:

Important note

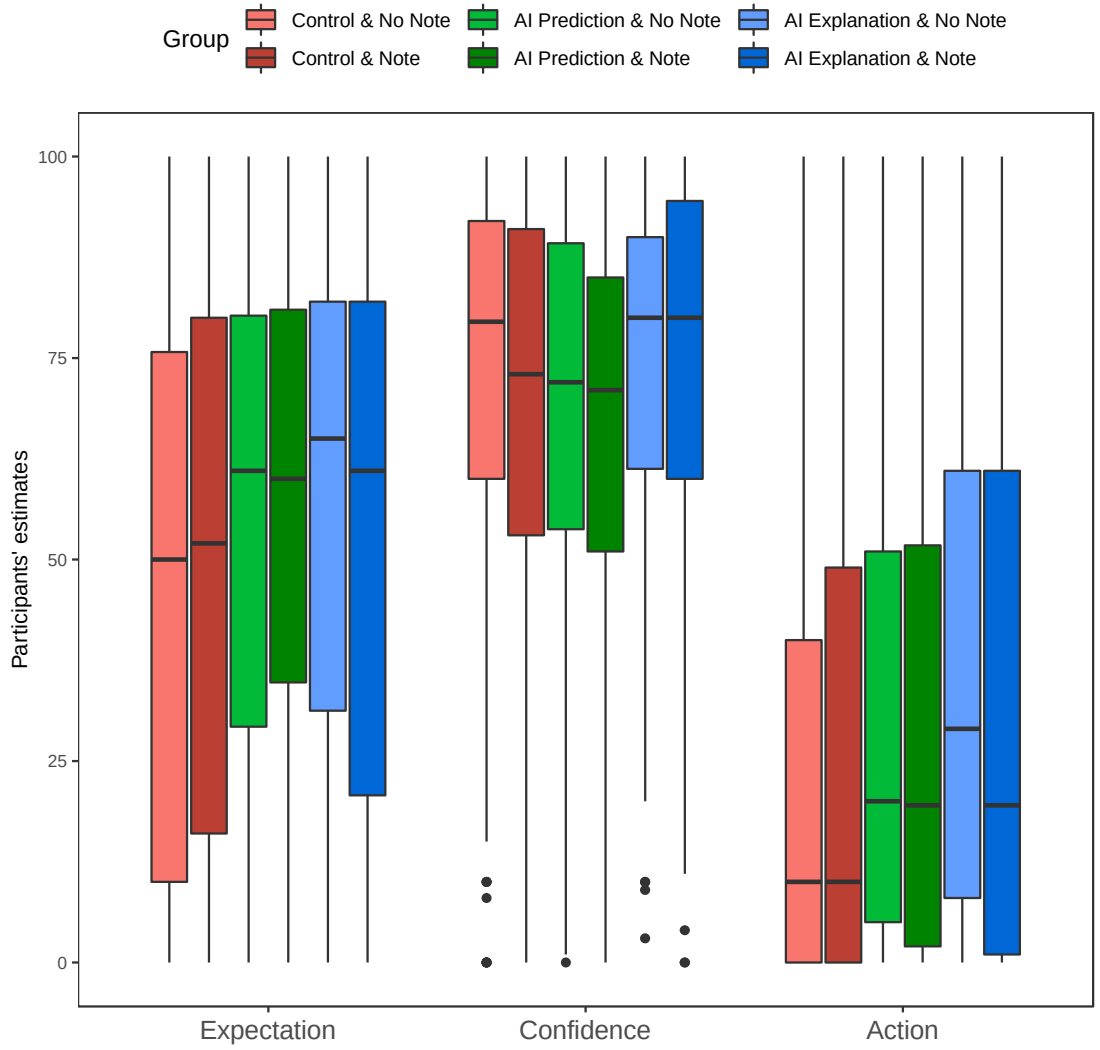
AI systems learn ***correlations*** in data. Even though the factors the AI system uses are potentially ***correlated*** with higher salary that **does not** mean that they are ***causing*** higher salary.

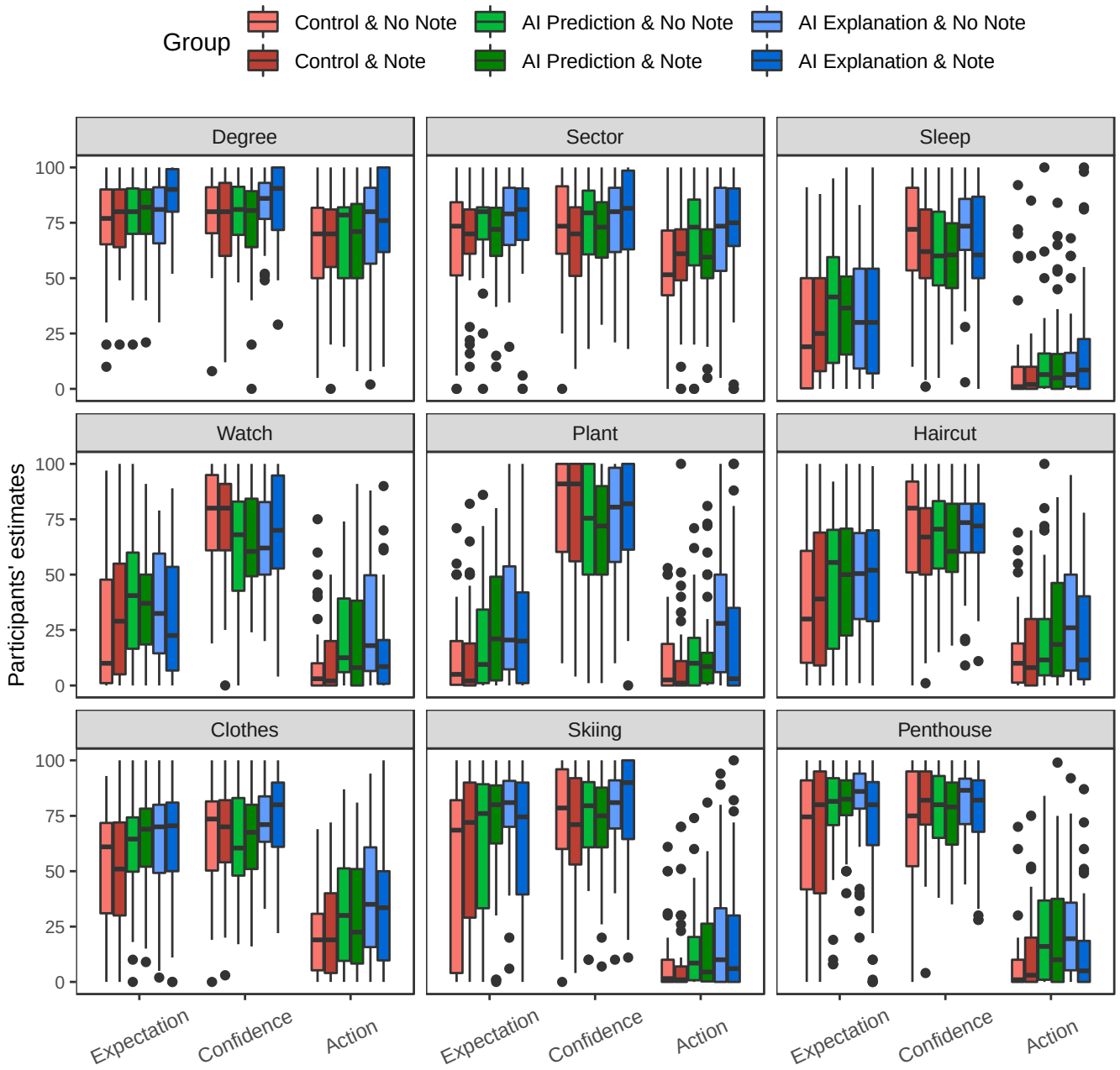
In AI Prediction condition the note read:

Important note

AI systems learn ***correlations*** in data. Even though the factors the AI system uses are potentially ***correlated*** with higher salary that **does not** mean that they are ***causing*** higher salary.
Similarly, the **explanations** of the AI systems' predictions are about the *correlations* the AI system has identified and not about which factors are *actually causing* higher salary.

Results

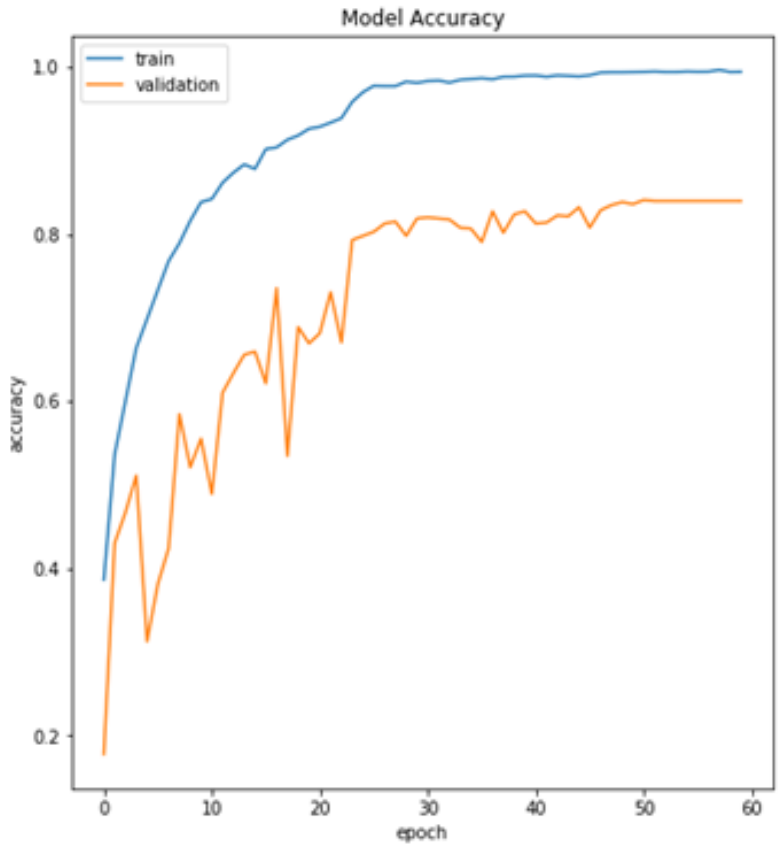

Group
Control & No Note
Control & Note
AI Prediction & No Note
AI Prediction & Note
AI Explanation & No Note
AI Explanation & Note
Participants' estimates
100
75
50
25
0
Expectation
Confidence
Action

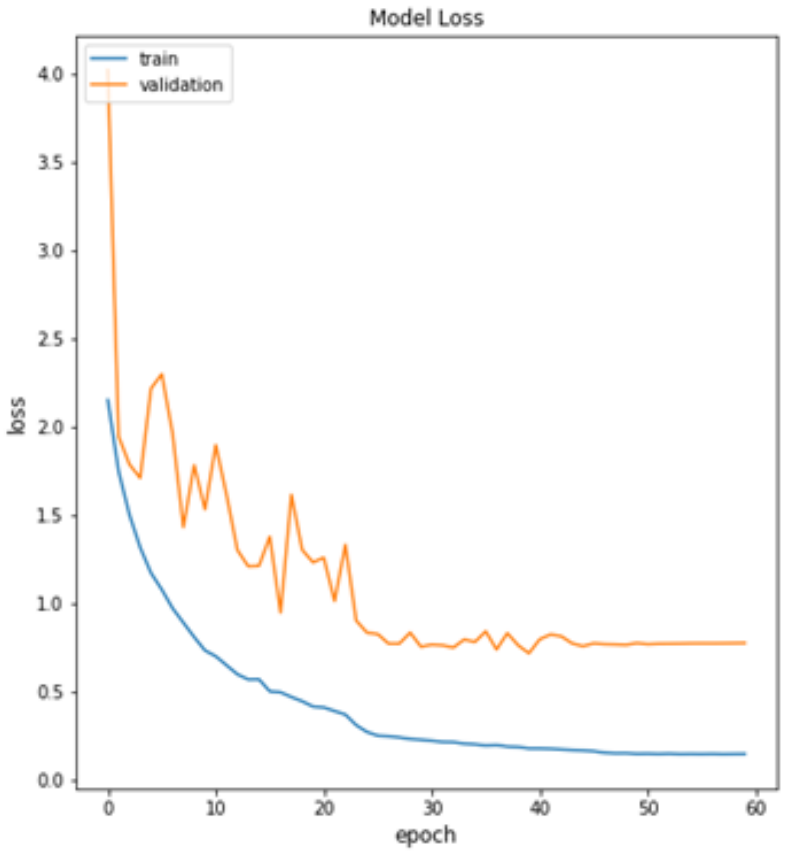

Group
Control & No Note
Control & Note
AI Prediction & No Note
AI Prediction & Note
AI Explanation & No Note
AI Explanation & Note
Degree
Sector
Sleep
Watch
Plant
Haircut
Clothes
Skiing
Penthouse
Participants' estimates
100
75
50
25
0
Expectation
Confidence
Action

Summary I

- Explanations of AI systems can unjustifiably impact lay people's causal intuitions
- It may be possible to correct for this by communicating to people that AI system's capture correlations and not causal relationships
- In general, we should be telling/reminding users of AI's associative and correlational nature
- We should be applying insights from cognitive science and psychology carefully as they may undesirably translate in the domain of AI

(Perceived) AI accuracy





Research questions II

- Does (perceived) AI accuracy play a role when providing lay users with counterfactual explanations?

- Is the undesirable effect of counterfactual explanations on our causal beliefs more pronounced when (we think) AI is highly accurate?

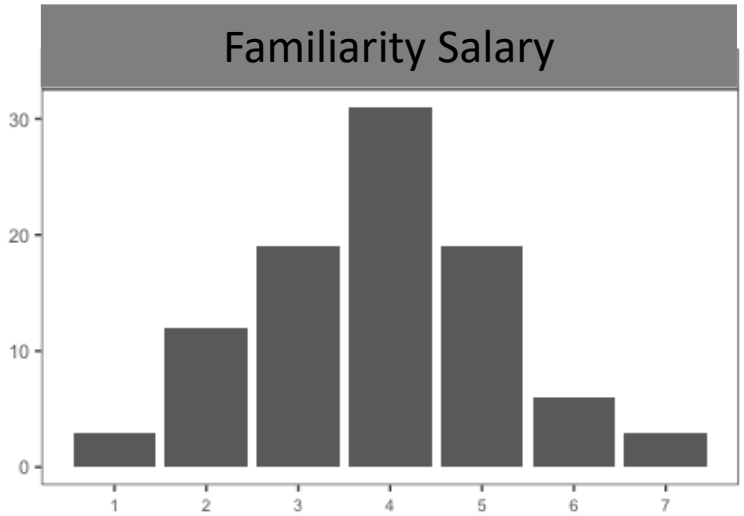
Experiment 3

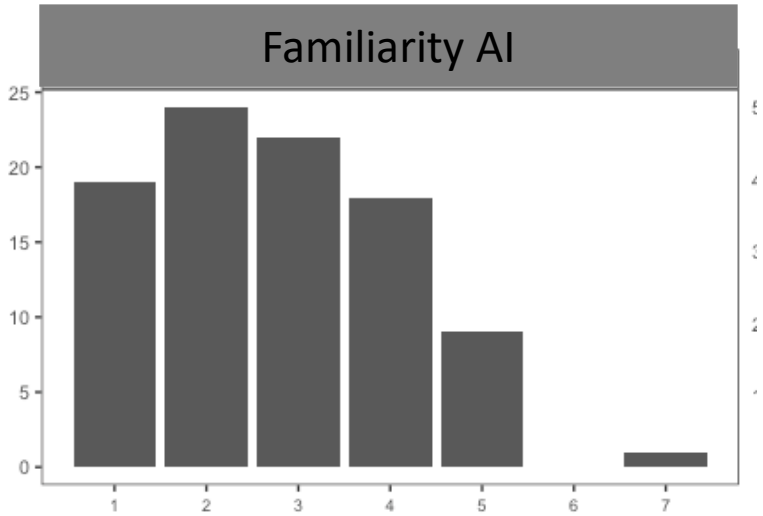
Three groups: Control,
AI Prediction, AI explanation
Unfamiliar scenario: predicting rice yield
-> easer to manipulate AI accuracy
-> replicate finding in a different domain

Familiarity

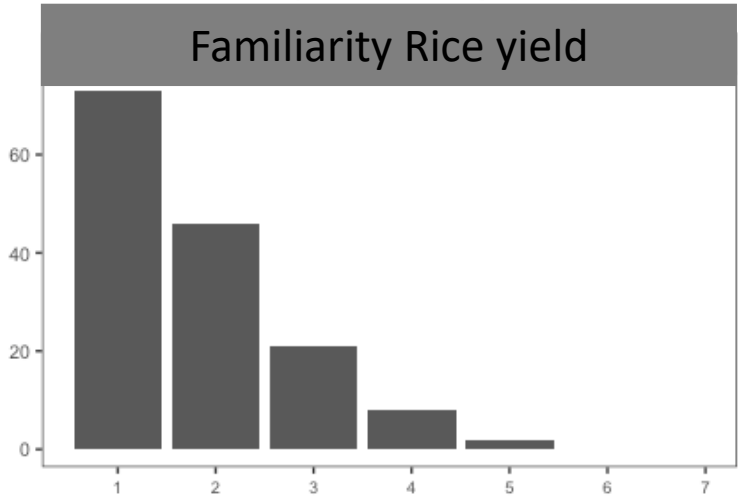
How familiar are you with the factors that may affect **salary/the yield of rice**?
Scale from 1 – Not at all familiar to 7 – Extremely familiar

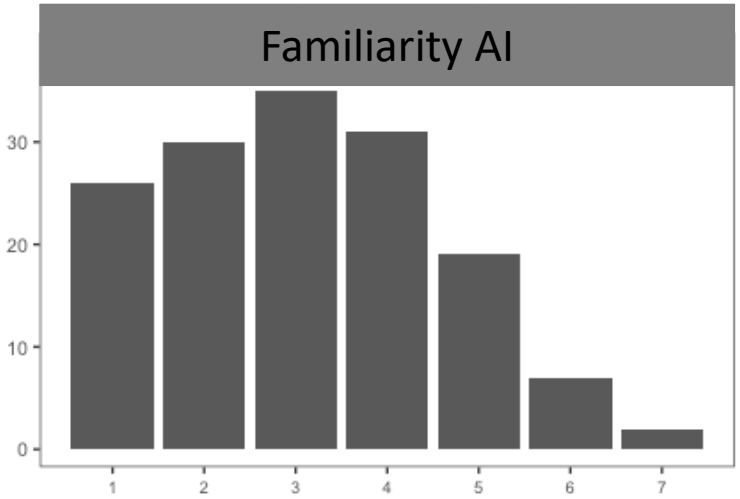
Experiment 1





Experiment 3





Experiment 3

Reminder: The AI system predicts that Lam's per acre rice yield is going to be ***lower than the average***.

Factor: **Shrimps**

Explanation: If Lam had **grown shrimps in the same field as rice**, the AI system would have predicted that his per acre **rice yield** is going to be **higher than/equal to average**.

Q. Would you expect that farmers who **grow shrimps in the same field as rice** have a **higher rice yield**?

Please rate your answer from 0 (No, not at all) to 100 (Yes, absolutely).

No, not at all — Yes, absolutely

0 10 20 30 40 50 60 70 80 90 100

Q. How confident are you in your response?

Not at all — Fully

0 10 20 30 40 50 60 70 80 90 100

Q. Assuming Lam has the resources (office space, money, etc.), would you recommend he **grows shrimps in the same field as rice** with the hope of increasing **the rice yield**?

Please rate your answer from 0 (not at all) to 100 (totally).

Not at all — Totally

0 10 20 30 40 50 60 70 80 90 100

Your good friend Lam recently started growing rice in Mekong River Delta, Vietnam. He is looking for ways to **increase the rice yield** on his rice fields.

Lam is cultivating deepwater rice, a common variety of rice in Vietnam grown in ponds with water over 20 inches deep. The picture below is an example of a rice field in Mekong River Delta, Vietnam (credit: J. Sammut, UNSW).

In your search for ways to help your friend you have found an **AI system** that can predict whether farmers' per acre **deepwater rice yields** are going to be *higher than/equal to the average rice yield* or *lower than the average* in Mekong River Delta, Vietnam.

The AI system uses a number of **factors** to make the prediction. The AI system also has an option to provide you with **explanations** regarding its predictions.

You input Lam's rice field details for all factors into the AI system and it predicts that his per acre rice yield is going to be ***lower than the average***.

The AI system now provides you with explanations with respect to each factor as to why it predicts that Lam's per acre rice yield salary is going to be lower than the average.

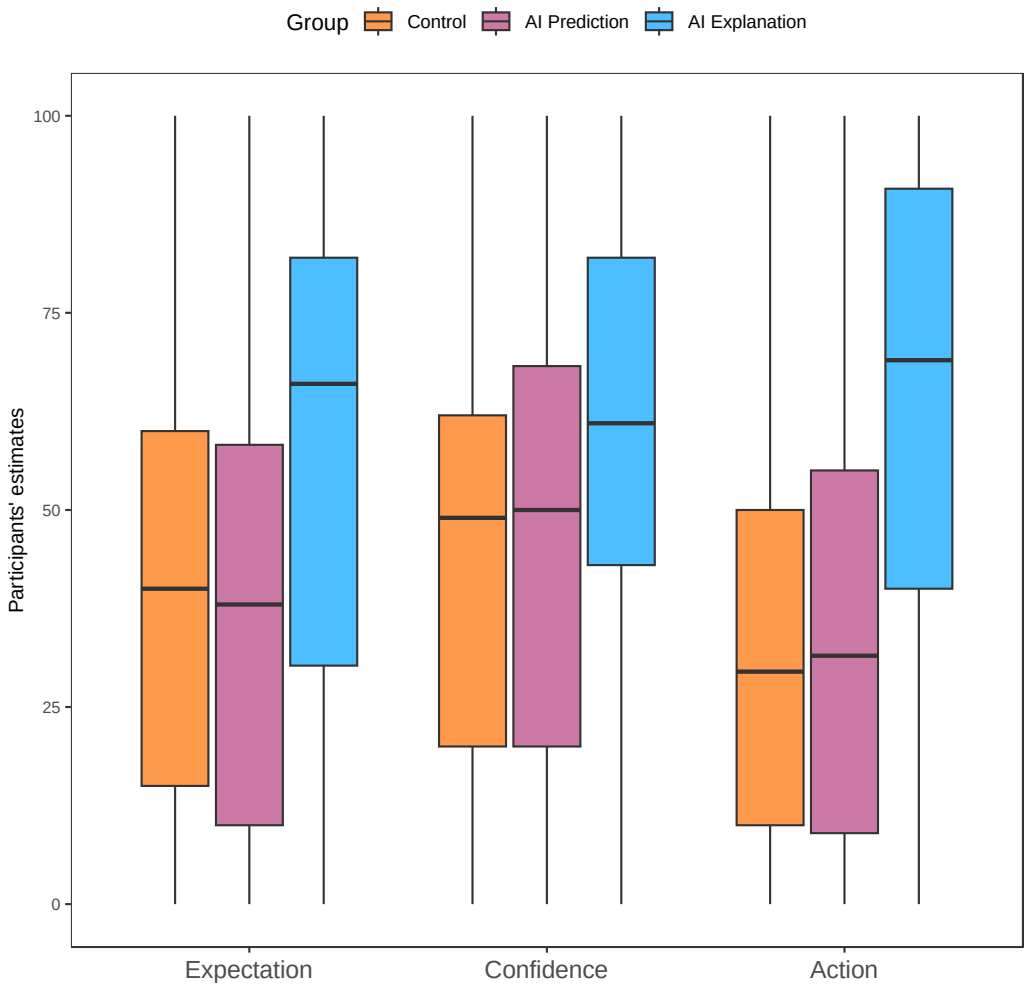
You will now be asked questions related to the factors that the AI system used to make the prediction.

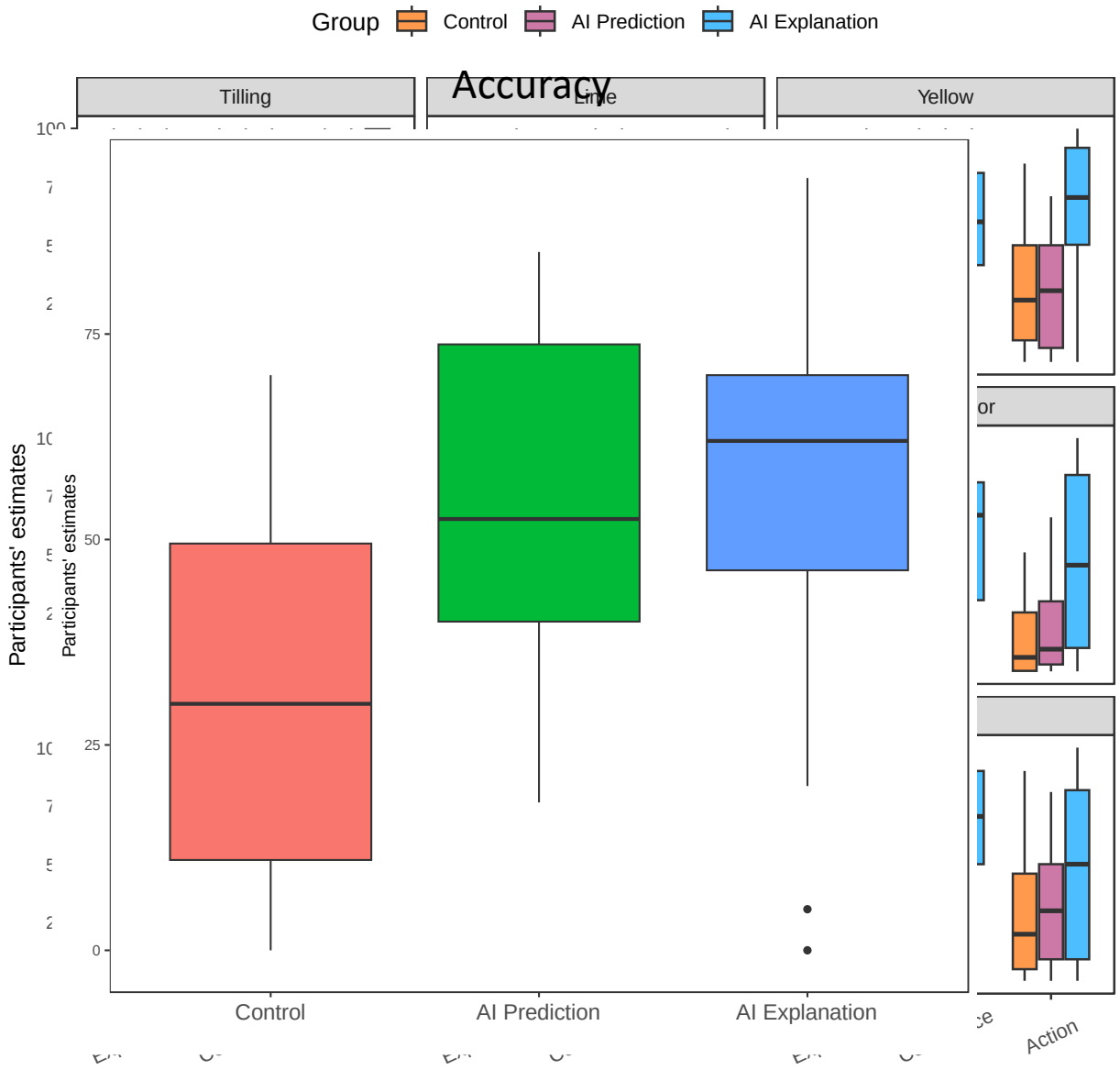
On the scale from 0 (extremely inaccurate) to 100 (extremely accurate), please rate how **accurate** you believe the **AI system** is in predicting whether the per acre *rice yield* will be *higher than/equal to the average* or *lower than the average*.

Extremely inaccurate — Extremely accurate

0 10 20 30 40 50 60 70 80 90 100

Results


Group
Control
AI Prediction
AI Explanation
Participants' estimates
100
75
50
25
0
Expectation
Confidence
Action


Group
Control
AI Prediction
AI Explanation
Accuracy
Tilling
Yellow
Participants' estimates
75
50
25
0
Control
AI Prediction
AI Explanation
Action

Experiment 4

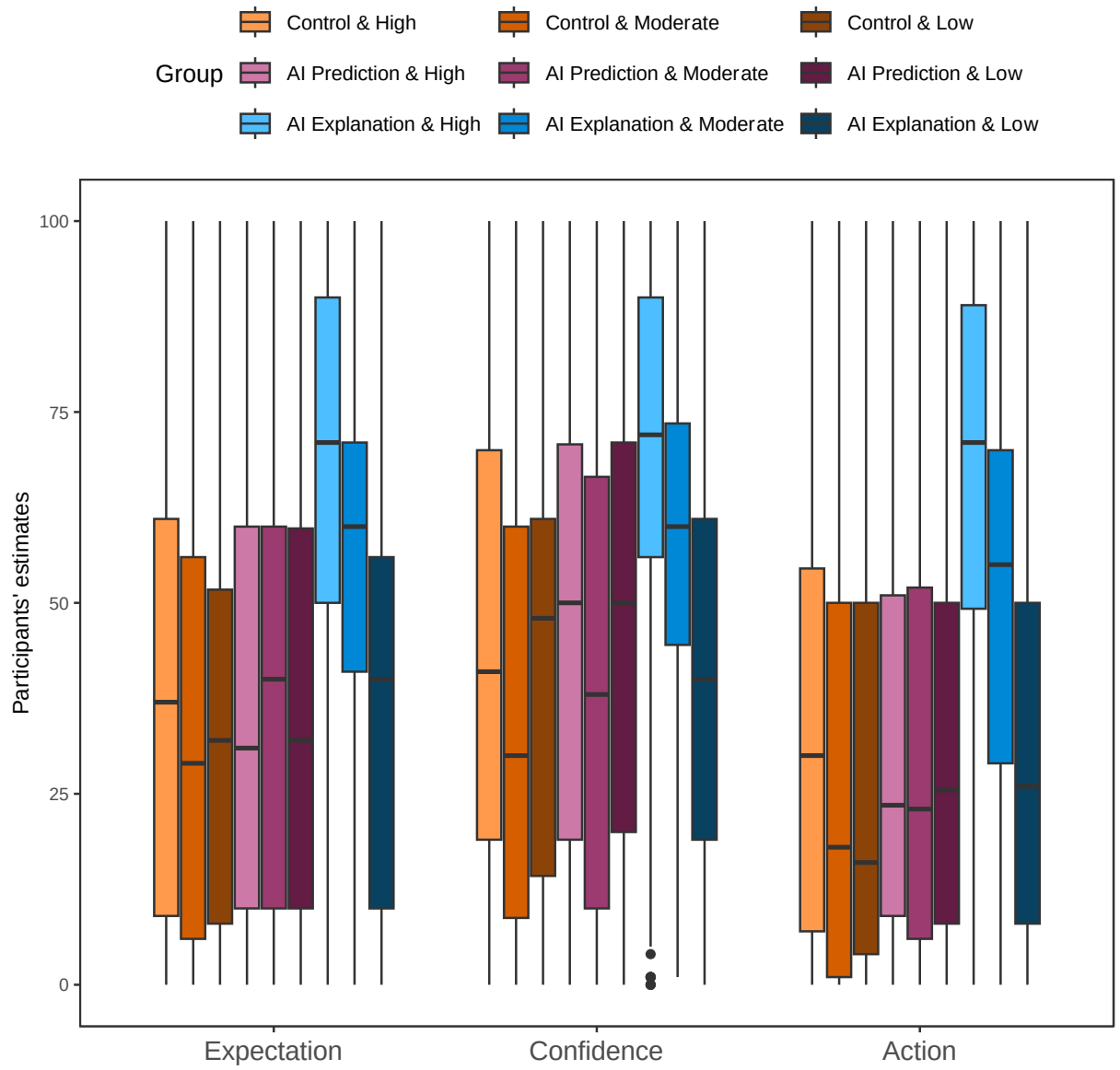
Manipulate accuracy: High, Moderate, Low
9 groups: Control AI Prediction, AI explanation X High, Moderate, Low

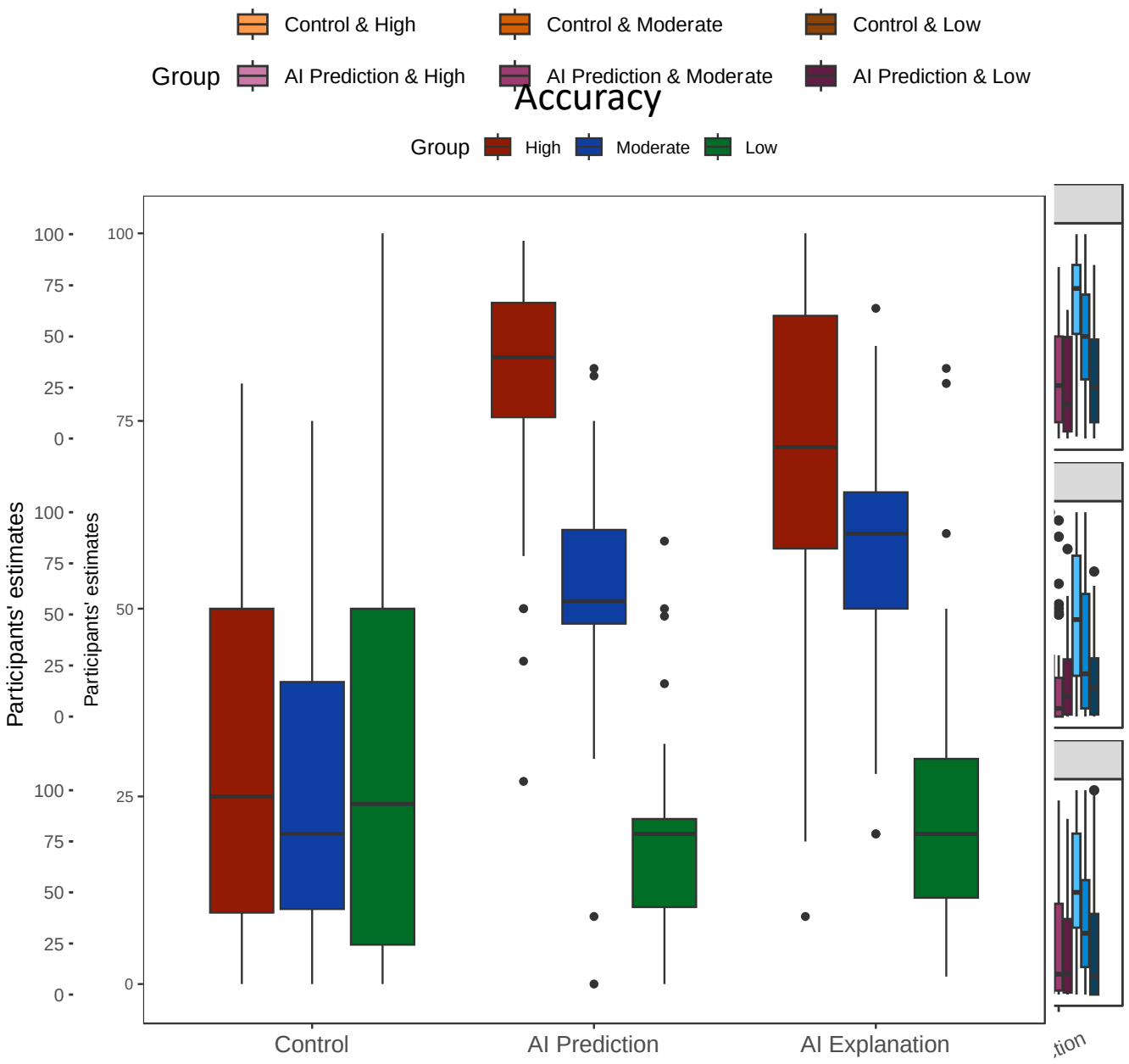
You also found out that robust testing over an extended period of time has shown that the AI system's **accuracy** is **very high**: it makes almost all predictions correctly, with only a very few incorrect ones.

You also found out that robust testing over an extended period of time has shown that the AI system's **accuracy** is **moderate**: it makes most predictions correctly, however a notable proportion of its predictions are incorrect.

You also found out that robust testing over an extended period of time has shown that the AI system's **accuracy** is **low**: it makes some predictions correctly, however the vast majority its predictions are incorrect.

Results


Group
Control & High
Control & Moderate
Control & Low
AI Prediction & High
AI Prediction & Moderate
AI Prediction & Low
AI Explanation & High
AI Explanation & Moderate
AI Explanation & Low
Participants' estimates
100
75
50
25
0
Expectation
Confidence
Action


Accuracy
Group
Control & High
Control & Moderate
Control & Low
AI Prediction & High
AI Prediction & Moderate
AI Prediction & Low
Group
High
Moderate
Low
Participants' estimates
100
75
50
25
0
Control
AI Prediction
AI Explanation

Summary II

- Explanation of AI systems affect our causal beliefs about the world, even when we don’t have strongly established causal beliefs about the domain
- AI accuracy amplifies the effect of explanations on our causal beliefs in unfamiliar domains
- AI accuracy does not seem to affect the predictive power of factors in unfamiliar domains

Conclusions

- Counterfactual explanations of predictive AI systems affect our causal beliefs about the real world
- They do so whether we are familiar or unfamiliar with the domain
- More accurate AI systems have a larger effect on our causal beliefs
- We may be able to correct for some of these effects
- Applying insights from cognitive science and psychology to AI should be done with caution