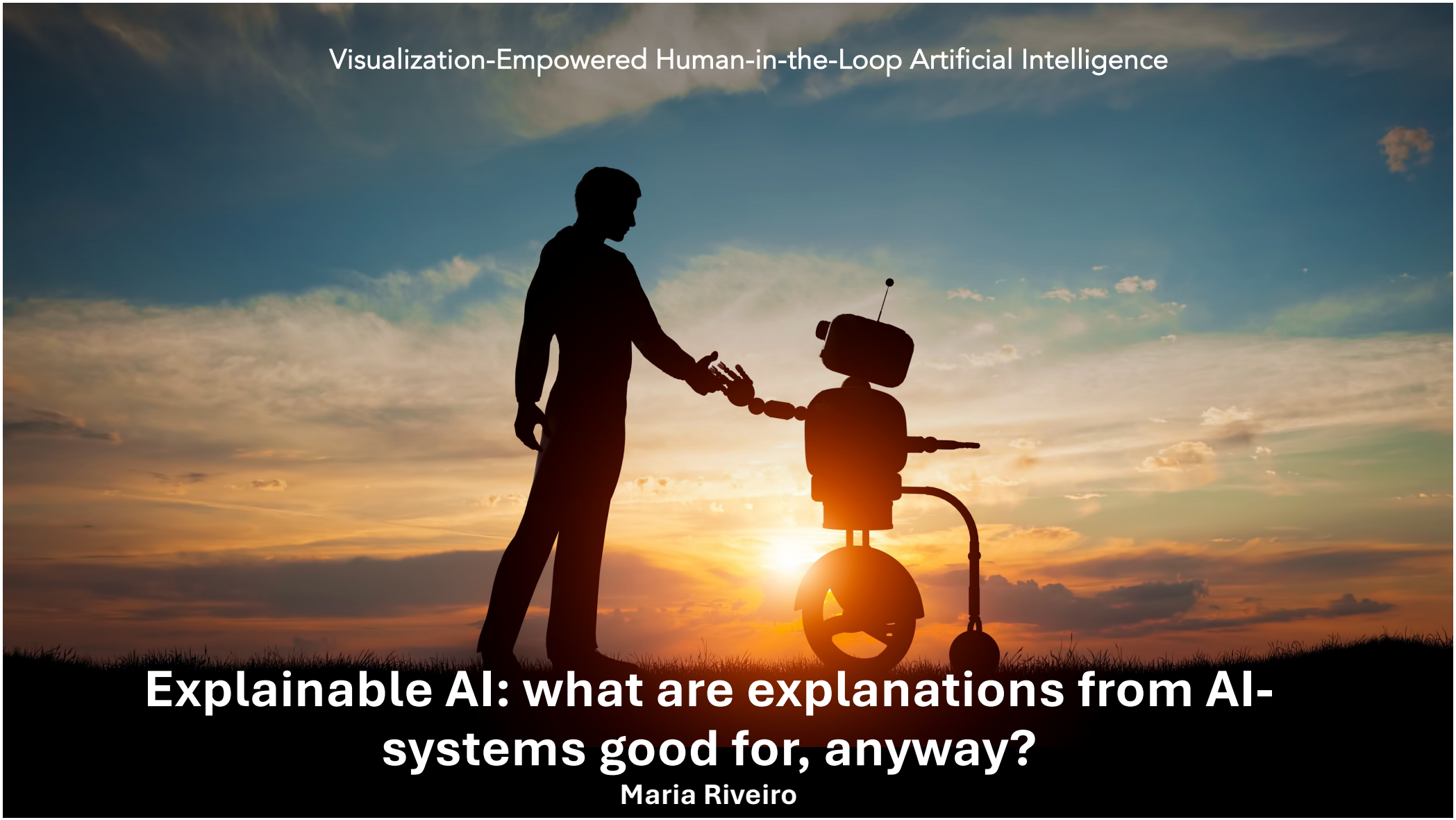
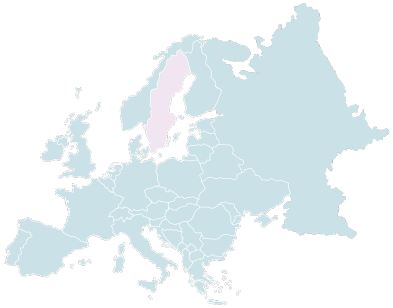


Explainable AI: what are explanations from AI-systems good for, anyway?

Maria Riveiro

JÖNKÖPING UNIVERSITY



Human-Centered Technology
Computer Science and Informatics
School of Engineering

Outline

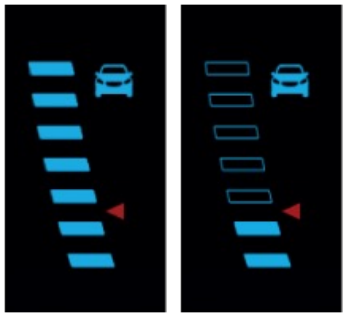
- Human-Machine Collaboration.

  Explainable AI
- Empirical Studies
- Future

Visualization-Empowered Human-in-the-Loop AI and Explainable AI

- **Visualization and Explainable AI provide transparency**
  - Visualization as a powerful mediator
  - Explanations (visual, textual, sound, multimodal) encourage user engagement and facilitate informed decision-making
  - Support human-in-the-loop interaction
  - Strengthen user trust and confidence in AI-generated outcomes

Autonomous driving



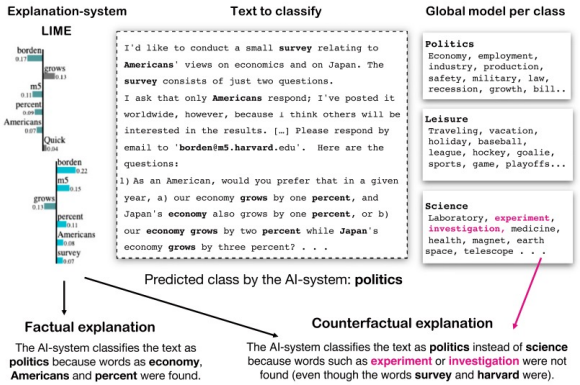
Results:

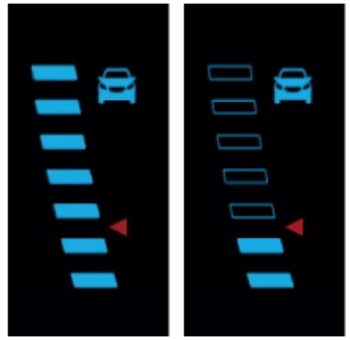
- Faster take-over
- More look away
- Better trust calibration

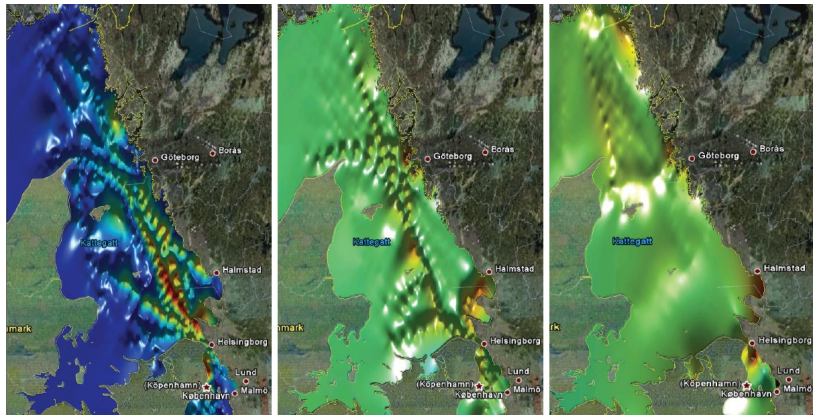
Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013, October). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications* (pp. 210-217).

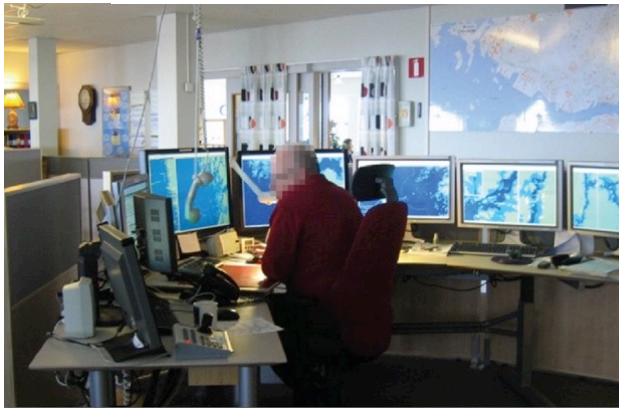
EXPLAINABLE AI & VISUALIZATION

- Explainable methods
- Visualization
- Human-Centered AI







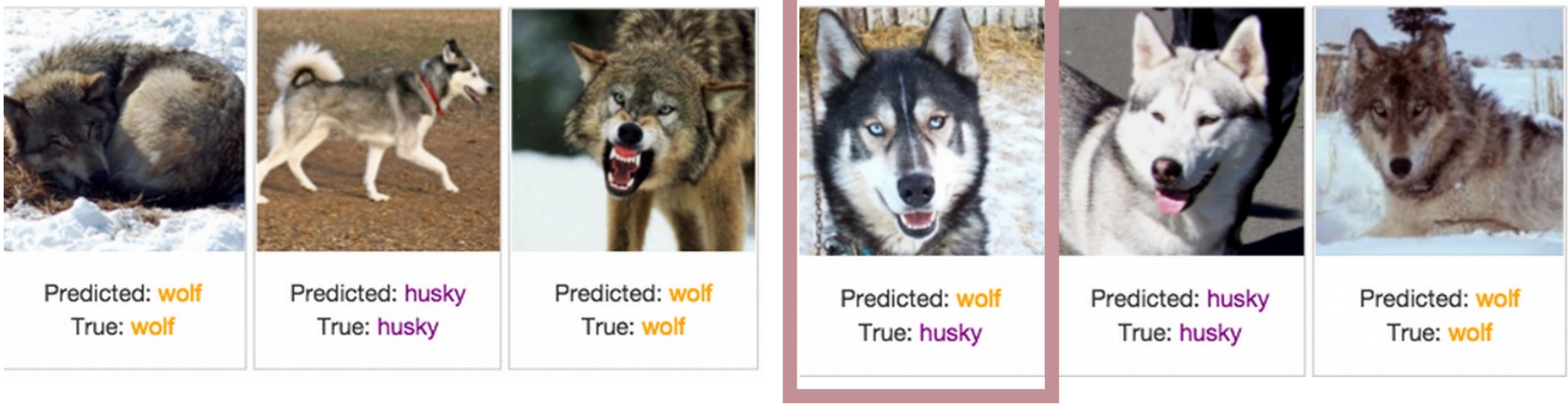


Human-machine collaboration

1. AI-systems need to support humans in understanding them
2. AI-systems need to be able to understand humans

Transparency and explainability

- AI systems often operate as 'black boxes', lacking transparency.
- This makes it difficult for users to understand and trust the system.
- Explainable AI (XAI) seeks to provide clarity and justification for AI decisions.



Explainable AI

(a) Husky classified as wolf (b) Explanation

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

AI used in many application areas (automation – human support, LoA)

Empirical studies

The brains…

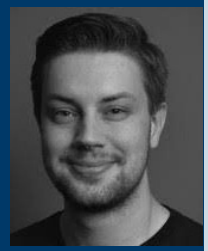
People train robots by demonstration, and robots train other robots.

Reinforcement learning

AI – human communication decision-making

What is my neural network actually learning?

Visualising concepts in NN



Mental models, theory of mind, expectations

Human Robot Interaction

Expectations

PostDoc

Detection of defects in manufacturing

Automatic product recognition.

Computer Vision, fine-grained techniques and data generation

ML for forecasting and planning

AI adoption

What are explanations from AI-systems good for, anyway?

- Explanations lead to **positive results** (better understanding, trust, higher confidence in own decisions, satisfaction, performance, better mental models) but also….

negative effects or trade-offs

- … revealing limitations led to negative heuristics, under reliance
- … persuasion (follow advice even if it is incorrect, advice-taking) and overreliance
- … unnecessary explanations lead to higher cognitive load, information overload, more time
- … confusion
- … perceived accuracy is more important than explainability, no explanations needed

Trust. The relationship between explainability and trust is difficult to comprehend… (trust calibration)

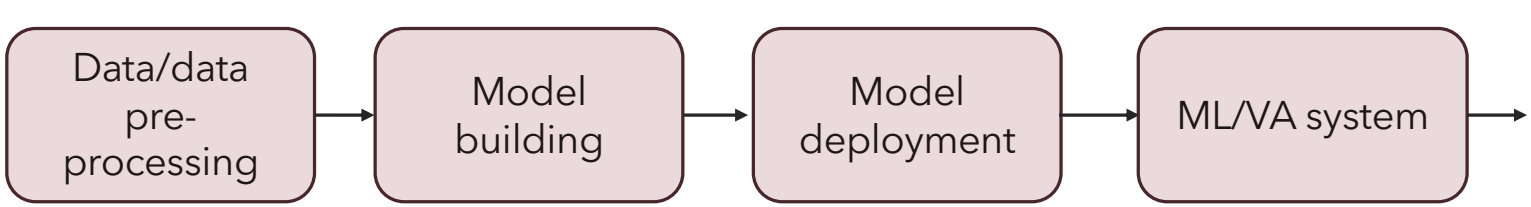
Explaining something to someone is a **complex cognitive process**… (not only XAI-ML)

XAI – a solution looking for a problem?

Expert systems (80s)

XAI design space is complex …

WHY, WHAT, WHEN, WHERE & HOW?



Type of user

ML-expert
Domain expert
Novice

Data types

Text
Images
Time-series
Geographical
Numerical
Categorical

Task to be supported (ML and user task)

Classification
Prediction
Clustering
Finding anomalies
Finding associations
Causal relations

Presentation

Visual encodings
Interaction methods
Multimodal

Decision-making

Context

Time to make a decision
Effects of wrong decision
Consequences –high stakes
Accountability
Expectations

Overall…

Better decisions?
Acceptance
Satisfaction
Understanding
Mental models
Trust/calibrated trust

Uncertainty

Uncertainty types (many concepts associated with uncertainty)
Propagates, aggregates..

Empirical studies

Expectations

- Riveiro, M., and Thill, S. (2021). “That's (not) the output I expected!” On the role of end user expectations in creating explanations of AI systems. Artificial Intelligence, Elsevier, ISSN 0004-3702, E-ISSN 1872-7921, Vol. 298, article id 103507
- Riveiro, M., and Thill, S. (2022). The challenges of providing explanations of AI systems when they do not behave like users expect. New York: Association for Computing Machinery (ACM), UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4-7, 2022

Task difficulty

- Ingesson, E. and Riveiro, M (2025 to be submitted) When Do We (Not) Want Explanations? A Study on Explanation Demand in Human-AI Decision-Making.

Human robot interaction – explaining errors

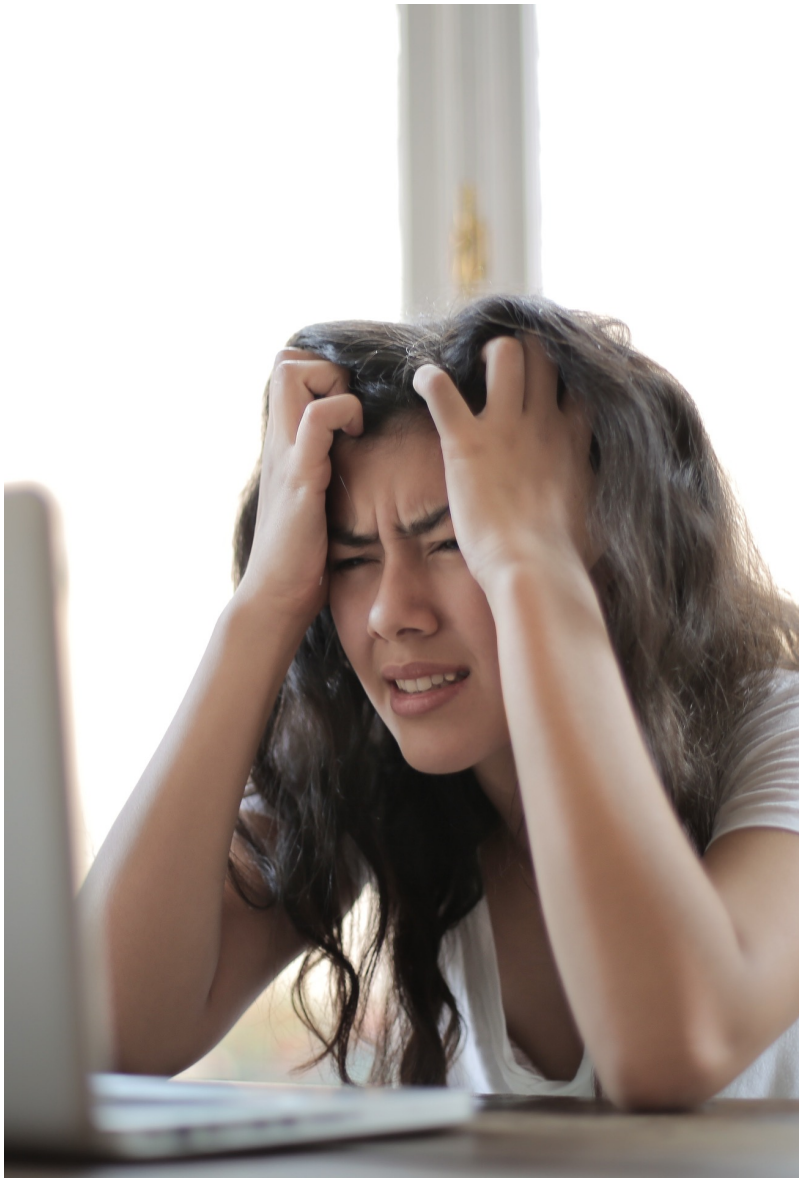
- Akalin, N. and Riveiro, M (2025 ROMAN). Let Me Explain Why I didn’t Take the Action You Wanted! Comparing Different Modalities for Explanations in Human-Robot Interaction

Chatbot Alba mental health support

- Holmberg, L., Sikström, S. and Riveiro, M. Speaking or Writing? Do Response Times Influence Anthropomorphism Differently for ADHD and Neurotypical Users in a Mental Health Chatbot? (2025 under review) Conversational User Interfaces.

How empirical studies inform theory

- Riveiro, M and Thill, S. (2025 under review) The diversity of empirical research on explainable artificial intelligence and implications for theory building. ACM Transactions on Interactive Intelligent Systems.



Expectations

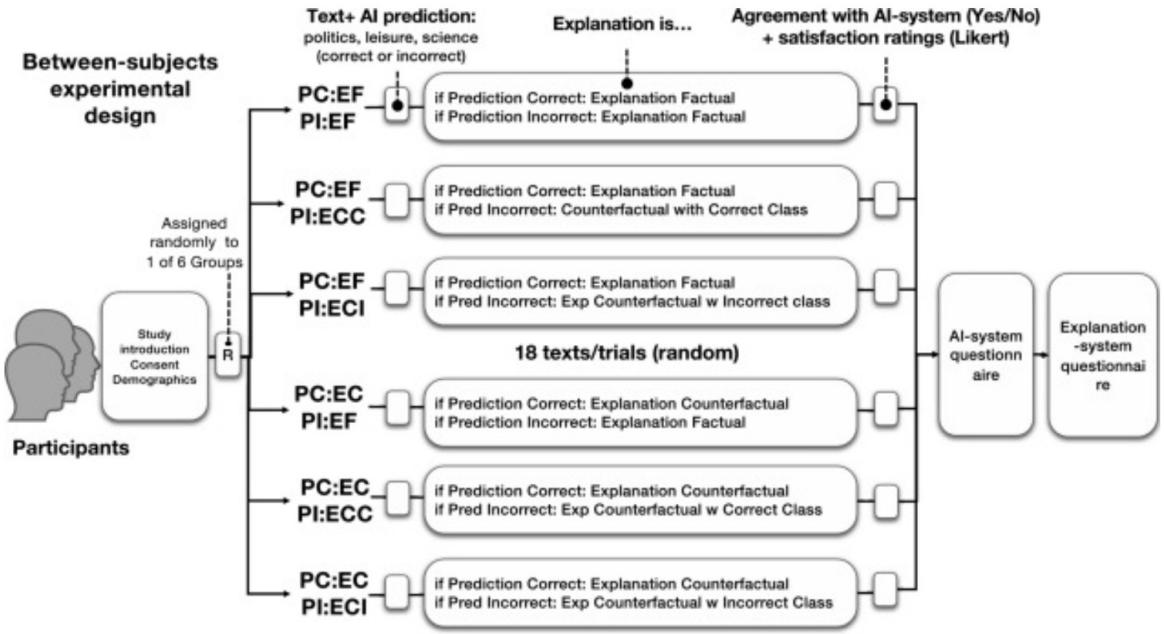
- In human-human interactions, explanations are often needed when an event is unexpected (Why?), and we need to explain the unexpected fact in relation to an implicit expected foil

- Do expectations play a role in when and what? Do they modulate the content of explanations? If we don't consider them, do we risk that you are not getting the explanation you are looking for?

- So…. what do we want to see in the explanations when we don't agree with the system/when it does something that we don't expect?

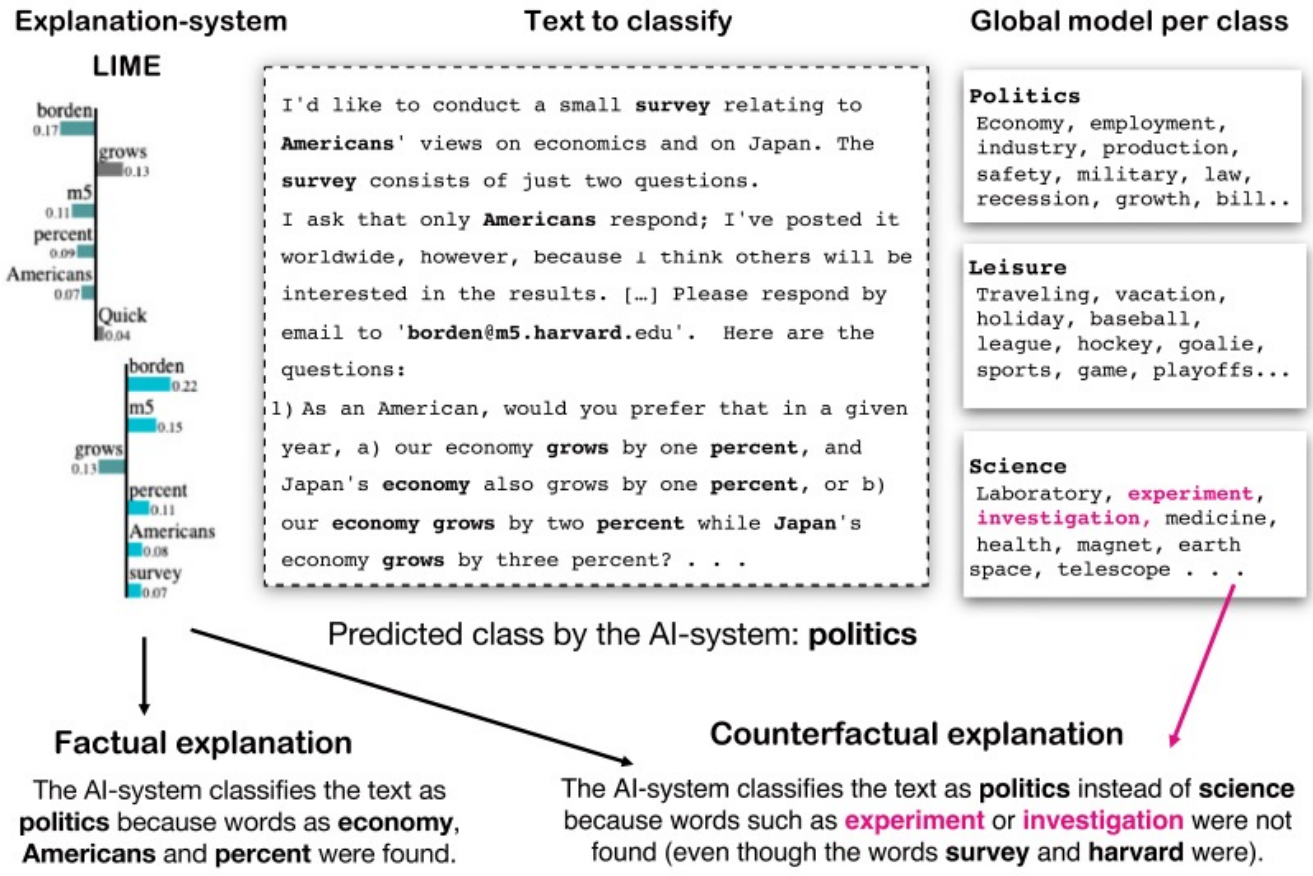
Measures/metrics

- System understanding
- Explanation satisfaction, completeness
- Performance
- Perceived need for interaction



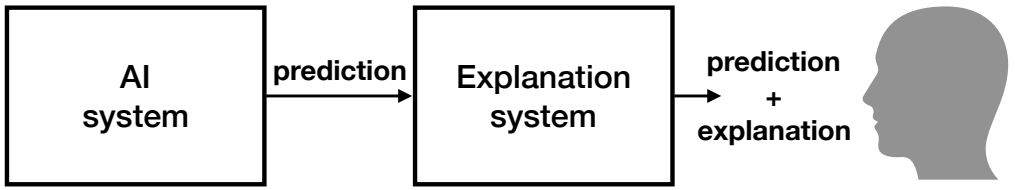
Example. Prediction Correct (PC): politics. Prediction Incorrect (PI): science or leisure.

- Explanation Factual (**EF**): The AI-system classifies the text as *politics* because words as *economy*, *Americans* and *percent* were found.
- Explanation Counterfactual (**EC**): The AI-system classifies the text as *politics* instead of *science* because words such as *experiment* or *investigation* were not found (even though the words *survey* and *harvard* were).
- Explanation Counterfactual with Correct Class (**ECC**): The AI-system classifies the text as *science/leisure* instead of *politics* because words such as *financial* or *growth* were not found (even though the words *American* and *percent* were).
- Explanation Counterfactual with Incorrect Class (**ECI**): The AI-system classifies the text as *leisure* instead of *science* because words such as *experiment* or *investigation* were not found (even though the words *survey* and *harvard* were).


Explanation-system
LIME
borden 0.17
grows 0.13
m5 0.11
percent 0.09
Americans 0.07
Quick 0.04
borden 0.22
m5 0.15
grows 0.13
percent 0.11
Americans 0.08
survey 0.07
Text to classify
I'd like to conduct a small survey relating to Americans' views on economics and on Japan. The survey consists of just two questions.
I ask that only Americans respond; I've posted it worldwide, however, because i think others will be interested in the results. […] Please respond by email to 'borden@m5.harvard.edu'. Here are the questions:
1) As an American, would you prefer that in a given year, a) our economy grows by one percent, and Japan's economy also grows by one percent, or b) our economy grows by two percent while Japan's economy grows by three percent? . . .
Global model per class
Politics
Economy, employment, industry, production, safety, military, law, recession, growth, bill..
Leisure
Traveling, vacation, holiday, baseball, league, hockey, goalie, sports, game, playoffs...
Science
Laboratory, experiment, investigation, medicine, health, magnet, earth space, telescope . . .
Predicted class by the AI-system: politics
Factual explanation
The AI-system classifies the text as politics because words as economy, Americans and percent were found.
Counterfactual explanation
The AI-system classifies the text as politics instead of science because words such as experiment or investigation were not found (even though the words survey and harvard were).

Role of expectations in explanations

- Do expectations determine explanation content?
- Are counterfactuals preferred when outcomes from AI-system are unexpected?



- Factual and counterfactual explanations

✓ • H1: Factual explanations are appropriate for correct predictions because the system output is in line with the expected output.

✗ • H2: Counterfactual explanations that contain the expected foil are appropriate when the system prediction is incorrect

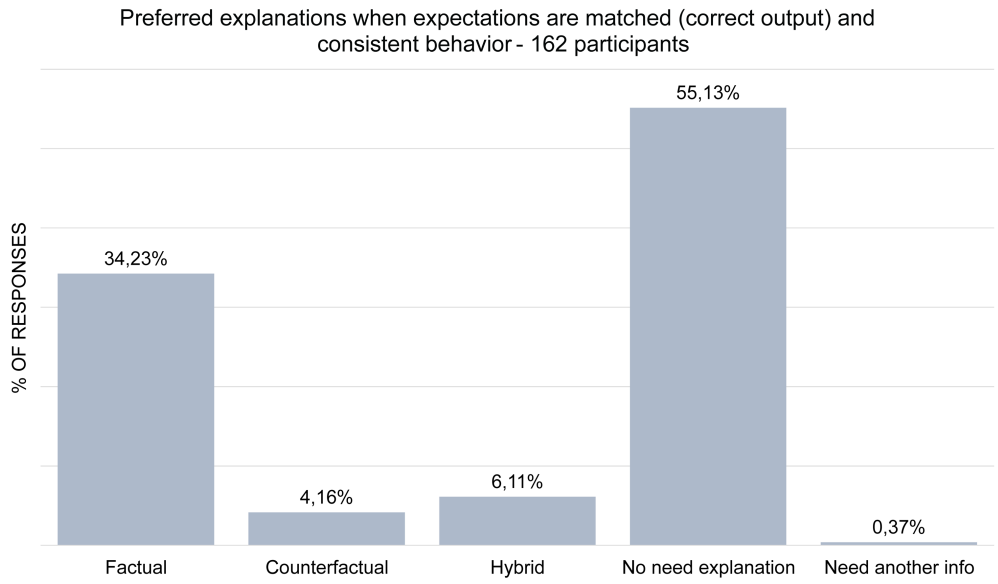
So…. what do we want to see in the explanations when the system does something that we don’t expect?

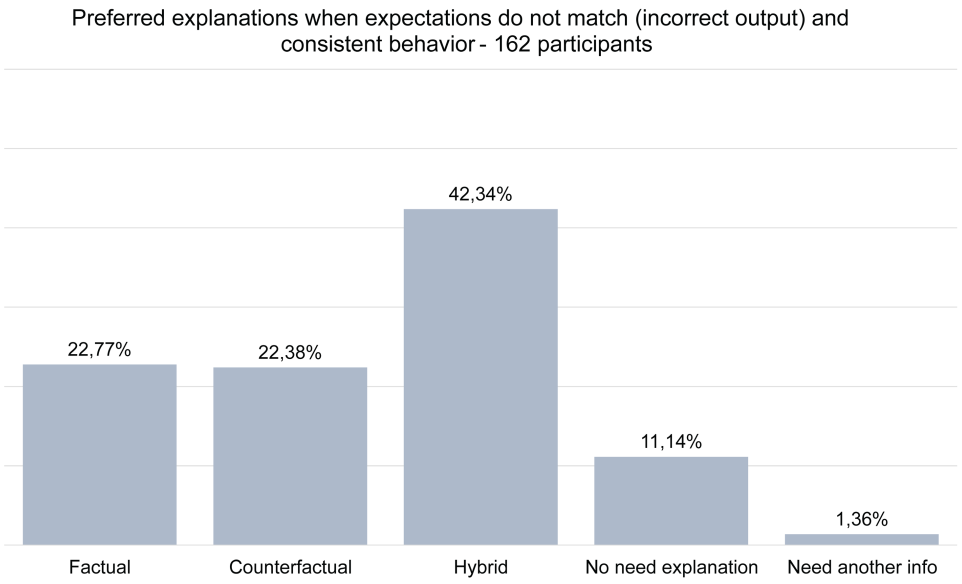
Method

- We presented participants with various scenarios involving a text classifier and then asked them to indicate their preferred explanation for each scenario
- One group of participants chose the type of explanation from a multiple-choice questionnaire (Study I), the other had to answer using free text (Study II)

162
15 (+ 2)

STUDY I





Matched expectations

Mismatched expectations

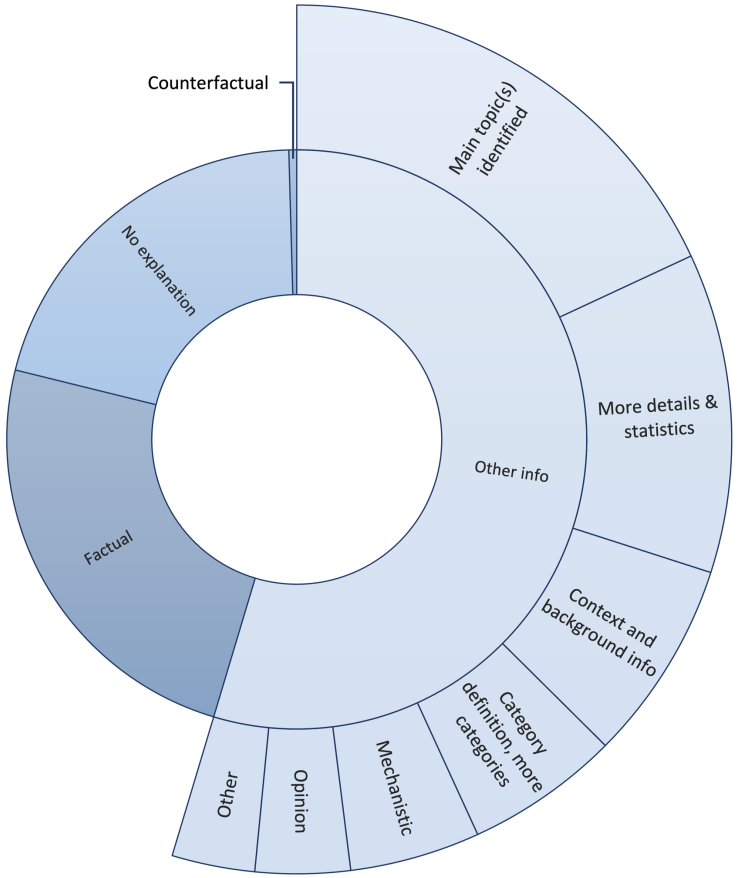
Table 2. Main categories found under class "other information" in response to"What kind of information you would like to see in the explanations from the AI system in order to understand this prediction?" when the AI system output matched user expectations. n=126.

Category	Description	Count	Example evidence from responses
Main topic(s) identified	A summary of the main themes, overall topic found in the text	41	*"Identify theme and central ideas".* *"About UN to establish a police force for Haiti".* *"Politics, colonial, occupation".*
More details and statistics	More details, more evidence, statistics	27	*"More information in regards to the smaller details of the text".* *"I'd like to know the ingredients of the medicine. Perhaps their specific function."* *"Statistics."* *"Also, the percentage of key words that were mapped to the category (assuming that the AI is running on some sort of key word matching logic)."*
Context and background info	Context, background info on article, where this text comes from, reference	17	*"The source of the article and the names of the people having the conversation."* *"Research and references, something that can actually prove this is the case.'* *"It seems like it's missing context."'*
Category definition, more categories	Unclear how the AI system defined each category, more categories needed	13	*"Explain how Sports is Leisure."* *"Sub topic of politics."* *"What kind of politics."*
Mechanistic	How the AI system reached that conclusion, reasoning	11	*"How it came to this conclusion."* *"Showing the reasoning behind the decisions if able to do so."*
Opinion about the text	Participants expressed their opinions about topics in text	8	*"It should be said how the UN predicts policies."* *"I would love to see the breakdown of how taxes are being spent ."*
Other/Not relevant	Comments not related to explanations	7	*"Don't understand the text ."*

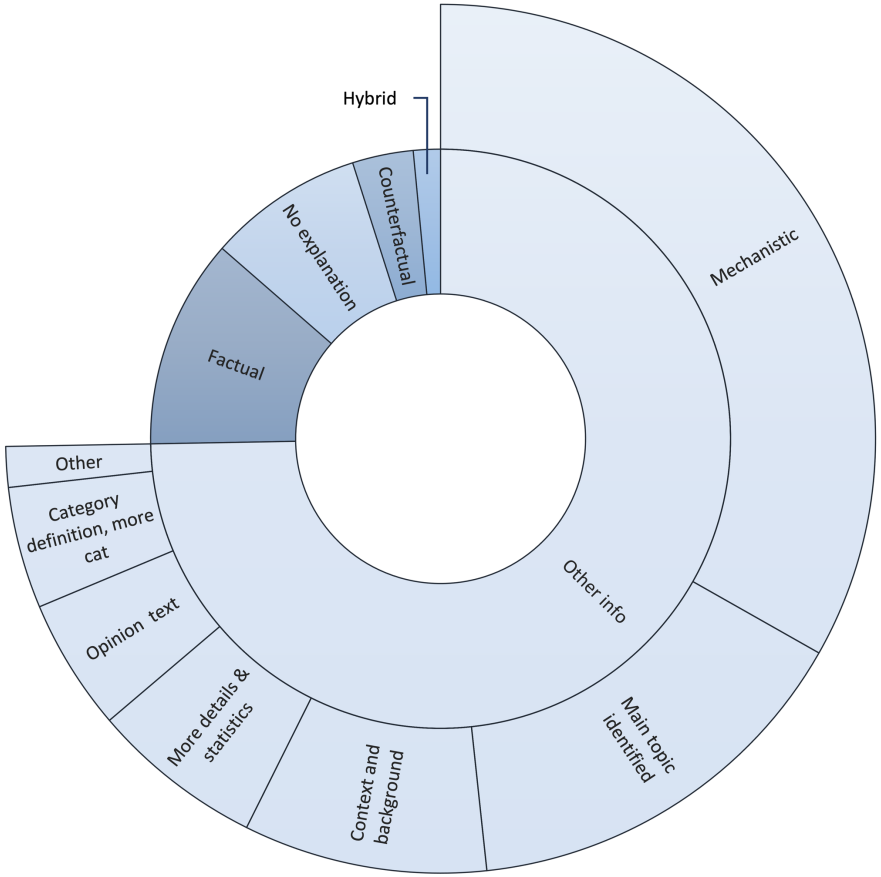
Table 3. Main categories found under class "other information" in response to "What kind of information you would like to see in the explanations from the AI system in order to understand this prediction?" when the output from the AI system did not match user expectations. n=198.

Category	Description	Count	Example evidence from responses
Mechanistic	How the AI system reached that conclusion, reasoning	88	*"I don't understand what the AI system would think this is leisure. How exactly would a gun buyback program be of any leisure ?"* *"What got you to believe this was politics?"* *"I would like the AI system to list reason for suggesting the prediction as Politics."*
Main topic(s) identified	A summary of the main themes, overall topic found in the text	40	*"The effects of gun buyback program and how it affects citizens."* *"It's about been tour round the world."*
Context and background info	Context, background info on article, where this text comes from, reference	24	*"I would love to know information source about the fail-safe mechanism."* *"I would like to see some reasons that justifies the prediction like the context in which the discussion was made."* *"How the ecosystem in Utah works, and the climate of Utah."*
More details and statistics	More evidence and statistics	17	*"I would love to know more about NOOP operation."* *"More clearer information such as locations."* *"Statistics."*
Opinion about the text	Participants expressed their opinions about topics in text	13	*"Gun sport maybe leisure to some people but this is his opinion more than anything else."*
Category definition, more categories	Unclear how the AI system defined each category, more categories needed	12	*"Language selected to recognise leisure, what is 'leisure' by definition."* *"What leisure activity is referred to."*
Other/Not relevant	Comments not related to explanations	4	*"Better paragraph structure."*

Content of explanations when expectations are matched



Content of explanations when expectations are **not** matched



Matched expectations

Mismatched expectations

Conclusions from expectations

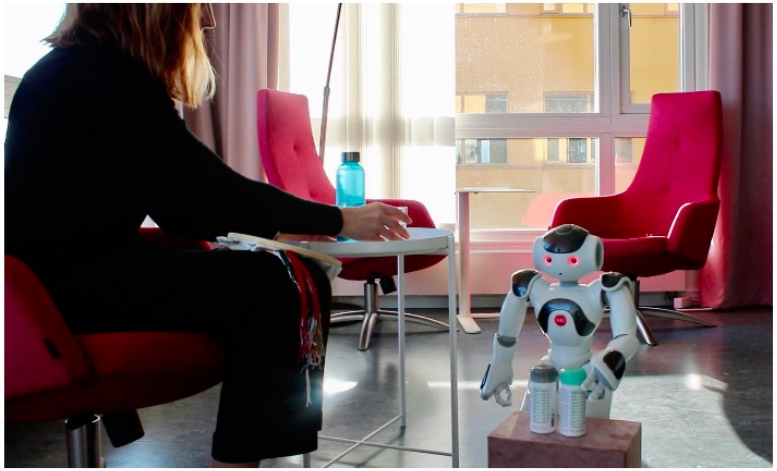
- For matched expectations, an explanation is often not required at all, while if one is, it is of the factual type
- Providing explanations when system output does not match user expectations is a challenging matter, primarily because there does not seem to be a unique strategy, although mechanistic explanations are requested more often than other types
- No one size fits all
- Overall, user expectations are a significant variable in determining the most suitable content of explanations (including whether an explanation is needed at all)

Akalin, N. and Riveiro, M (2025 ROMAN). **Let Me Explain Why I didn't Take the Action You Wanted!** Comparing Different Modalities for Explanations in Human-Robot Interaction

- Aim: Explore preferred modalities for robot explanations when robots decline user requests. We focus on explanations in scenarios where a user makes a request to the robot, but the robot does not perform the requested action for various reasons
- Method: User study assessing various explanation modalities (visual, auditory, gesture).

Results:

- Participants strongly preferred speech-based explanations for clarity, naturalness, and ease of understanding.
- Multimodal explanations (combining speech with lights, sounds, or gestures) were preferred for critical or urgent situations to ensure attention and clarity.
- User preferences for explanation modalities varied according to context, such as urgency, ambient noise, or user's distance from the robot.
- Overall, users found explanations clear and helpful, but desired improvements for more human-like or customizable voices.
- Users requested more interactive explanations, such as allowing follow-up questions.



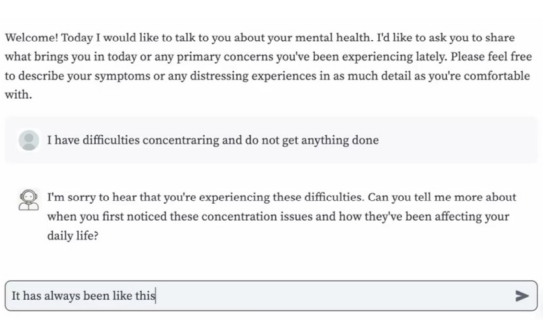
Holmberg, L., Sikström, S. and Riveiro, M. Speaking or Writing? **Do Response Times Influence Anthropomorphism Differently for ADHD and Neurotypical Users in a Mental Health Chatbot?** (2025 under review) Conversational User Interfaces.

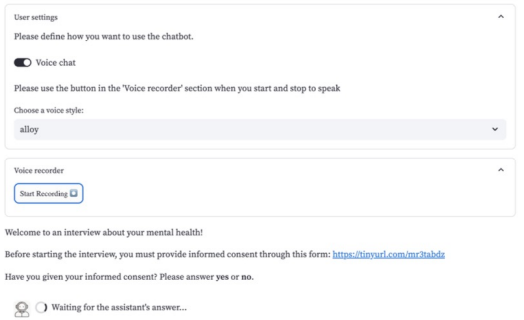
- Aim: Investigate how interaction modality (text vs. speech) affects anthropomorphism of chatbots among neurotypical and ADHD users.
- Method: User study with 107 participants interacting via text or speech.
- Results: Neurotypical users equally anthropomorphize chatbots in both conditions, while ADHD users anthropomorphize less in voice due to longer response times.

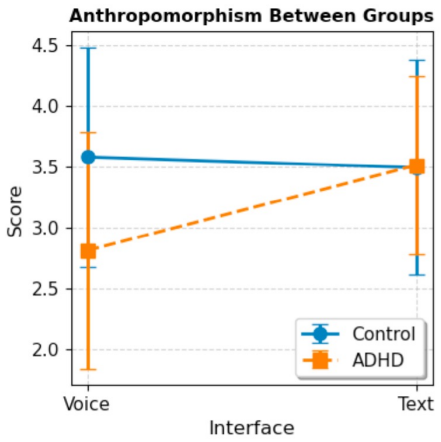
Chatbot Alba mental health support

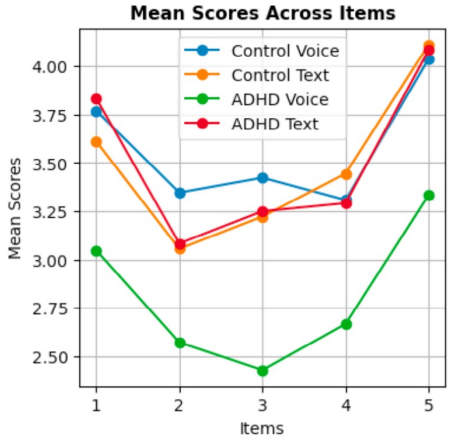
AI CBT: Cognitive behavioral therapy with an AI-therapist

talktoalba

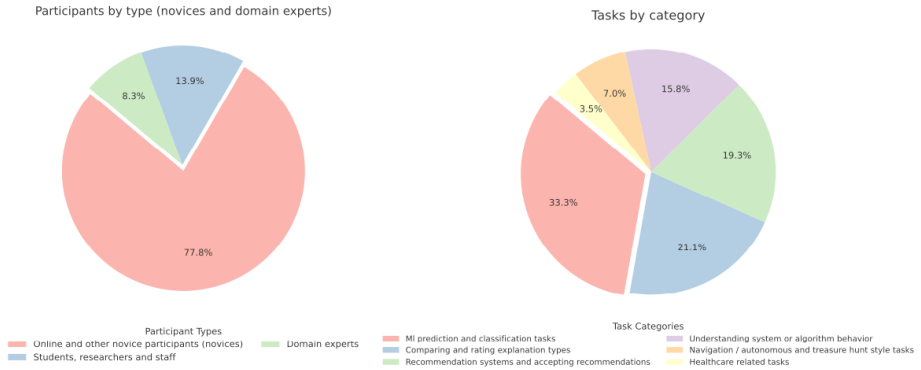








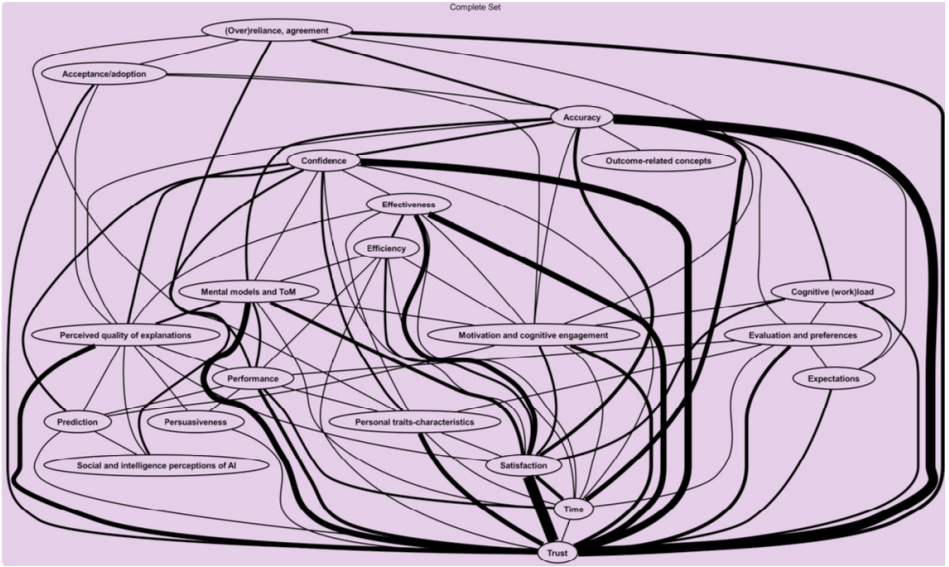
Riveiro, M and Thill, S. (2025 under review) **The diversity of empirical research on explainable artificial intelligence and implications for theory building**. ACM Transactions on Interactive Intelligent Systems.



- Aim: Overview and summary of empirical studies in XAI, analyzing contexts, purposes, and effects of explanations.
- Method: Reviewed 95 empirical studies evaluating explanations in human-AI interaction.

Results:

- Lack of ecological valid experiments.
- XAI research is highly diverse, lacking a common theoretical framework.
- Explanations serve multiple purposes (trust, understanding, decision support) but show varied effectiveness.
- ... principles for deriving a general theory of explanations from AI-systems?

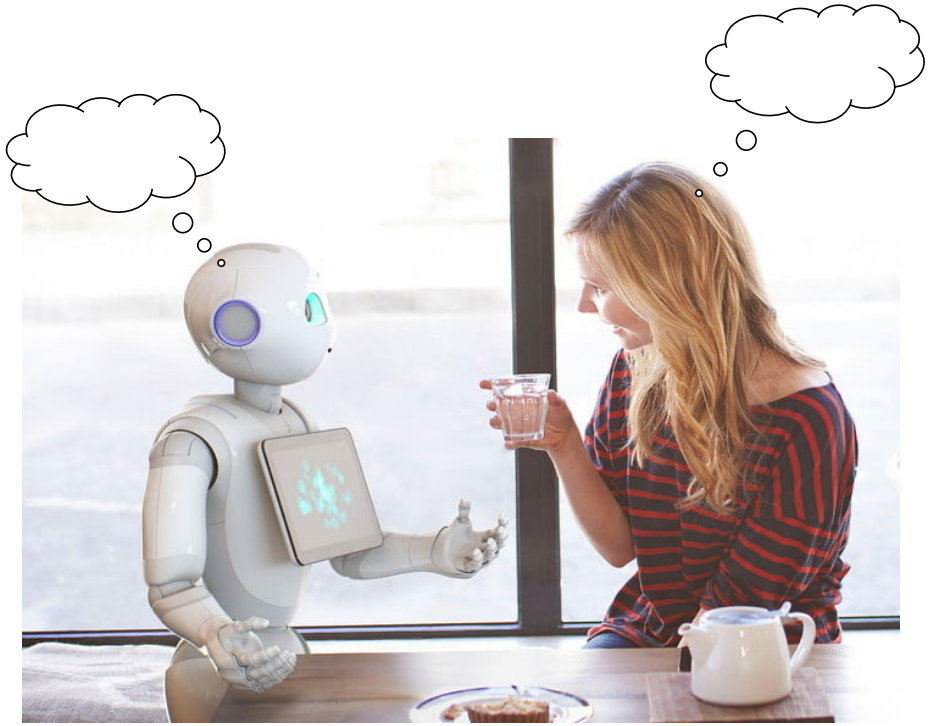


Human-machine collaboration

1. AI-systems need to support humans in understanding them

2. AI-systems need to be able to understand humans

AI-systems need to be able to understand humans



Adaptation

- Human-AI collaboration (my stand is that it mirrors human-human interactions)
- Understand users (needs, expectations, abilities, personality traits) and adapt interactions accordingly

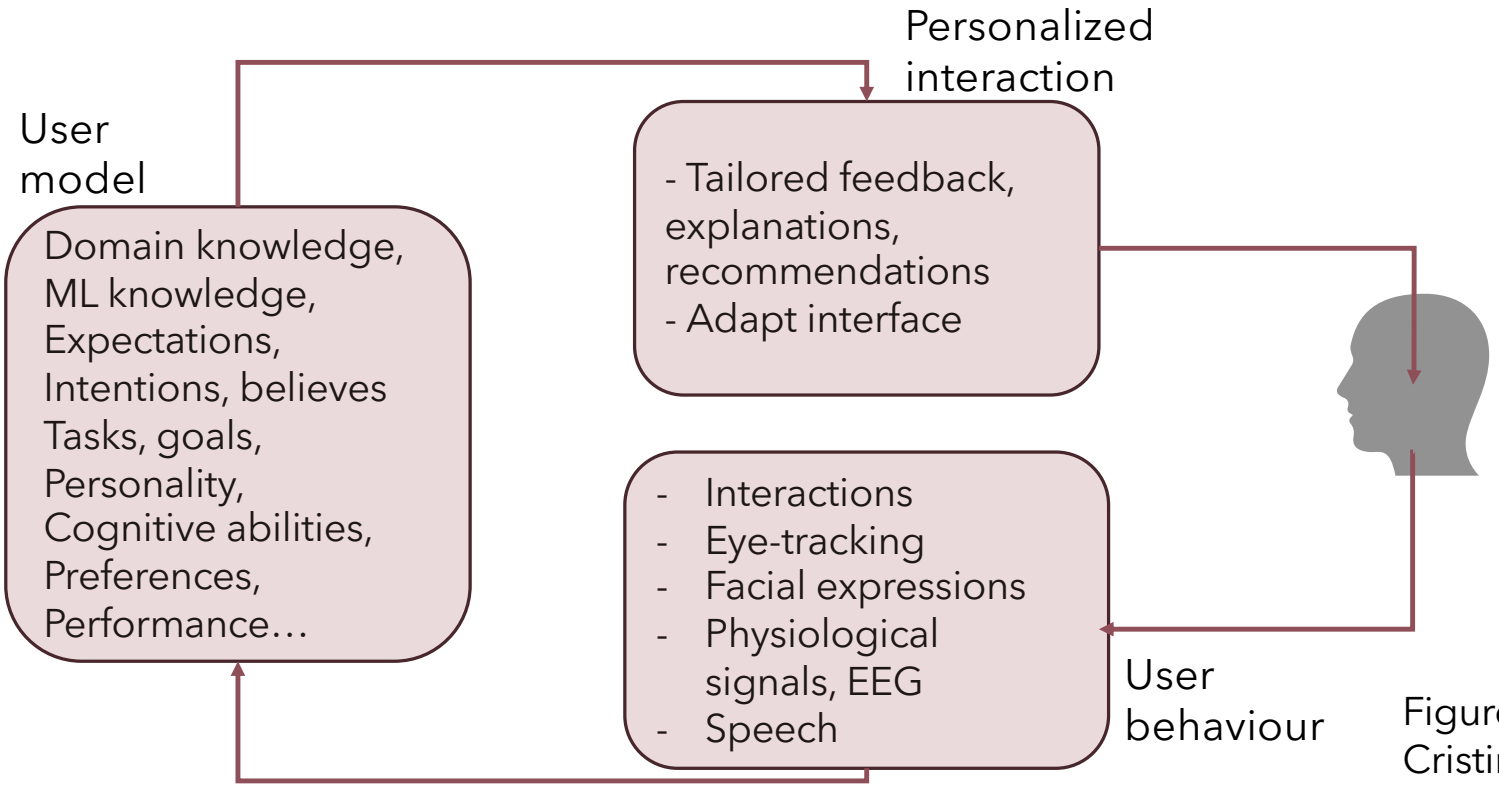
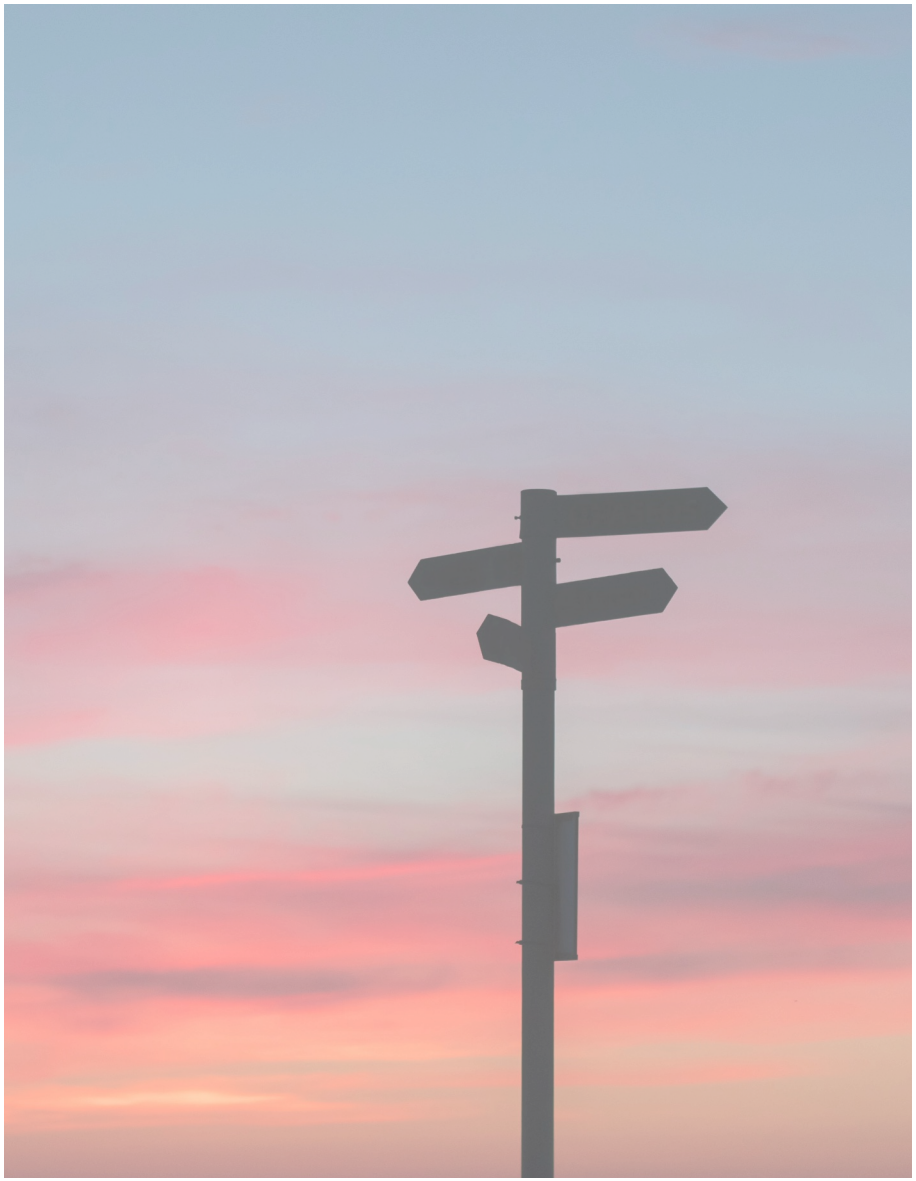


Figure inspired by Cristina Conati's work



Future

- User models, mental models
- Expectations
- Theory of Mind
- Intention recognition
- Visualization, presentation and interaction modalities
- Adaptation: engagement, curiosity, knowledge
- Use cases!


Thanks for listening!
JÖNKÖPING UNIVERSITY

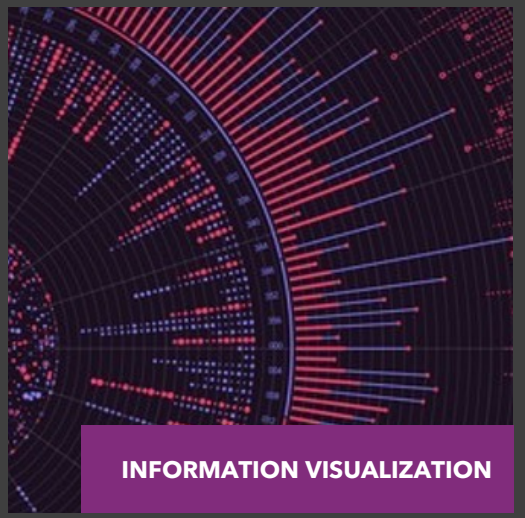
JÖNKÖPING UNIVERSITY

Human-Centered Technology (HCT)

HUMAN-COMPUTER COMMUNICATION

EDUCATION & HEALTHCARE

SMART INDUSTRIES



INFORMATION VISUALIZATION