Evaluatory psychological framework for XAI explanations

Matija Franklin
Causal Cognition Lab, UCL

Background

In 2016, DARPA launched the Explainable AI program

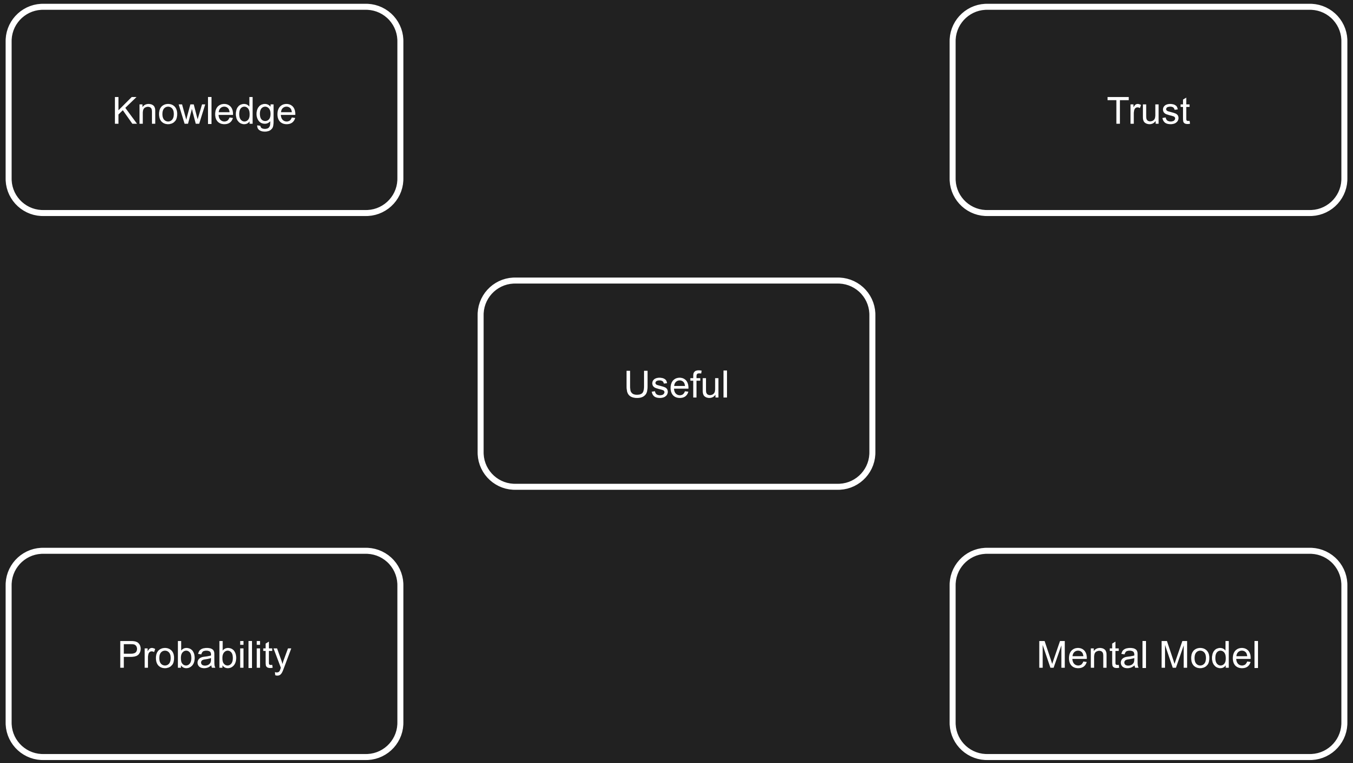
(1) How to produce more explainable models

(2) How to design the explanation interface

(3) How to understand the psychological requirements for effective explanations

How to measure whether an explanation is effective?

The Framework



Knowledge

Trust

Useful

Probability

Mental Model

Knowledge

Knowledge

- Provide knowledge
- Standard AI would allow people to make inferences from its decisions and detect patterns
- The learning from XAI can be more direct
- Measured through learning - does people’s behaviour change when the XAI tool is removed or does it stay the same

Trust

- Increases trust in AI

- Bertram Malle and Daniel Ullman. (2021) "A multidimensional conception and measure of human-robot trust."

- Predictive of whether people choose to use AI

- Alignment between AI suggestions and human decisions

- Self-reported trust

Usefulness

- Positively impacts performance i.e., increases accuracy

Useful

- Behavioural measure of domain-specific task performance
- Compared to doing the task without an explanation
- Better at the upper end or less mistakes?

Probability

- Update the receiver's estimation about the probability of events occurring

- Measure of confidence or certainty in the users’ own predictions or the AI’s predictions

- Sliding scale out of 100

Probability

Mental Model

- Change the receiver’s mental models about:
  - The task
  - Broader domain of expertise
  - The AI

- This in turn should influence other aspects of their cognition e.g., whether or not the blame AI for certain mistakes

Mental Model

What do we get?

A paradigm

Comparing human-XAI to human-AI (and no AI) “teams” on different tasks

Training Phase: measure of “baseline performance”

Test Phase:
Part 1: Participants given XAI tools, measure of "performance",

Part 2: measure of "knowledge" (removal of AI)

Evaluation Phase: trust, causal models, individual differences

Predictive framework

- What explanations are appropriate for different
  - Domains
  - Tasks
  - Individuals e.g., seniority, capacity
  - Ordering

Useful for optimising the deployment of XAI methods

Adaptive Explainable Artificial Intelligence (AXAI)

A recommender system capable of predicting what the preferred explanations would be for a specific domain-expert on a particular task.

Achieving this with a simple model is possible given that the data contains high level, well-understood variables.

Further, by distilling the data through the paradigm, a simple model can also produce salient explanations of its own process by default — a kind of ‘meta-interpretability'.

Thank you!

matija.franklin@ucl.ac.uk
@FranklinMatija