«Human-Centered Data Science»

Exercise 12

Lars Sipos

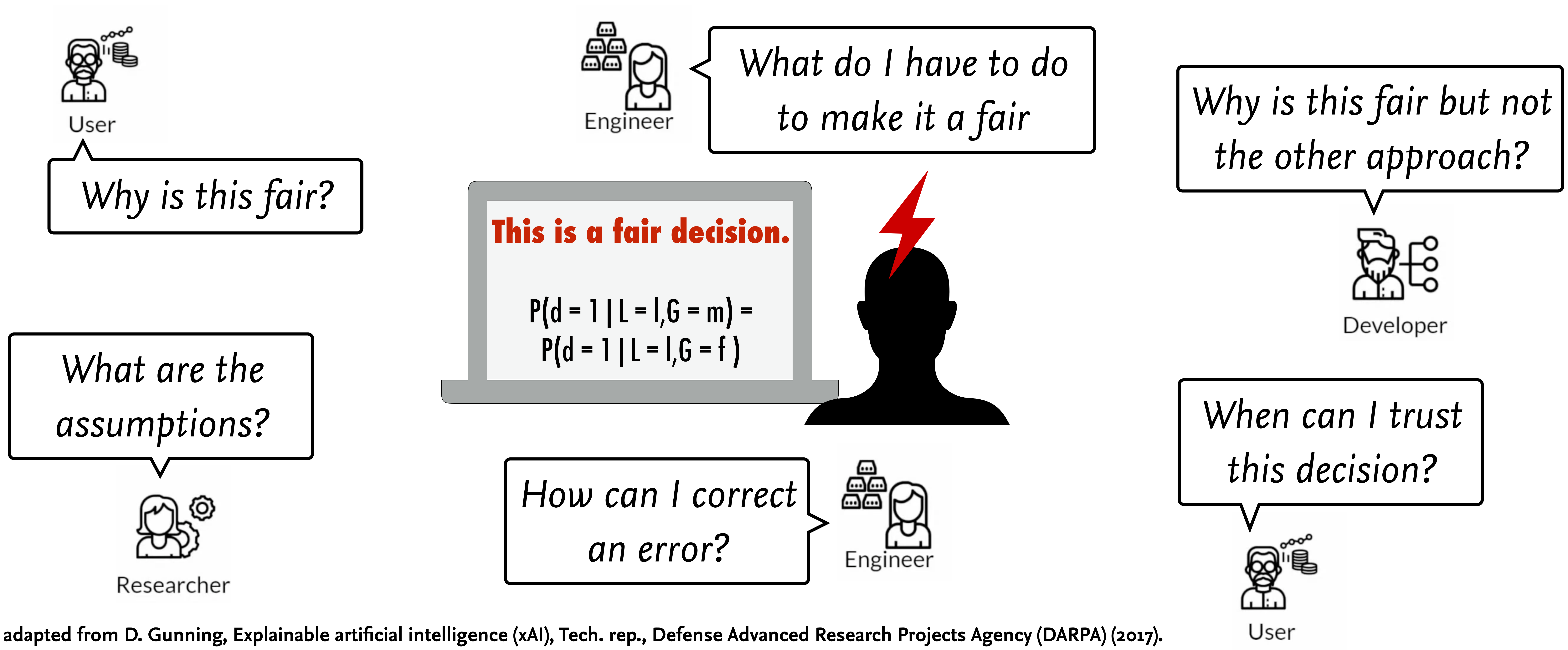
Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

12.07.2022

Introductory Question

What are **explanation needs**?



User

Why is this fair?

Engineer

What do I have to do to make it a fair

Why is this fair but not the other approach?

Developer

This is a fair decision.

$$P(d = 1 \mid L = l, G = m) = P(d = 1 \mid L = l, G = f)$$

What are the assumptions?

Researcher

How can I correct an error?

Engineer

When can I trust this decision?

User

Questions adapted from D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017). https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

What to Explain?

How Explanations

Why Explanations

Why-Not Explanations

What-If Explanations

How-To Explanations

What-Else Explanations

Six **common types of explanations** used in XAI system designs

Task: Discuss and draw how this would look like for the scenario from Assignment 6.

S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3, p. 45.

What to Explain?

How Explanations

Why Explanations

Why-Not Explanations

What-If Explanations

How-To Explanations

What-Else Explanations

Holistic representation of the machine learning algorithm to explain *How* the model works.

Possible Representations:

- » Model graphs
- » Decision boundaries
- » Collection of explanations from multiple individual instances

S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3, p. 45.

What to Explain?

How Explanations

Why Explanations

Why-Not Explanations

What-If Explanations

How-To Explanations

What-Else Explanations

Why a prediction is made for a particular input. Communicate what features in the input data, or what logic in the model has led to a given prediction by the algorithm.

Possible Representations:

» Model agnostic

» Model dependent

» Feature importance

S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3, p. 45.

What to Explain?

How Explanations

Why Explanations

Why-Not Explanations

What-If Explanations

How-To Explanations

What-Else Explanations

Demonstrate *why* a specific output was *not* in in the output of the system (also called **Contrastive Explanations**). Characterize the reasons for differences between a model prediction and the user's expected outcome.

Possible Representations:

» Model agnostic

» Model dependent

» Feature importance

S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3, p. 45.

What to Explain?

How Explanations

Why Explanations

Why-Not Explanations

What-If Explanations

How-To Explanations

What-Else Explanations

Demonstration of how different algorithmic and data changes affect model output given new input, manipulation of inputs, or changing model parameters.

Possible Representations:

» Recommendation of different *what-if* scenarios

» Exploration through interactive user control

» Parameter tuning

» Comparison with simpler data or models

S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3, p. 45.

What to Explain?

How Explanations

Why Explanations

Why-Not Explanations

What-If Explanations

How-To Explanations

What-Else Explanations

Spell out hypothetical adjustments to the input or model that would result in a different output (ideally to the user's specified output of interest).

Possible Representations:

» Ad-hoc

» Model-agnostic

» Model structure and internal values are not part of the explanation

» Interactive

S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3, p. 45.

What to Explain?

How Explanations

Why Explanations

Why-Not Explanations

What-If Explanations

How-To Explanations

What-Else Explanations

Present users with similar instances of input that generate the same or similar outputs from the model (also called *Explanation by Example*).

Possible Representations:

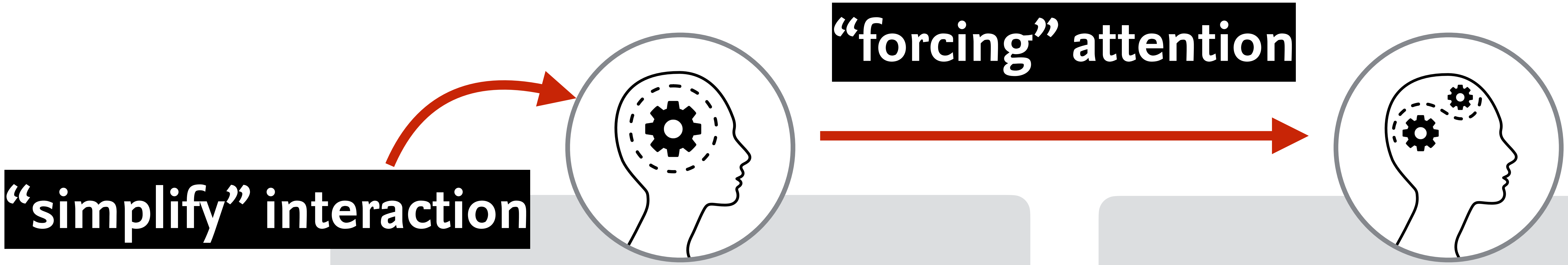
» Samples that are similar to the original input in the model representation space

S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3, p. 45.

Cognitive Forcing Functions

and study discussion

Improving XAI Outcome by Using Interventions



“simplify” interaction

“forcing” attention

Automatic Thinking

- » Intuitive understanding
- » Decisions are made instinctively, emotionally and unconsciously
- » Finds application in repeated and practiced actions

Reflective Thinking

- » Logical Reasoning
- » Decisions are based on a rational process that is consciously carried out
- » Finds application in deliberate, slow and strenuous actions

Kahneman, Daniel. Thinking, Fast and Slow. London: Penguin Books, 2012.

Cognitive Forcing Functions

An umbrella term for interventions that elicit “slow” thinking at the decision-making time.
These functions are often designed to explicitly disrupt the quick, heuristic (i.e., System 1) decision-making process.

Possible Design Approaches:

» Asking the person to make a decision before seeing the AI’s recommendation.

» Slowing down the process.

» Letting the person choose whether and when to see the AI recommendation.

Challenge: Studies (e.g., in visualization, education) suggest that people prefer simple interventions, even though they learn more and perform better with more complex interventions.

Buçinca, Zana, Maja Barbara Malaya, und Krzysztof Z. Gajos. „To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making“.
Proceedings of the ACM on Human-Computer Interaction_ 5, Nr. CSCW (2021): 188:1-188:21.

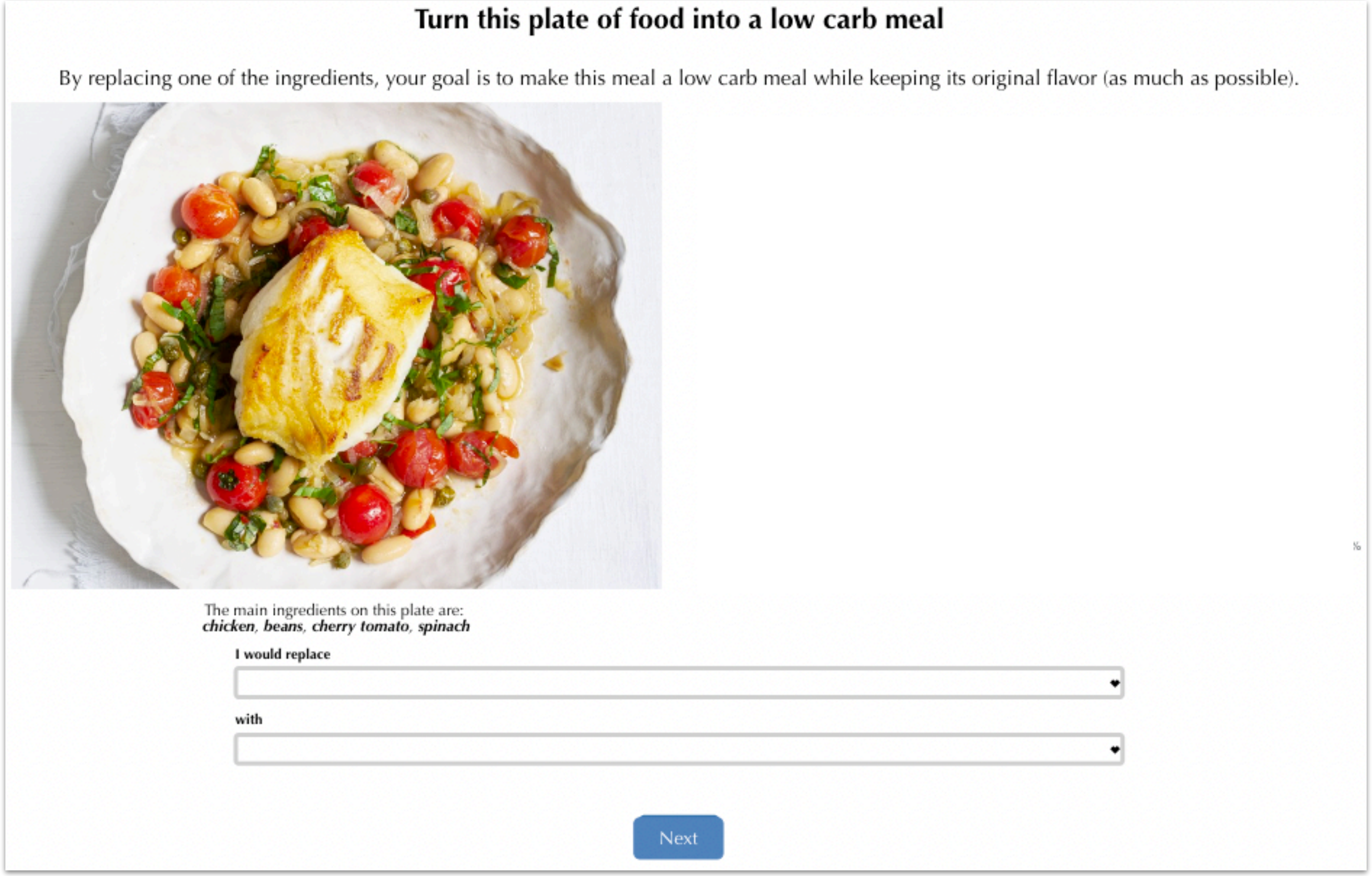
Study - Experimental Setup

Six conditions of **a nutrition task** for laypeople

- One No AI
- Two Simple Explainable AI (SXAI)
- Three Cognitive Forcing Functions (CFF)

The nutrition task was realized by a simulated AI that has a accuracy of 75%.

Included mistakes: AI did not “recognize” the ingredient highest in carbohydrates.



Turn this plate of food into a low carb meal

By replacing one of the ingredients, your goal is to make this meal a low carb meal while keeping its original flavor (as much as possible).

The main ingredients on this plate are:
chicken, beans, cherry tomato, spinach

I would replace

with

Buçinca, Zana, Maja Barbara Malaya, und Krzysztof Z. Gajos. „To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making“. Proceedings of the ACM on Human-Computer Interaction_ 5, Nr. CSCW (2021): 188:1-188:21.

Experimental Conditions

No AI (Condition 1)

Simple Explainable AI (SXAI)

» Condition 2a: Explanation

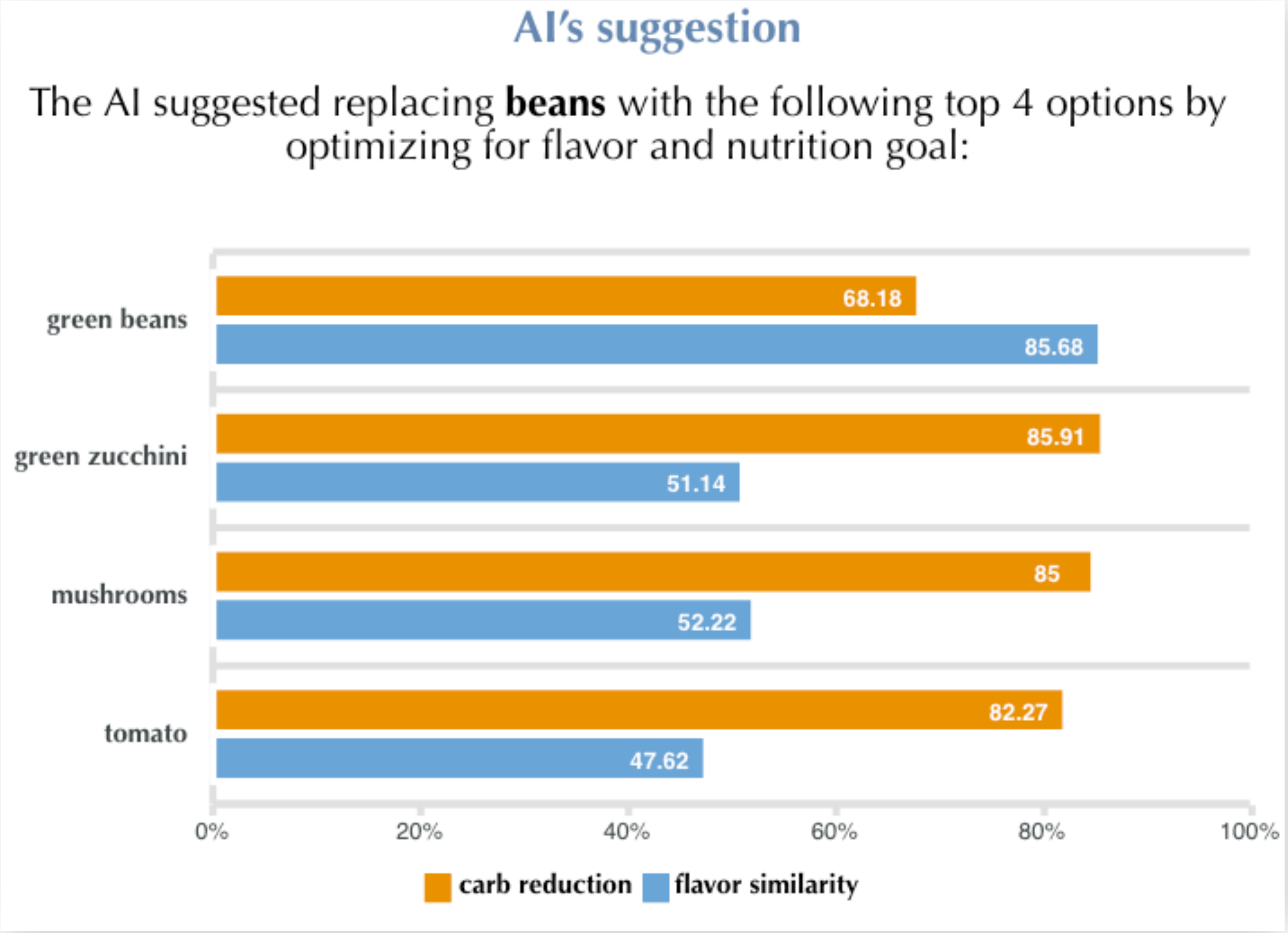
» Condition 2b: Uncertainty

Cognitive Forcing Functions (CFF)

» Condition 3a: On demand

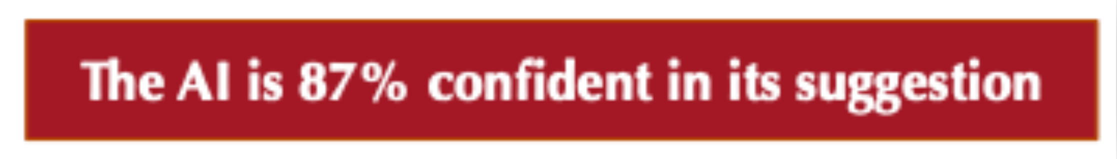
» Condition 3b: Update

» Condition 3c: Wait

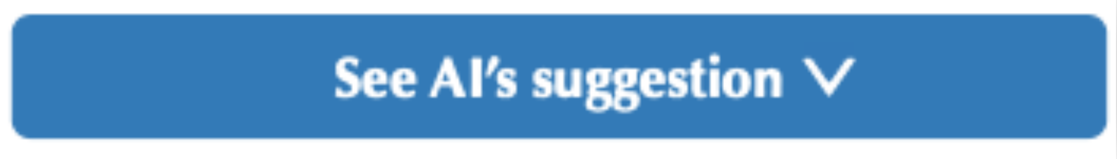


AI's suggestion

The AI suggested replacing **beans** with the following top 4 options by optimizing for flavor and nutrition goal:



The AI is 87% confident in its suggestion



Condition 1 then 2a, decision update possible

The AI is processing the image

Followed by condition 2a

Buçinca, Zana, Maja Barbara Malaya, und Krzysztof Z. Gajos. „To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making“. Proceedings of the ACM on Human-Computer Interaction_ 5, Nr. CSCW (2021): 188:1-188:21.

Experimental Setup

H1a: Compared to simple explainable AI approaches (SXAI), cognitive forcing functions (CFF) will improve the performance of human+AI teams in situations where the AI's top prediction is incorrect.

H1b: Compared to simple explainable AI approaches, cognitive forcing functions will improve the performance of human+AI teams.

H2: There will be a negative correlation between the self-reported acceptability of the interface and the performance of human+AI teams in situations where the AI's prediction is incorrect.

Study Procedure and Participants

Amazon Mechanical Turk with nine conditions (three unreported) and 26 questions (split into two blocks a 13 questions)

Each participant interacted with two of the nine conditions

260 Participants who needed 15 min on average to complete the study, 199 participants were considered

Reward for ‘top performer’ of each batch of 3 USD (by 2.5 USD for task)

Presented Results

Objective Measures

Performance on all task instances where top AI predictions are correct or incorrect

» both SXAI and CFF improved participants' performance compared to the no AI baseline

» no significant differences SXAI and CFF

Performance of task instances where top AI predictions were incorrect

» CFF improved the objective metrics significantly more compared to SXAI

» **but** performance of participants in no AI category was significantly higher than that of participants' in either CFF or SXAI (however, participants in CFF overrelied significantly less and made significantly more correct decisions than participants in SXAI)

Subjective Measures

» Participants reported higher trust in the AI in SXAI conditions compared to CFF (not significantly)

» Participants preferred completing the task significantly more with AI assistance (SXAI, CFF) than without it (no AI)

» Participants found the no AI condition to be significantly more mentally demanding than CFF and SXAI conditions

» Participants perceived the system as significantly less complex in SXAI conditions compared to CFF

Need for Cognition

Need for Cognition (NFC) is a stable personality trait that reflects how much a person enjoys engaging in cognitively demanding activities .

Participants with a **high-NFC** seek out more information and process it more deeply

Low-NFC participants are more likely to resort to cognitive shortcuts such as relying on the surface cues to assess the information such as the authority or celebrity of the source of the information, or the aesthetics of the presentation.

Ethical Challenge

> In the context of human-AI collaboration on decision making
>
> we considered individuals with high NFC to be the already privileged group and
>
> we investigated whether cognitive forcing functions were equally effective for people with low NFC
>
> as they were for people with high NFC
>
> or whether they increased the performance gap (i.e., inequality) between the two groups.

Results - Objective Measures

- High-NFC participants demonstrated significantly higher overall performance than low-NFC
- Incorrect model predictions: high-NFC participants benefited from CFF as they significantly improved both their overall and carb source detection performance.
- Incorrect model predictions: low-NFC participants benefited from CFF in carb source detection performance only
- High-NFC participants in CFF conditions over-relied on the AI significantly less for carb source detection than those in simple explainable AI conditions

Results - Subjective Measures

- Low NFC participants generally found the task significantly more mentally demanding than high NFC participants.
- Low NFC participants found the system to be significantly more complex than high NFC participants
- Low NFC participants reported on average higher trust than high NFC participants, albeit not significantly.
- High NFC participants reported significantly higher trust for SXAI conditions compared to CFF, and they also perceived them as less complex.

✓ H1a: Cognitive forcing functions reduced overreliance on AI compared to the SXAI approaches.

✗ No support H1b: There was no significant difference in performance between SXAI approaches and CFF.

✓ Support H2: There is a trade-off between the acceptability of a design of the human-AI collaboration interface and the performance of the human+AI team.

Insights

Human’s **cognitive motivation influences the effectiveness** of XAI solutions.

Future interventions might be tailored to **account for the differences in intrinsic cognitive motivation**: stricter interventions might benefit more and still be accepted by people with lower intrinsic cognitive motivation.

Developing adaptive strategies for providing different interventions based on models that predict the performance of human+AI teams on particular task instances.

Buçinca, Zana, Maja Barbara Malaya, und Krzysztof Z. Gajos. „To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making“. Proceedings of the ACM on Human-Computer Interaction_ 5, Nr. CSCW (2021): 188:1-188:21.

Discussion

What does this imply for the implementation of your explanation system?

Next Time

you will have …

1. actively participated in the lecture
2. submitted the sixth programming assignment
3. peer-reviewed assignment 5

Have fun!