Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Post-hoc Interpretability: Understanding Human Cognitive Abilities

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 30, 2022

Lecture Overview

Recap

Explaining vs. Interpreting (Interpretation, Eliciting Explanation Strategies by using Participatory Design)

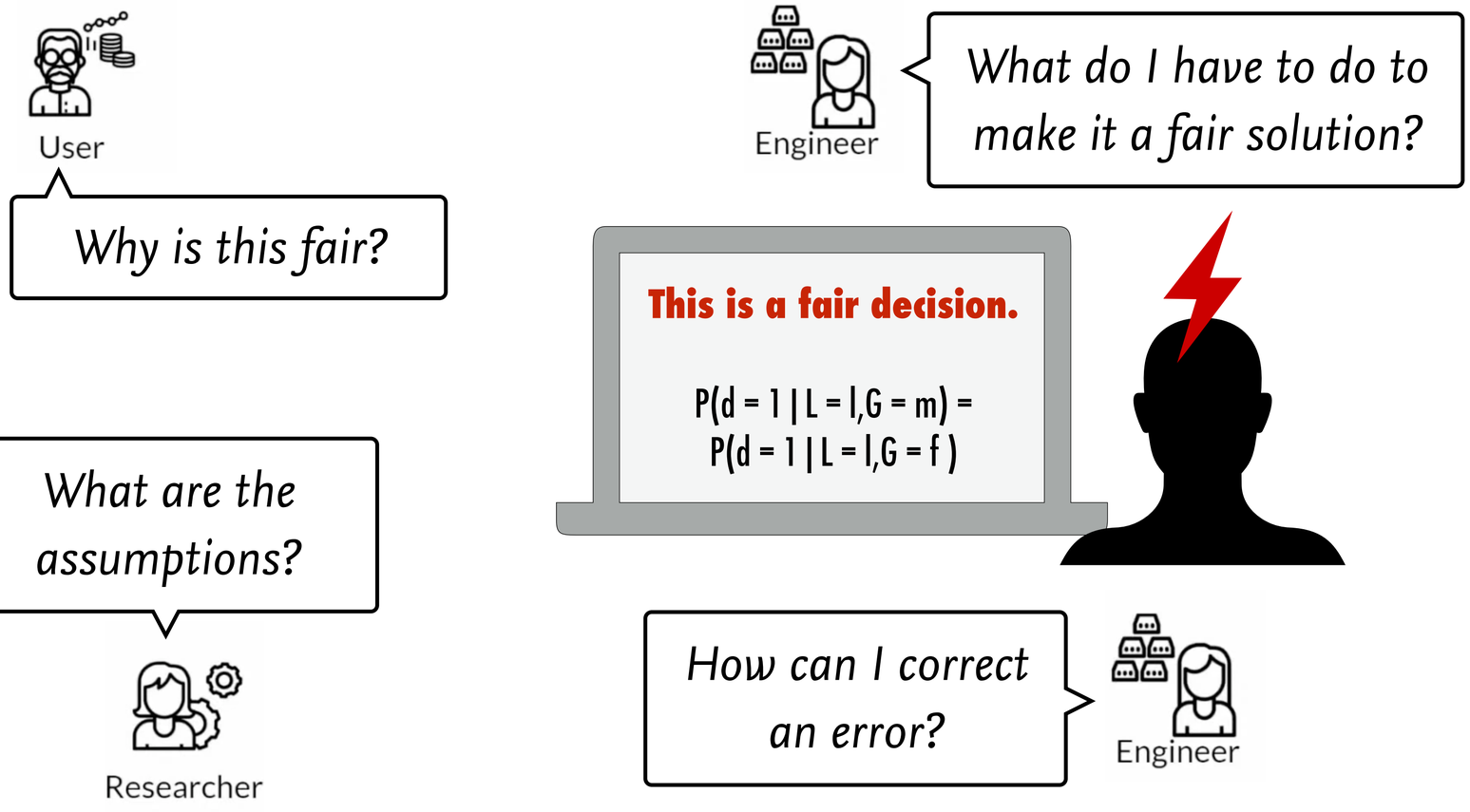
Designing for Interpretion (Human Processor Model, Sensation, Cognition, Model of Interaction, Principles of Good Design, Slip vs. error)

Break

Designing Interventions for Ensuring Interpretation (Educational Strategies , Cognitive Forcing Functions)

Recap

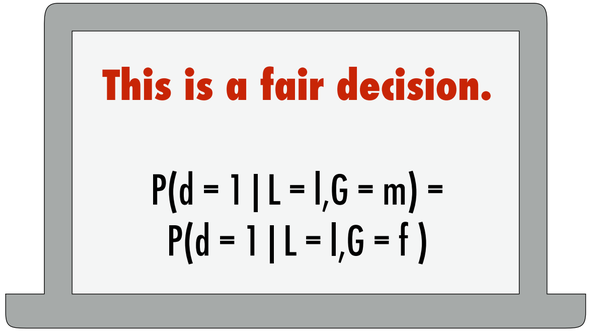
Motivating the Need for Transparency



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

Intrinsic vs. Post-hoc Interpretability Techniques

Use models that are intrinsically interpretable and known to be easy for humans to understand.



Intrinsic Interpretability Techniques

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

Challenge of Interpretability vs. Accuracy/Complexity

	Representational Complexity	Interpretability	Predictive Accuracy
Simple Models (e.g., linear regression, logistic regression and decision trees)	↓	↑	↓
Complex Models (e.g., random forests, support vector machines, deep learning)	↑	↓	↑

Bamman, D. (2016). Interpretability in human-centered data science. In CSCW Workshop on Human-Centered Data Science. https://cscw2016hcds.files.wordpress.com/2015/10/bamman_hcds.pdf
Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. Decision Support Systems, 51(4), 782-793.

Intrinsic vs. Post-hoc Interpretability Techniques

Use models that are intrinsically interpretable and known to be easy for humans to understand.

This is a fair decision.

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

Train a black box model and apply post-hoc interpretability techniques to provide explanations.

Intrinsic Interpretability Techniques

Post-hoc Interpretability Techniques

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

Interpretability Techniques - Explanations

> [...] an explanation of a model result is a description of how a model's outcome came to be. Explanations thus seek to describe the process, or rules, that were implemented to achieve an outcome independent of context. Typically, explanations are detailed, technical, and may be causative.
>
> Broniatowski (2021)

Broniatowski, D. A. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology.

A Participatory Process for Interpretability Techniques

WHAT to explain?

Expert Mental Model

- What happens to the best of our knowledge?
- What can be explained?
- What does an expert mental model of the system look like?

User Mental Model

- How do users currently make sense of the system?
- What is the user mental model of the system based on its current UI?
- How does it differ from the expert mental model?

Target Mental Model

- Which key components of the algorithm do users want to be made transparent in the UI?
- To what extent are users actually interested in the rationale behind the algorithms?

HOW to explain?

User Mental Model

- How can the target mental model be reached through UI design?
- How and where can transparency be integrated into the UI of the system?

Target Mental Model

- How has the users mental model developed?
- Has the target mental model been reached?

Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In Conf. on Intelligent User Interfaces, 211–223.

Defining a Mental Model

> **Mental models are internal representations that people build based on real world experiences. These models allow people to understand, explain, and predict phenomena.**
>
> Johnson-Laird (1983)

Johnson-Laird, P. (1983). Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge University Press.

Expert Mental Model vs. User Mental Model

Experts Mental Model

reflects explanation methods

Target Models

worst — better

User Mental Model

reflects explanation needs

Image from Cooper, A., Reimann, R., & Cronin, D. (2007). About face 3: the essentials of interaction design. John Wiley & Sons.

A Participatory Process for Interpretability Techniques

WHAT to explain?

Expert Mental Model

- What happens to the best of our knowledge?
- What can be explained?
- What does an expert mental model of the system look like?

User Mental Model

- How do users currently make sense of the system?
- What is the user mental model of the system based on its current UI?
- How does it differ from the expert mental model?

Target Mental Model

- Which key components of the algorithm do users want to be made transparent in the UI?
- To what extent are users actually interested in the rationale behind the algorithms?

HOW to explain?

User Mental Model

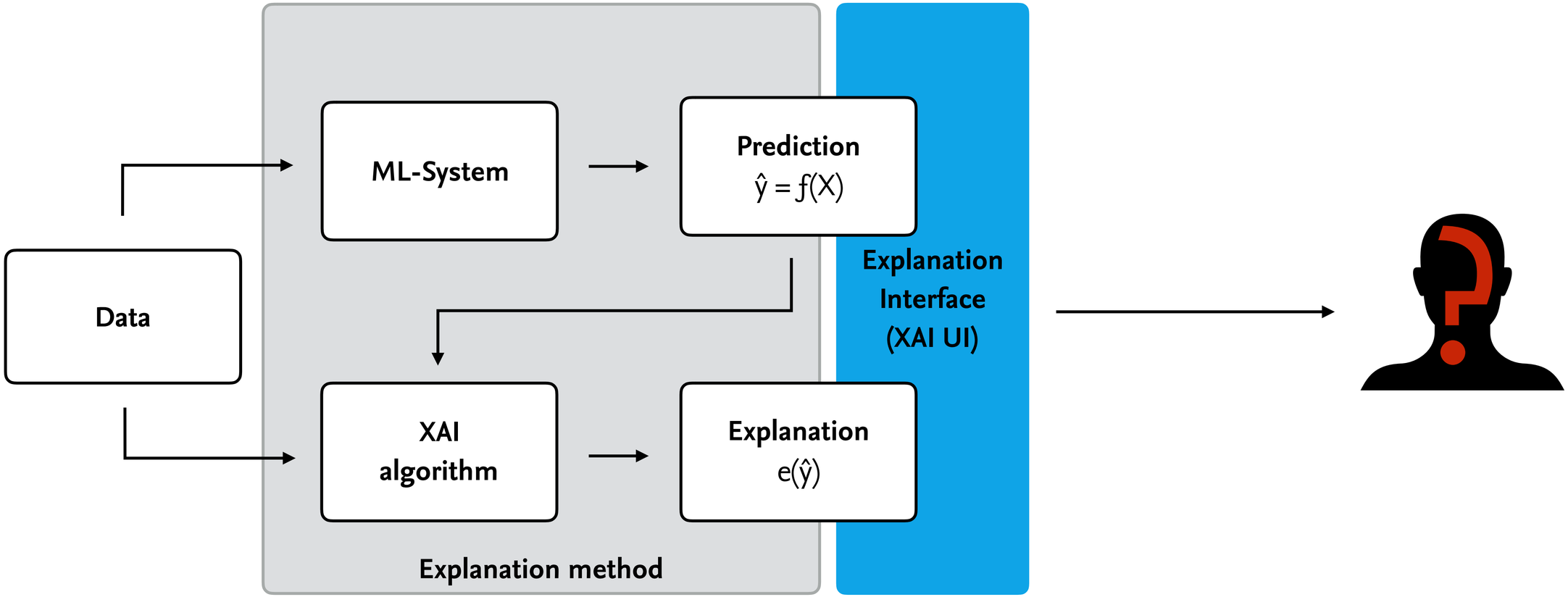
- How can the target mental model be reached through UI design?
- How and where can transparency be integrated into the UI of the system?

Target Mental Model

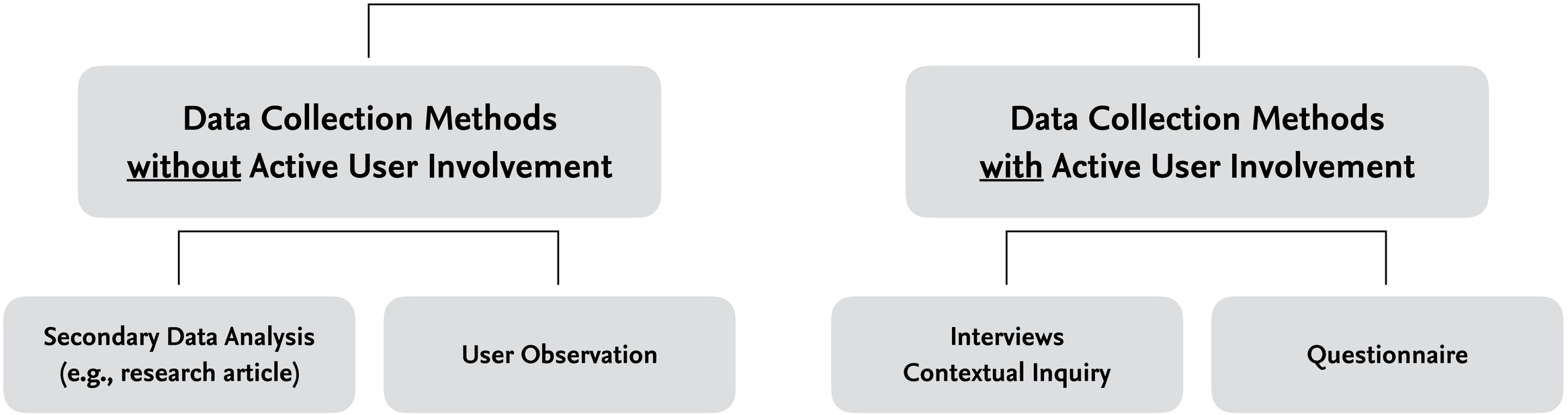
- How has the users mental model developed?
- Has the target mental model been reached?

Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In Conf. on Intelligent User Interfaces, 211–223.

From Explanations to Explanation User Interfaces



Eliciting Explanation Needs

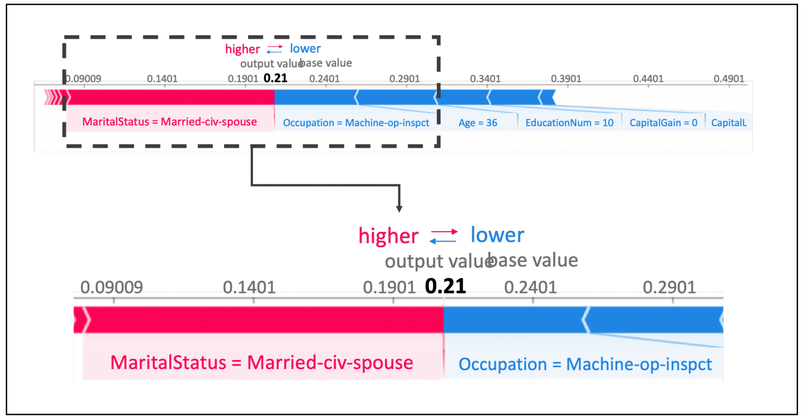


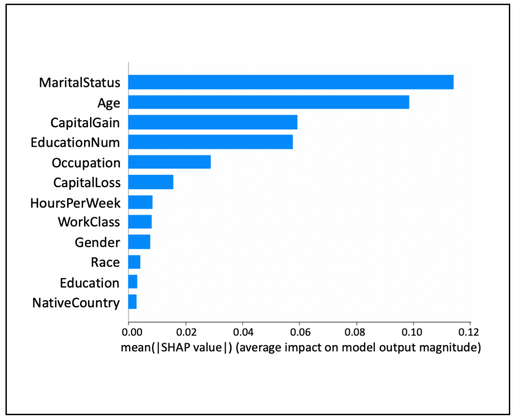
Insights from a Study Explanations based on SHAP

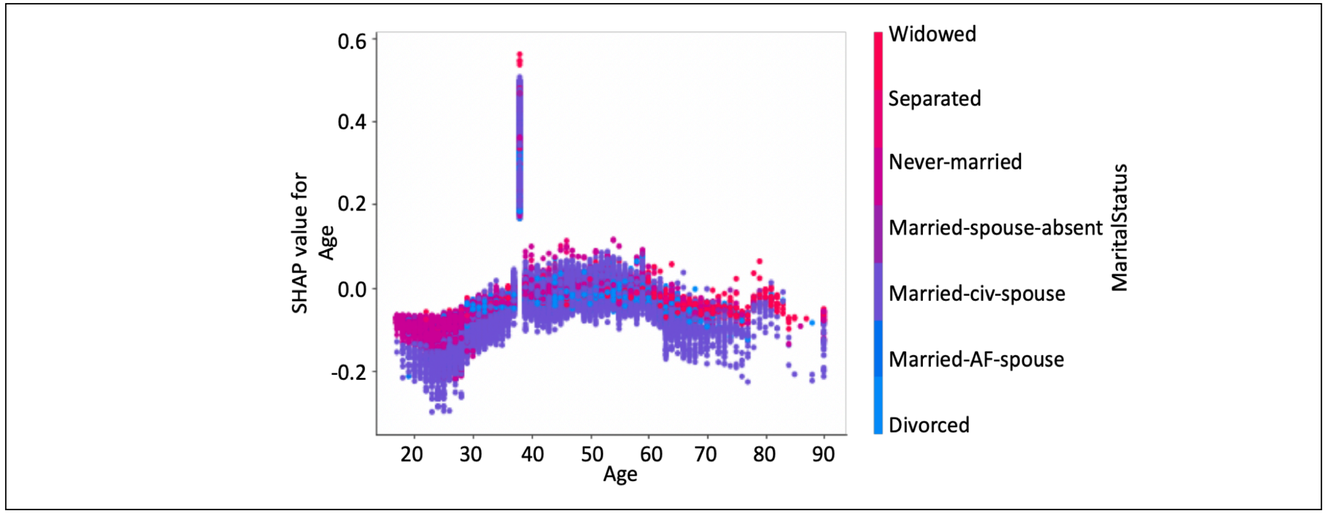
Participants often do not fully understand the underlying conceptional model of the tool.

The “ease of use” of interpretability tools (here SHAP tools) reduces participant’s critically thinking.

The diagrams conveyed trust in the quality of the underlying explanation method.



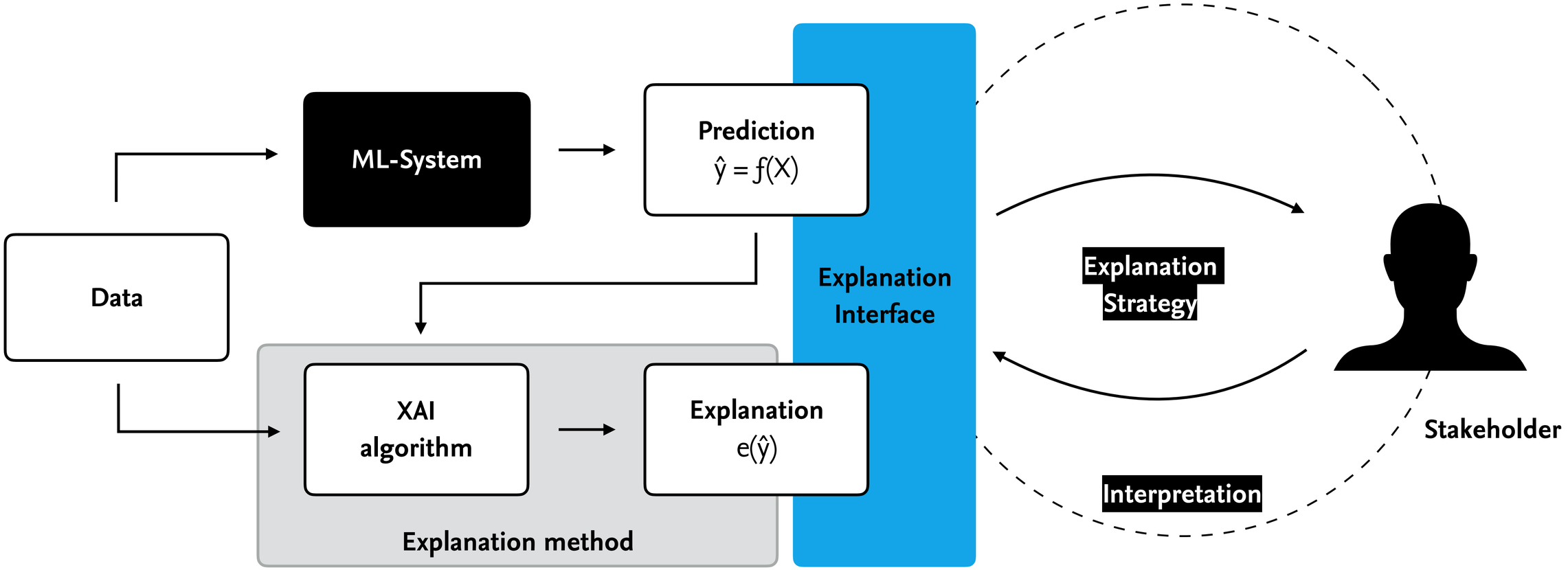




Harmanpreet Kaur, et al. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proce CHI. ACM, 1–14.

Explaining vs. Interpreting

Ensuring Interpretability by Explanation User Interfaces



Jesse Josua Benjamin, Christoph Kinkeldey, Claudia Müller-Birn, Tim Korjakow, and Eva-Maria Herbst. 2022. Explanation Strategies as an Empirical-Analytical Lens for Socio-Technical Contextualization of Machine Learning Interpretability. Proc. ACM Hum.-Comput. Interact. 6, GROUP.

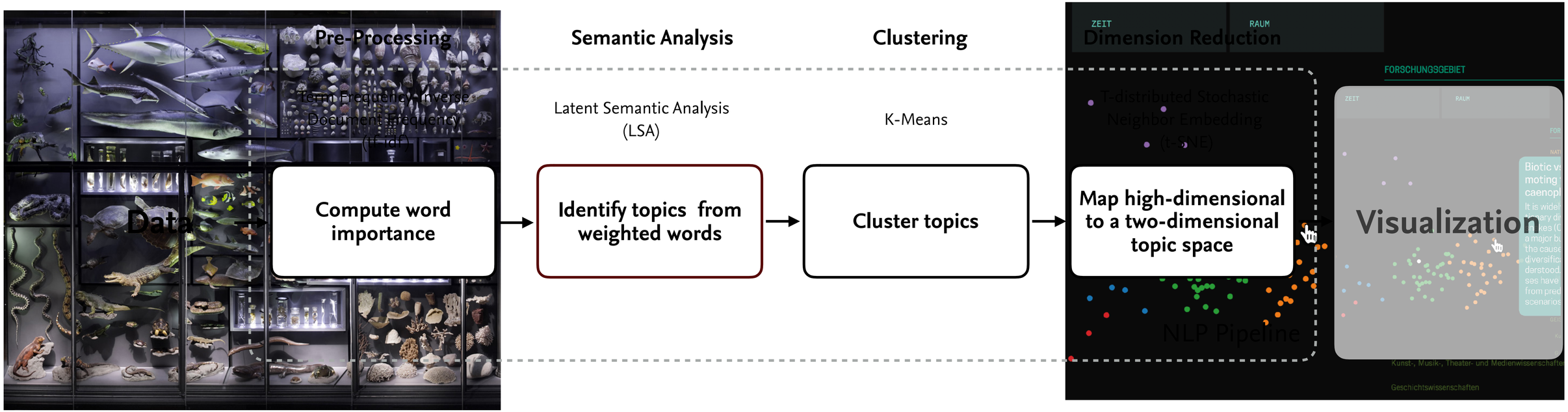
Interpretation

> **Interpretation refers to a human's ability to make sense, or derive meaning, from a given stimulus (e.g., a machine learning model's output) so that the human can make a decision.**
>
> Broniatowski (2021)

Broniatowski, D. A. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology.

Explanation Strategies — Understanding Interpretation in Context

Research Study: Contextualizing Interpretability



Research Information → NLP Pipeline → Human Interpretation

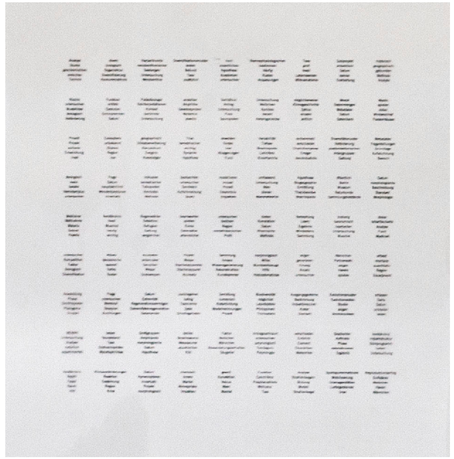
Used Explainability Methods

Explainability Methods

Local Techniques

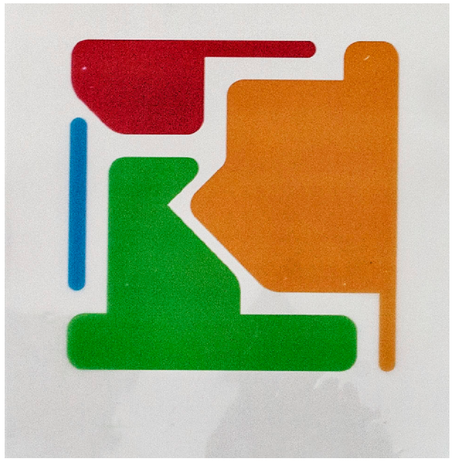
Feature Importance

Project Topwards



Similarity

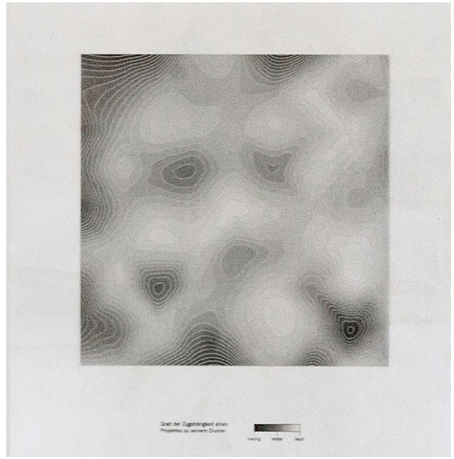
Cluster Regions



Global Techniques

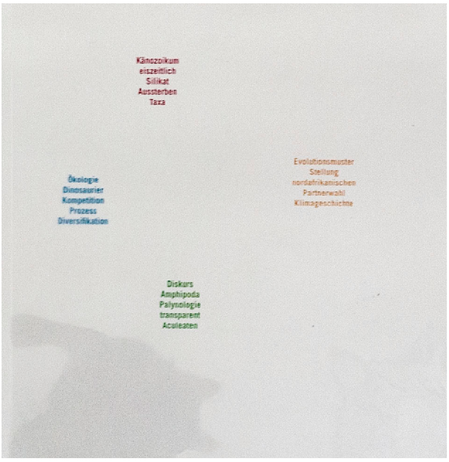
Regions of Error

Uncertainty Map



Feature Importance

Cluster Topwards



Used Method: Participatory Design

> **[…] participatory design sought not only to incorporate users in design, but also to intervene in situations of conflict through developing more democratic processes.**
>
> Liam Bannon, Jeffrey Bardzell, and Susanne Bødker. 2018. Reimagining participatory design. interactions 26, 1 (January - February 2019), 26–32. https://doi.org/10.1145/3292015

Participatory design (PD) emerged about 25 years ago as a distinct set of design and research practices rooted in a **Scandinavian approach** to systems design, commonly classed under the label of **cooperative design**.

Halskov, Kim, und Nicolai Brodersen Hansen. „The Diversity of Participatory Design Research Practice at PDC 2002–2012“. International Journal of Human-Computer Studies, Bd. 74, Februar 2015, S. 81–92. DOI.org (Crossref), doi:10.1016/j.ijhcs.2014.09.003.

Aspects of Participatory Design

Politics	People	Context	Methods	Product
» People who are affected by a decision should have an opportunity to influence it	» People play critical roles in design by being experts in their own lives » Users learn about technological means	» Use cases as fundamental starting point for the design process » Mutual learning through collective reflection	» Users influence in design processes » Learn about the users' situation	» Design alternatives, improving quality of life » Designing to respond to human need

Adopted from Halskov, Kim, und Nicolai Brodersen Hansen. „The Diversity of Participatory Design Research Practice at PDC 2002–2012". International Journal of Human-Computer Studies, Bd. 74, 2015, 81–92.

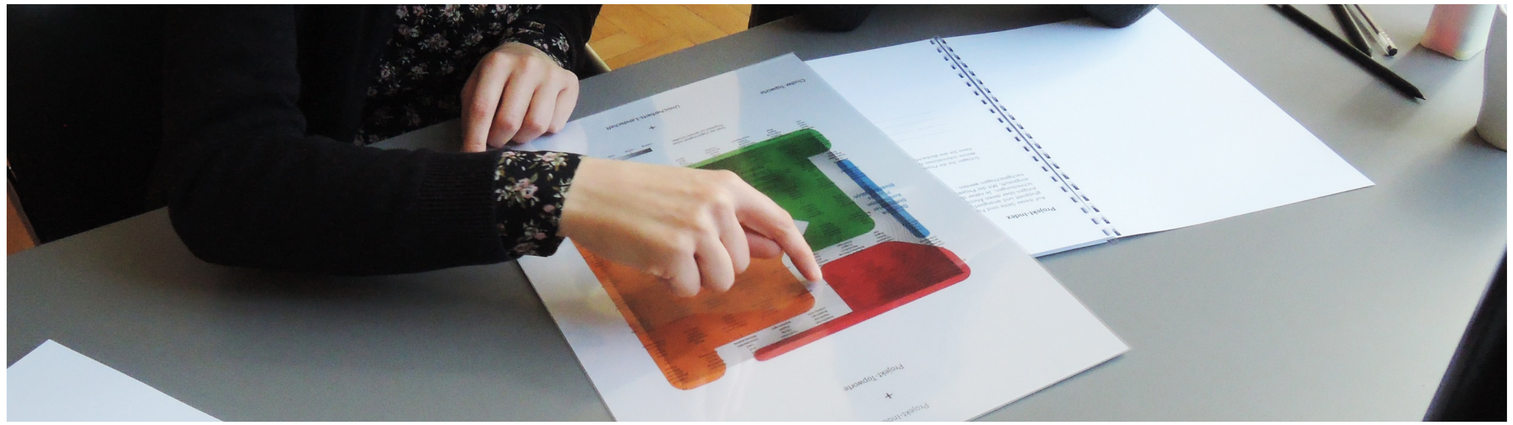
Participatory Design for Revealing Explanation Strategies

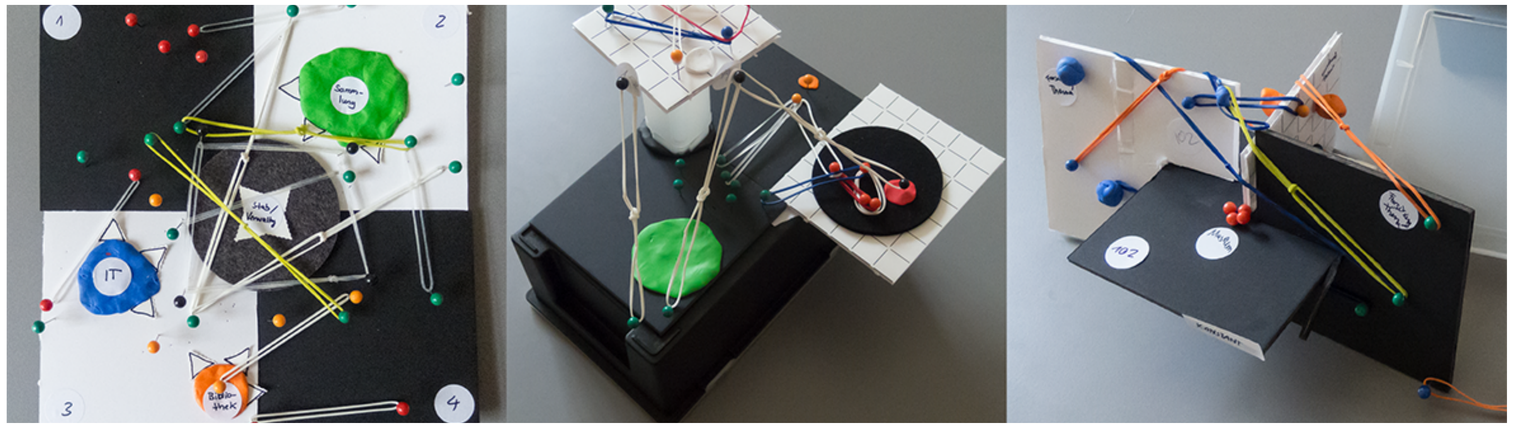
Reactive Part

- **Phase I: Application**
- **Phase II: Enactment**

Self-reflexive Part

- **Phase III: Reification**
- **Phase VI: Reflection**





Jesse Josua Benjamin, Christoph Kinkeldey, Claudia Müller-Birn, Tim Korjakow, and Eva-Maria Herbst. 2022. Explanation Strategies as an Empirical-Analytical Lens for Socio-Technical Contextualization of Machine Learning Interpretability. Proc. ACM Hum.-Comput. Interact. 6, GROUP.

Implications

Methodological Implications of Explanation Strategies

» The co-creation workshop is a proof of concept for studying human interpretation in context. Explanation strategies function as the “what” that links context, explanation method and explanations.

» Explanation strategies learned from a co-design workshop could be used within explainability scenarios, which could then be probed.

Design Implications for Explanations for Non-ML Experts

» Enable combinations of explanations to support explanation strategies.

» Supplement explanations with contextual cues.

Jesse Josua Benjamin, Christoph Kinkeldey, Claudia Müller-Birn, Tim Korjakow, and Eva-Maria Herbst. 2022. Explanation Strategies as an Empirical-Analytical Lens for Socio-Technical Contextualization of Machine Learning Interpretability. Proc. ACM Hum.-Comput. Interact. 6, GROUP.

Designing for Interpretion

Human Processor Model

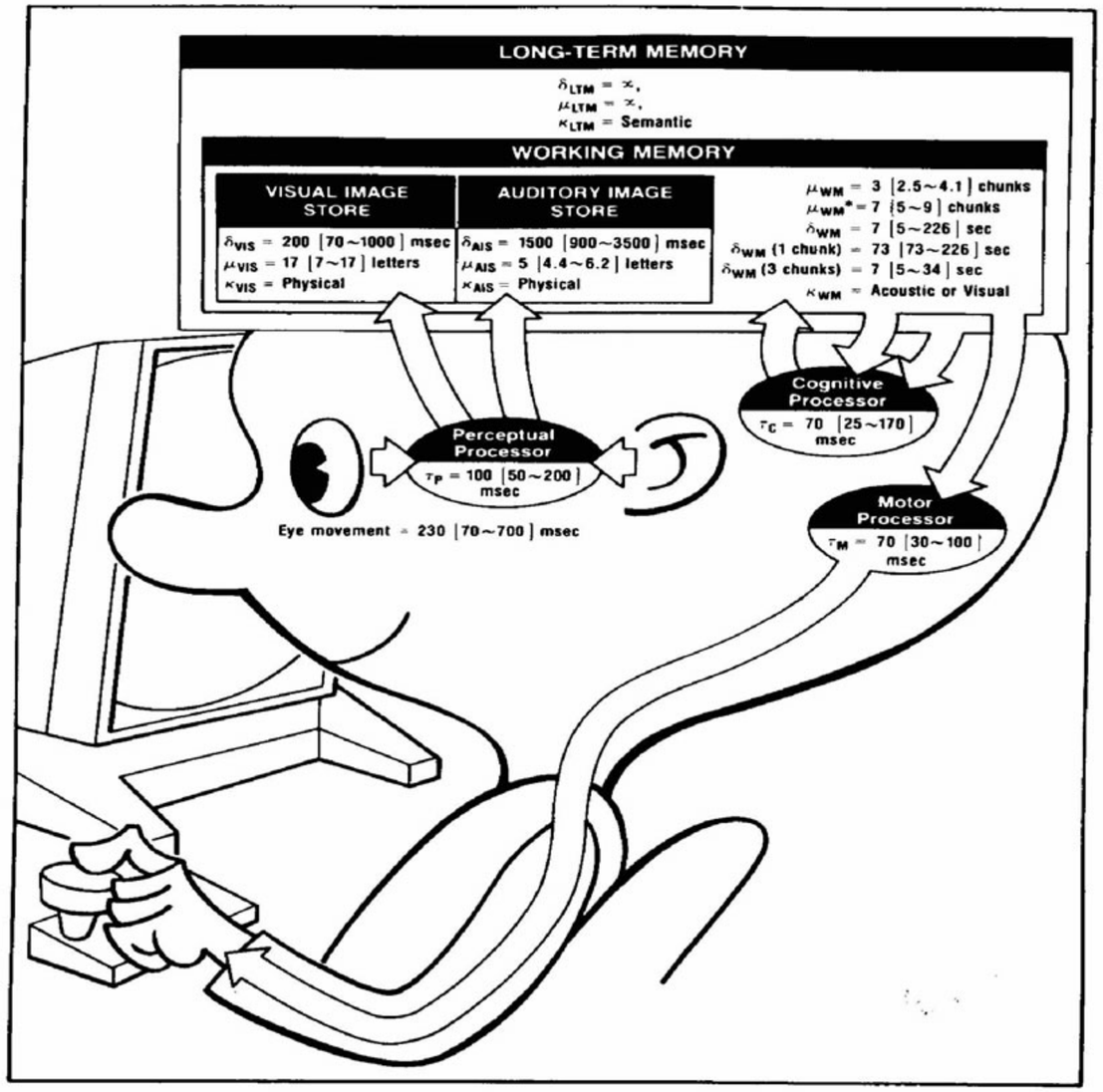
The model is based on years of basic psychology experiments found in the literature.

Models the information processes of a user interacting with a computer

» Most likely serial in action & parallel in recognition

» Skilled performance differs from novice performance

Focus on a single user interacting with some entity.



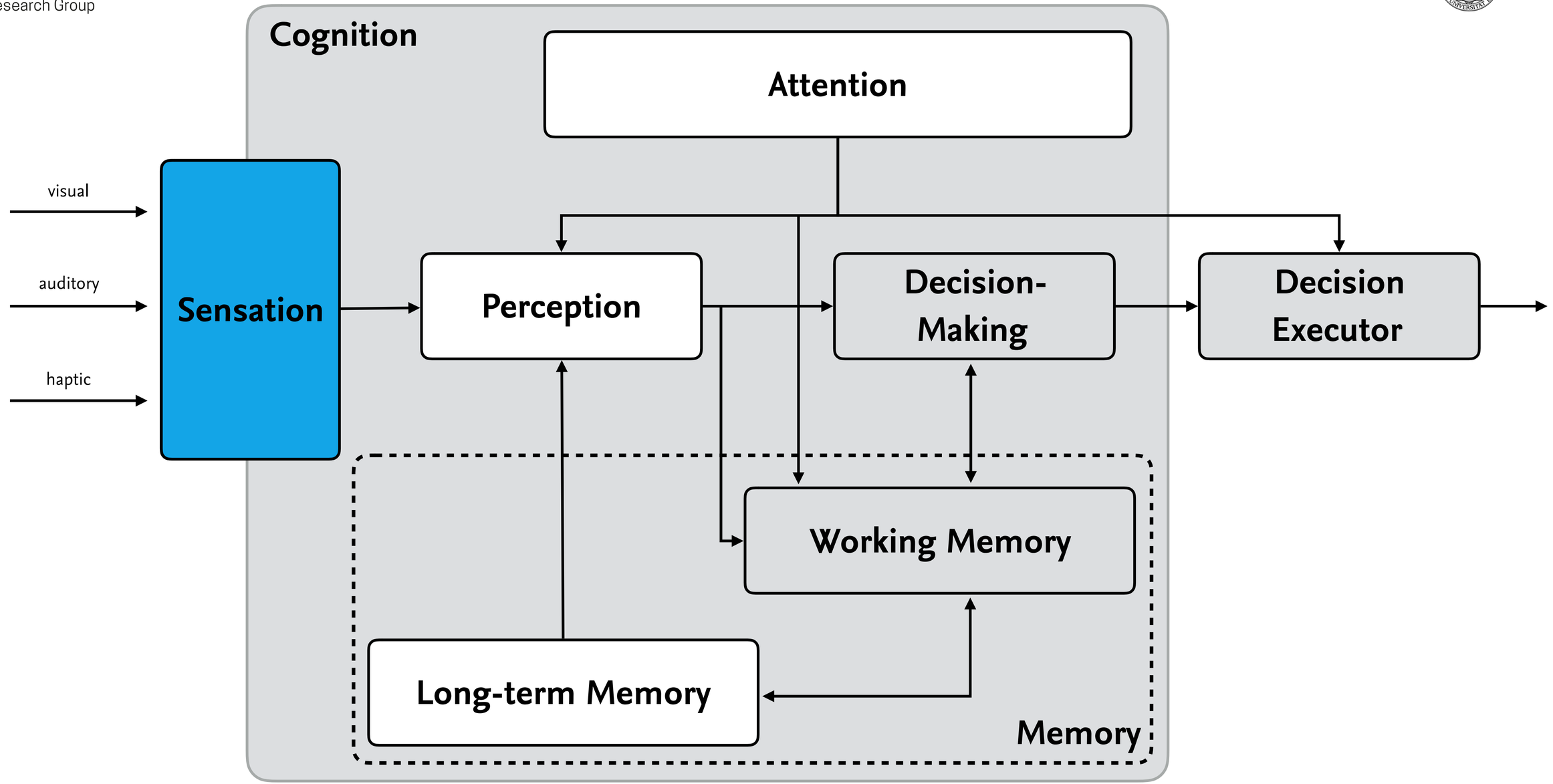
Card, S.K; Moran, T. P; and Newell, A. The Model Human Processor: An Engineering Model of Human Performance. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), Handbook of Perception and Human Performance. Vol. 2: Cognitive Processes and Performance, 1986, pages 1–35.

Understanding the Basics of Human Information Processing

We need an understanding of what can and cannot be expected of human beings.

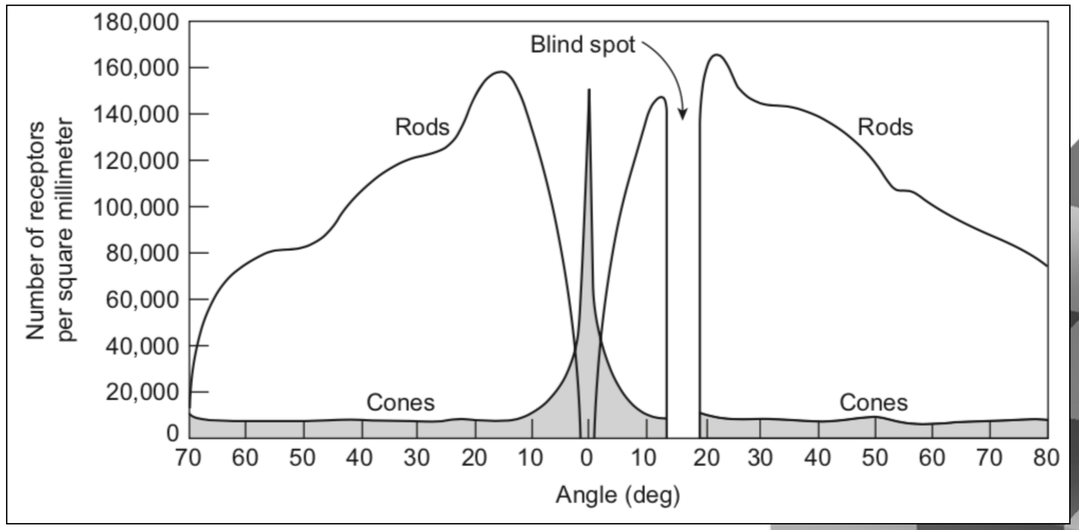
We identify and explain the nature and causes of problems that people encounter.

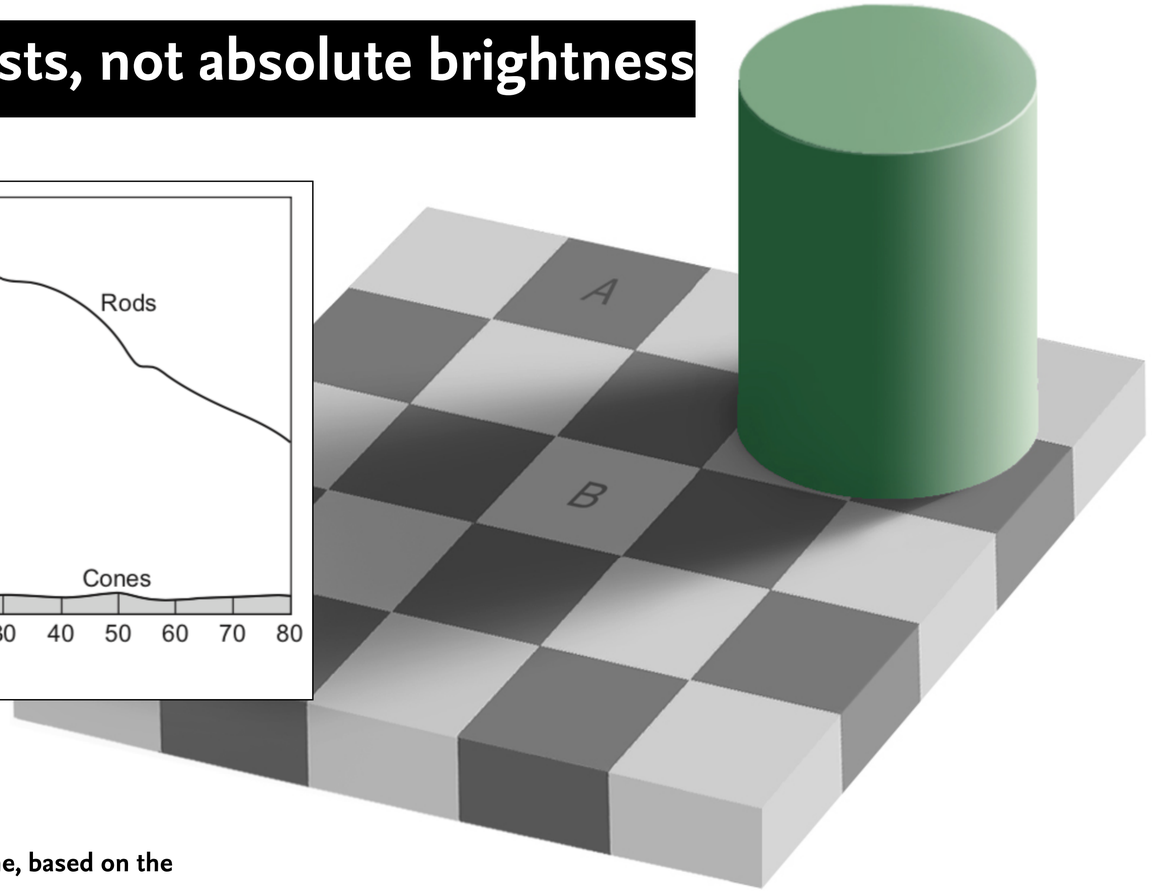
We know theories and methods that can lead to the design of XUIs.


Cognition
Attention
visual
auditory
haptic
Sensation
Perception
Decision-
Making
Decision
Executor
Working Memory
Long-term Memory
Memory

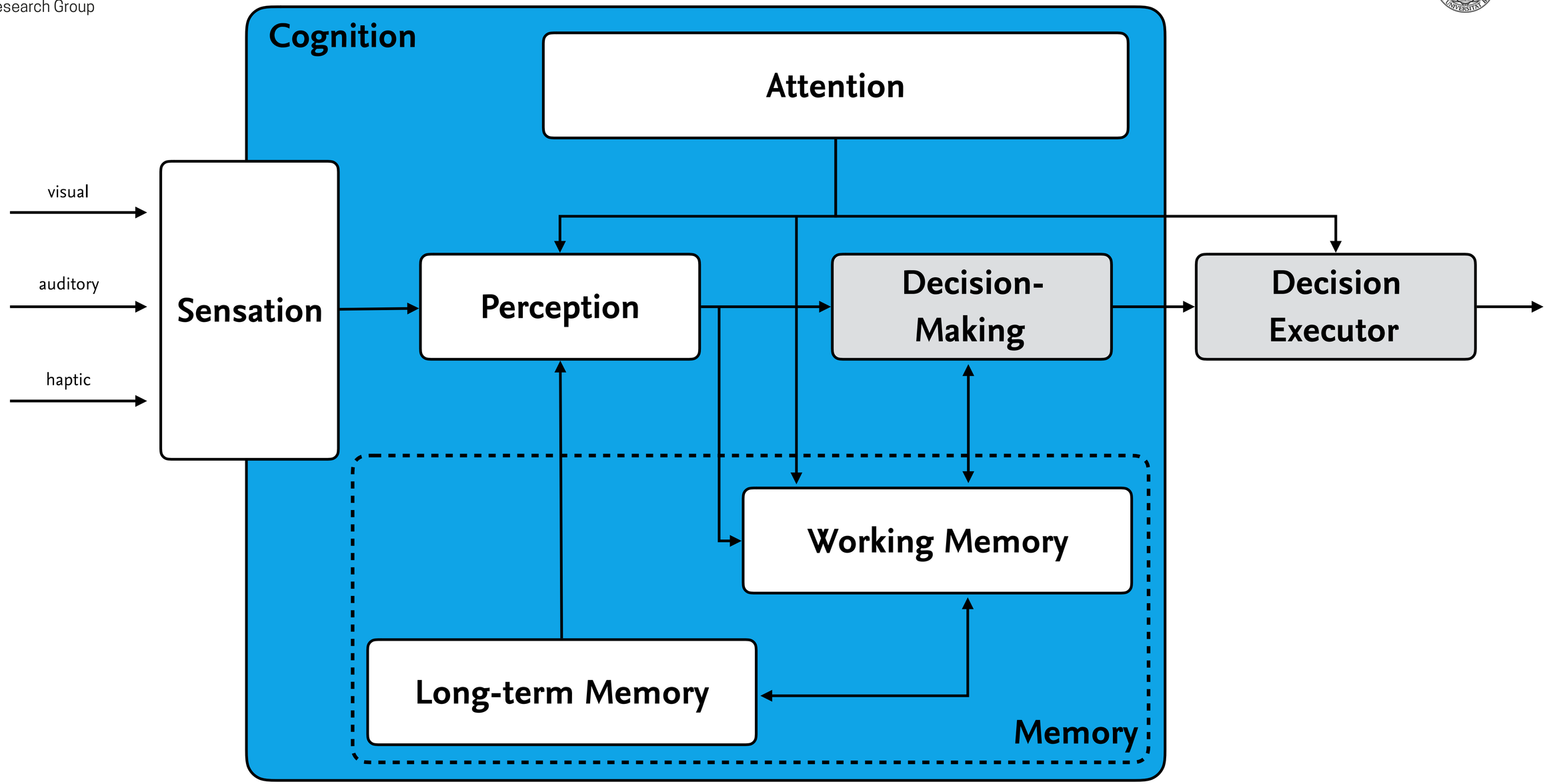
Why does Sensation Matters? - One example

Vision is optimized for contrasts, not absolute brightness





By Original by Edward H. Adelson - File created by Adrian Pingstone, based on the original created by Edward H. Adelson, Copyrighted free use, https://commons.wikimedia.org/w/index.php?curid=45737683


Cognition
Attention
visual
auditory
haptic
Sensation
Perception
Decision-Making
Decision Executor
Working Memory
Long-term Memory
Memory

The Role of Attention

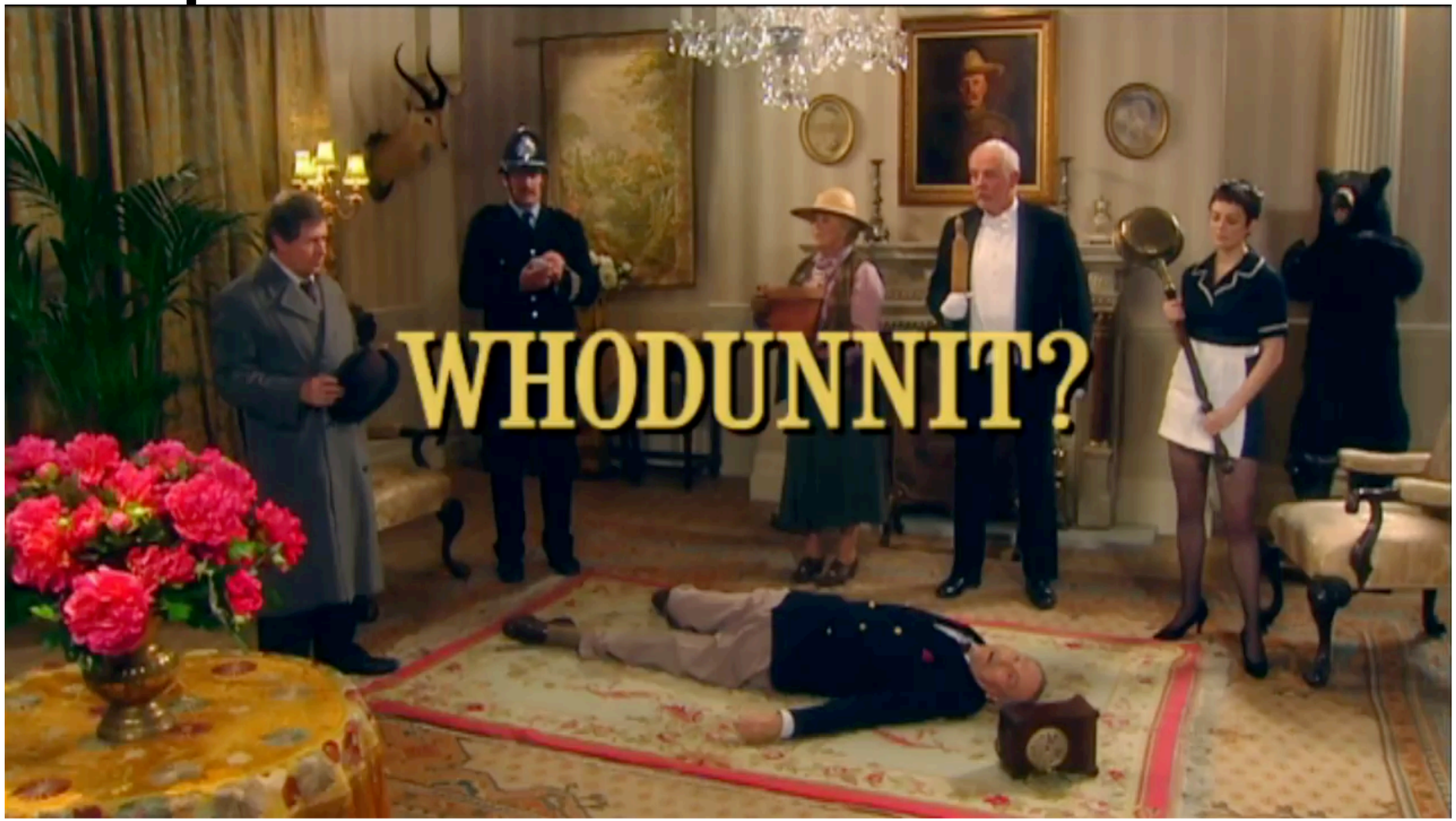
Attention is selecting things to concentrate on a point in time from the mass of stimuli around us. It allows us to focus on information that is relevant to what we are doing.

The focussed and divided attention enables us to be selective in terms of the mass of competing stimuli but limits our ability to keep track of all events.

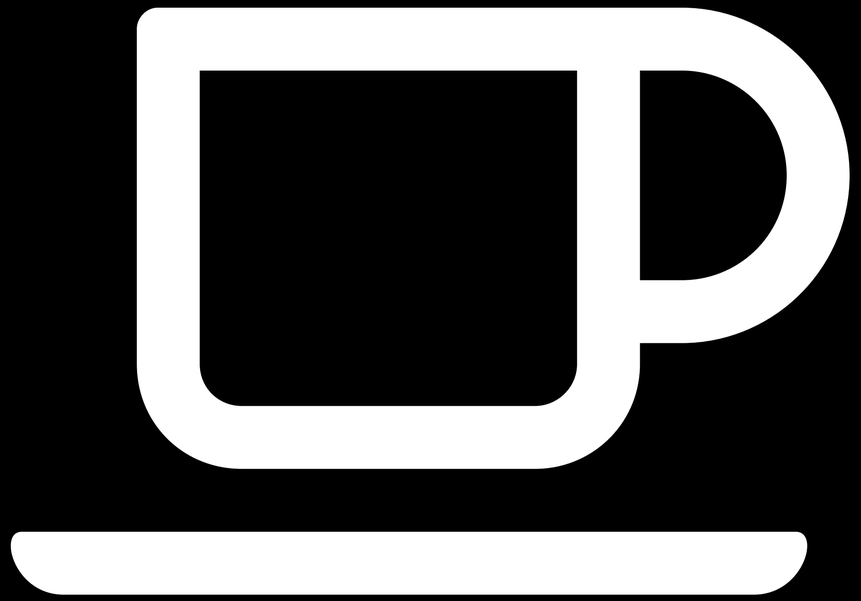
Exemplary Challenge:

» Inattentional Blindness which represents the selective attention in perception. It keeps humans focused on important aspects without distraction from irrelevant objects and events.

Example: Whodunnit?



Source: https://www.youtube.com/watch?v=ubNF9QNEQLA



5 minutes break

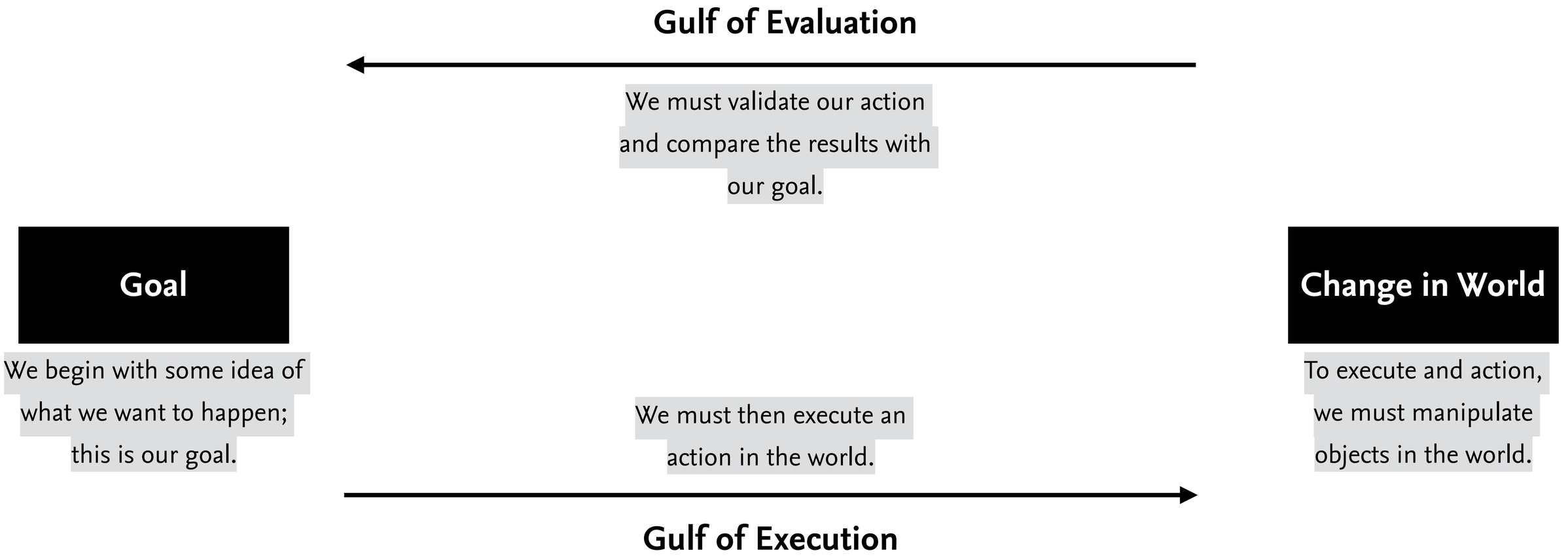
Understanding the Basics of Human Information Processing

We need an understanding of what can and cannot be expected of human beings.

We identify and explain the nature and causes of problems that people encounter.

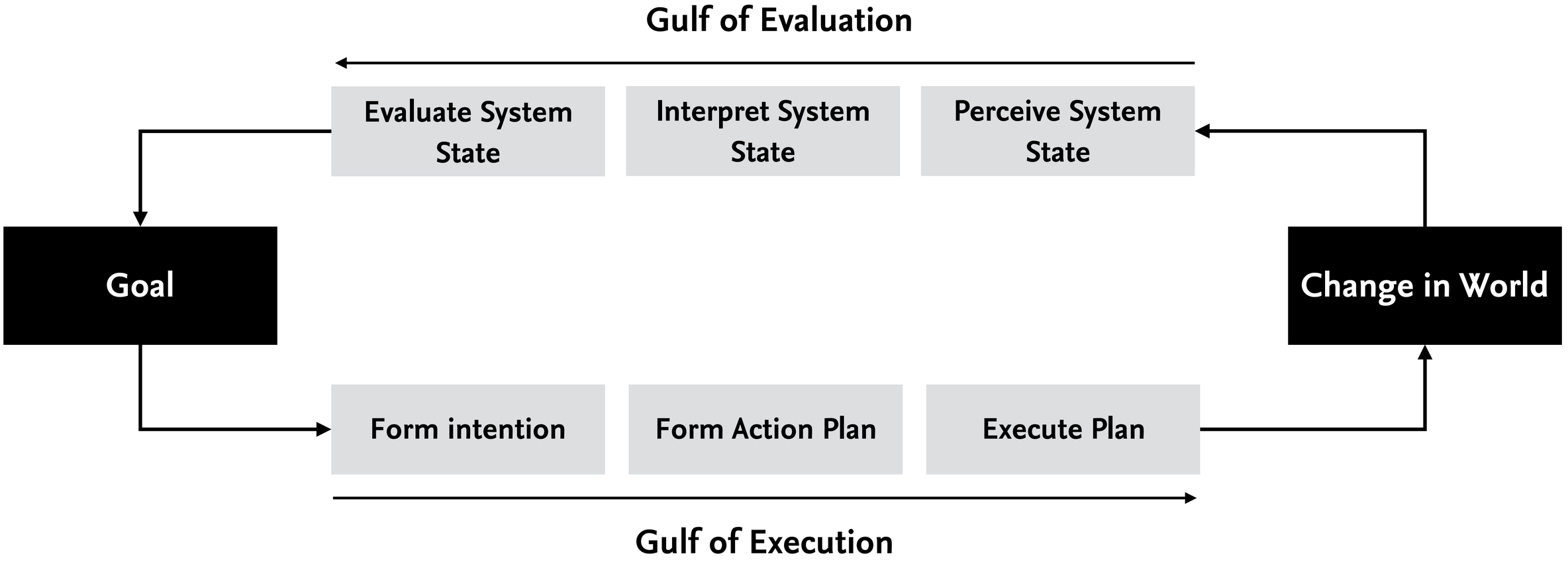
We know theories and methods that can lead to the design of XUIs.

Model of Interaction



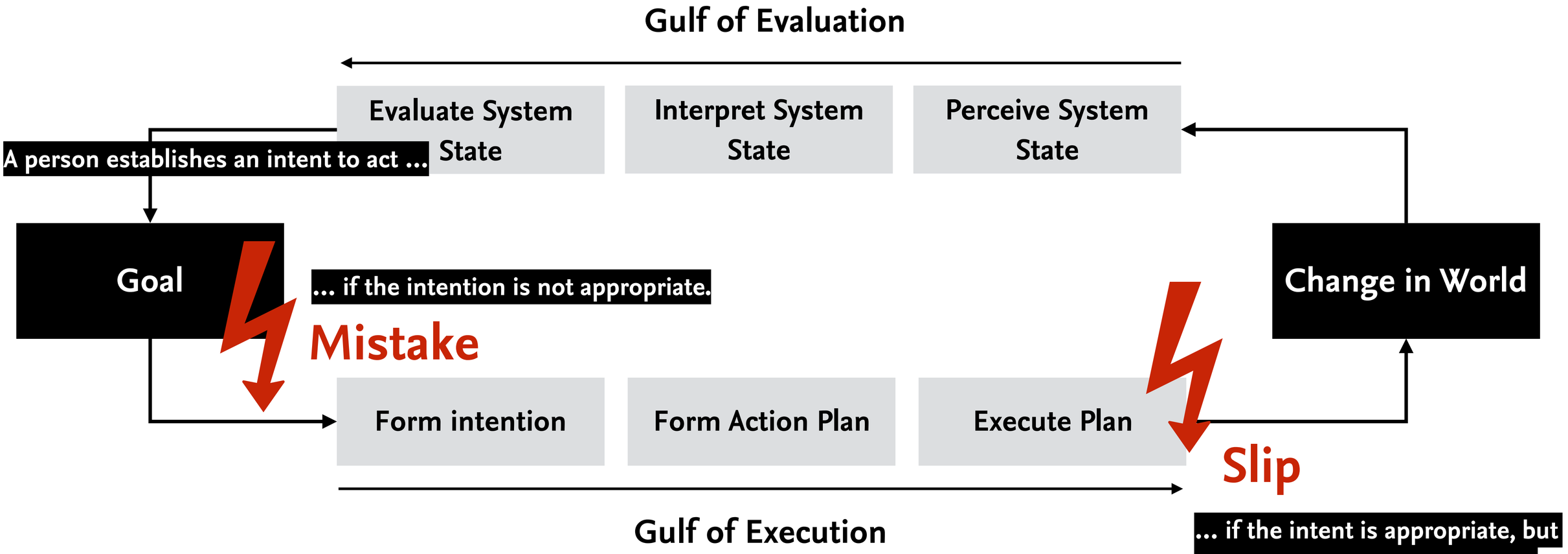
Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic books.

Model of Interaction (*cont.*)



Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic books.

To Err is Human



Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic books.

Slips vs. Errors

> **Mistakes are errors of conscious decision. … Slips … are unintended.**

Mistakes are “errors of intention” — the person intentionally decides to do something that turns out to be incorrect or to have an undesirable result. People make mistakes either because they have an incorrect understanding of the choices or because they have inaccurate or incomplete information, they have “a faulty mental model.”

A **slip** is an error a person does something they did not mean to do.

Principles of Good Design

Aim of a good design is to minimize the gulfs of execution and evaluation.

In order to do this the design should

- Help the user build the correct conceptual model of the system
- Make the right parts visible
- Provide memory aids to the user
- Provide good feedback
- Accommodate errors

Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic books.

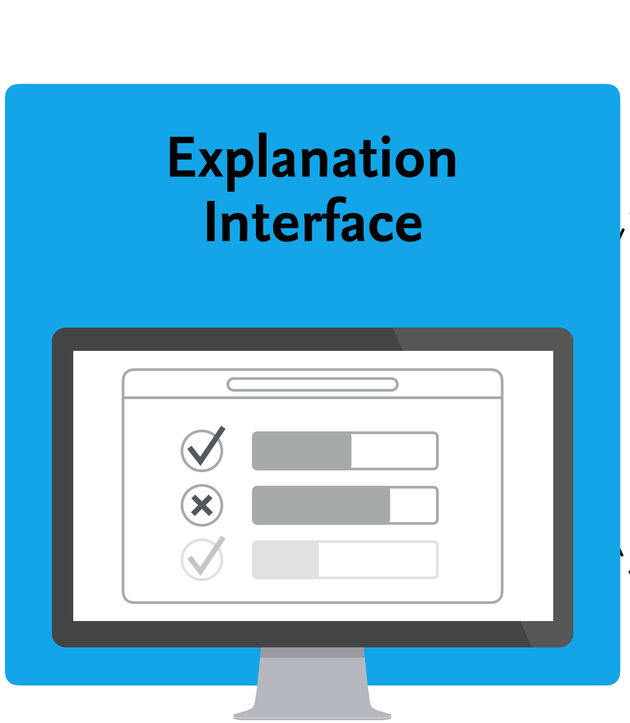
Understanding the Basics of Human Information Processing

We need an understanding of what can and cannot be expected of human beings.

We identify and explain the nature and causes of problems that people encounter.

We know theories and methods that can lead to the design of XUIs.

What Does it Mean for XAI?



Interpretation

Follow correct suggestions (appropriate reliance)

Ignote correct suggestions (self-reliance)

Follow incorrect suggestions (over-reliance)

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. 2015 International Conference on Healthcare Informatics, 160–169. ICHI.2015.26

Using Explanations does not result in better decision making

Relationship between a user's trust and their reliance on the system

» Users who trust the system highly are more likely to over-rely on the system's suggestions

» Users who distrust the system are more likely to rely on their own knowledge, even if it is poor

Effect of explanations on user's trust and reliance

» Without providing explanations there is a danger that users will rely too much on themselves

However, explanations are insufficient for addressing overreliance on imperfect ML algorithms.

Slip or Error ?

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. Translational Psychiatry, 11(1), 1–9.
Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. 2015 International Conference on Healthcare Informatics, 160–169. ICHI.2015.26

Dual Process Theory

Automatic Thinking (System 1)

» Intuitive understanding

» Decisions are made instinctively, emotionally and unconsciously

» Finds application in repeated and practiced actions

Reflective Thinking (System 2)

» 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.

Diverging Goals of Interventions

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.

Understanding the Basics of Human Information Processing

We need an understanding of what can and cannot be expected of human beings.

We identify and explain the nature and causes of problems that people encounter.

We know theories and methods that can lead to the design of XUIs.

Designing Interventions for Ensuring Interpretation

Overview on Interventions Supporting System 2 Thinking

Educational Strategies

- » metacognitive debiasing techniques
- » aim at enhancing future decision-making by increasing awareness about the existence of different decision-making pitfalls
- » examples are educational curricula, simulation training

Cognitive Forcing Functions

- » interventions which take place at the decision-making time
- » encourage the decision-maker to engage analytically with the decision at hand
- » Examples are checklists, diagnostic time-outs, and slow decision-making

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.

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.

Experimental Setup

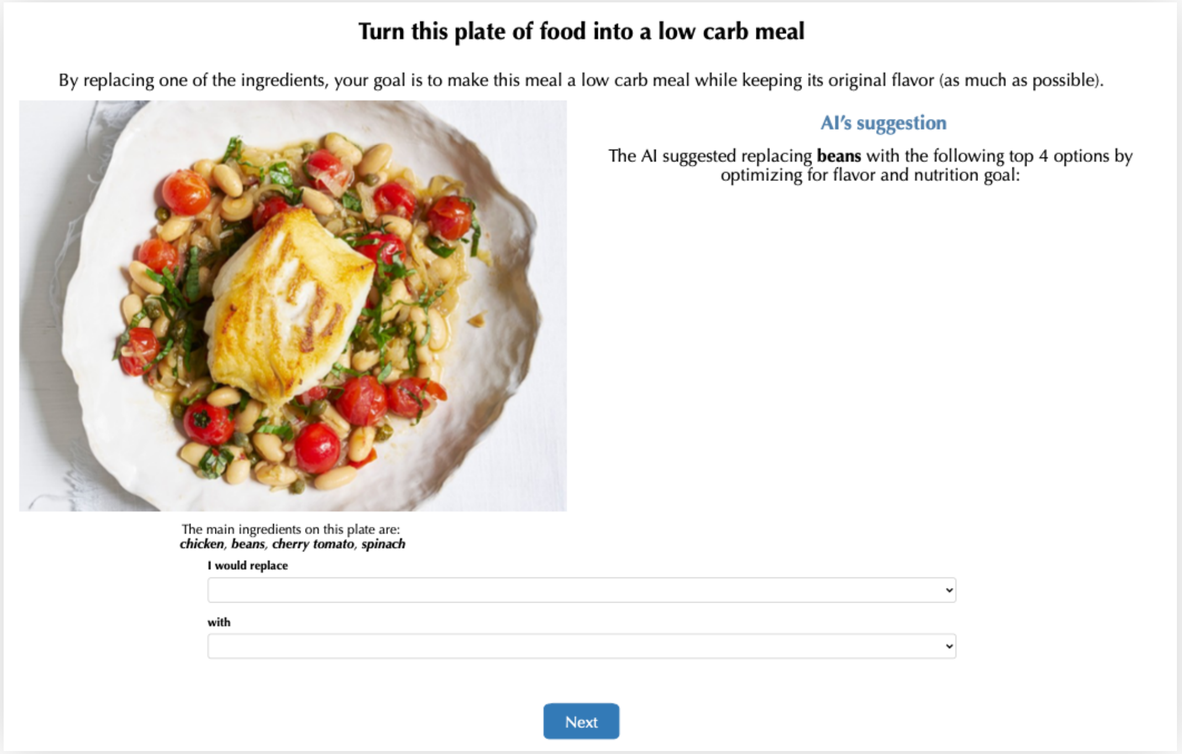
Six conditions of **a nutrition task** for laypeople

» 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%.



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

Simple Explainable AI (SXAI)

» Condition 2a: Explanation

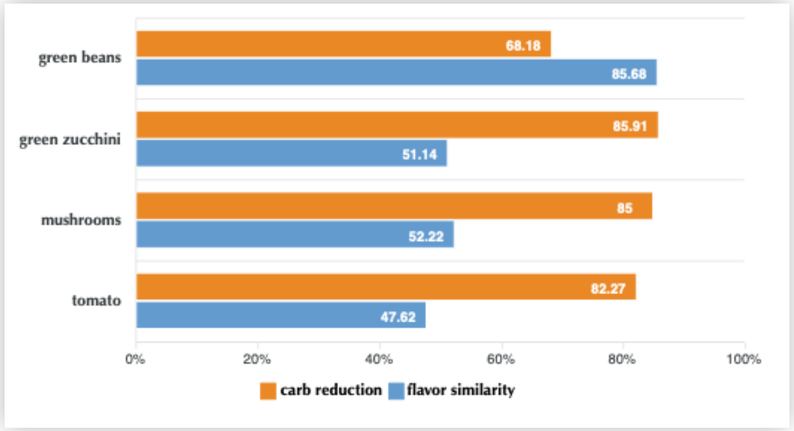
» Condition 2b: Uncertainty

Cognitive Forcing Functions (CFF)

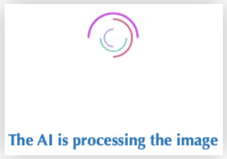
» Condition 3a: On demand

» Condition 3b: Update

» Condition 3c: Wait



Condition 1 and 2a, decision update possible



Followed by condition 2a

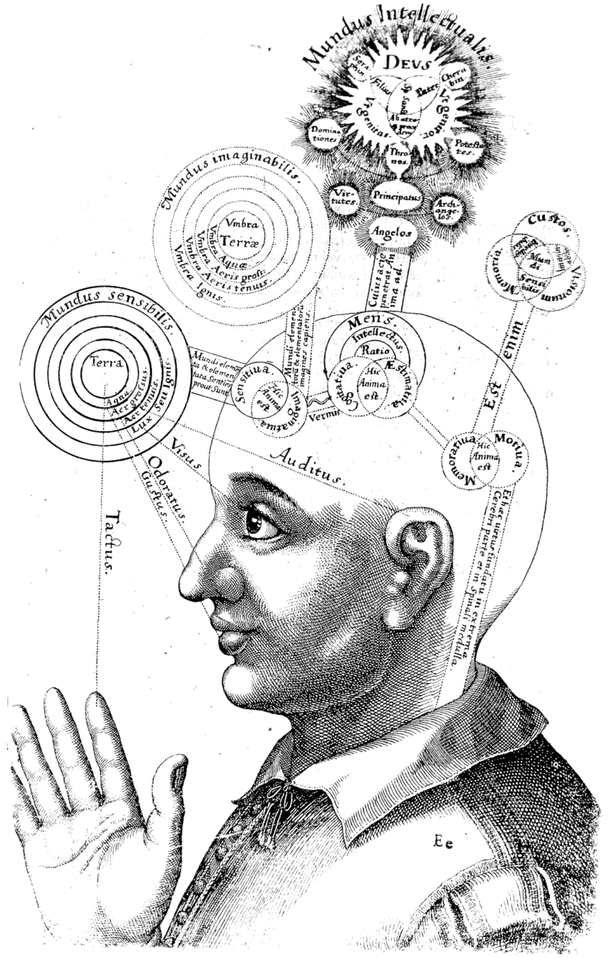
Buçinca, Zana, Maja Barbara Malaya, und Krzysztof Z. Gajos. „To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making“. Proceedings of the ACM on Human-Computer Interaction_ 5, Nr. CSCW (2021): 188:1-188:21.

Results

Cognitive forcing functions reduce overreliance on AI compared to the simple XAI approaches.

There was no significant difference in performance between simple XAI approaches and CFF.

There is a trade-off between the acceptability of a design of the explanation interface and the performance of the human+AI team.



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.
Image Source: https://de.wikipedia.org/wiki/Bewusstsein#/media%20/File:RobertFuddBewusstsein17Jh.png

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.

Check your Insights

Why does the mental model of the XAI developer differ from the mental model of a layperson?

What is the difference between interpretation and explanation?

What human-centered methods can help to elicit explanation needs?

How can you identify possible problems people encounter when interaction with XAI UIs?

What psychology theory did we discussed in the context of human thought and how does this theory influences peoples understanding of XAI UIs?

What are cognitive forcing functions and how can they applied to XAI UIs?

«Human-Centered Data Science»

Next week: Post-hoc Interpretability: Evaluating Explanation UIs

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

July 7, 2022