



Explanations, expectations and empirical studies

MARIA RIVEIRO

SCHOOL OF ENGINEERING
JÖNKÖPING UNIVERSITY, SWEDEN

MT4H WORKSHOP VALENCIA 2024

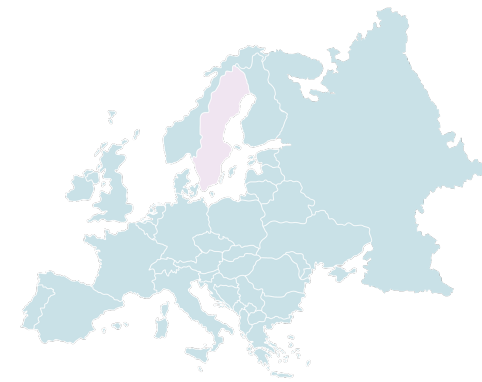


Explanations and evaluation
(EXPLAIN)

Tailoring explanations from AI
systems to users' expectations
(XPECT)

Collaboration with Serge Thill





Jönköping AI Lab

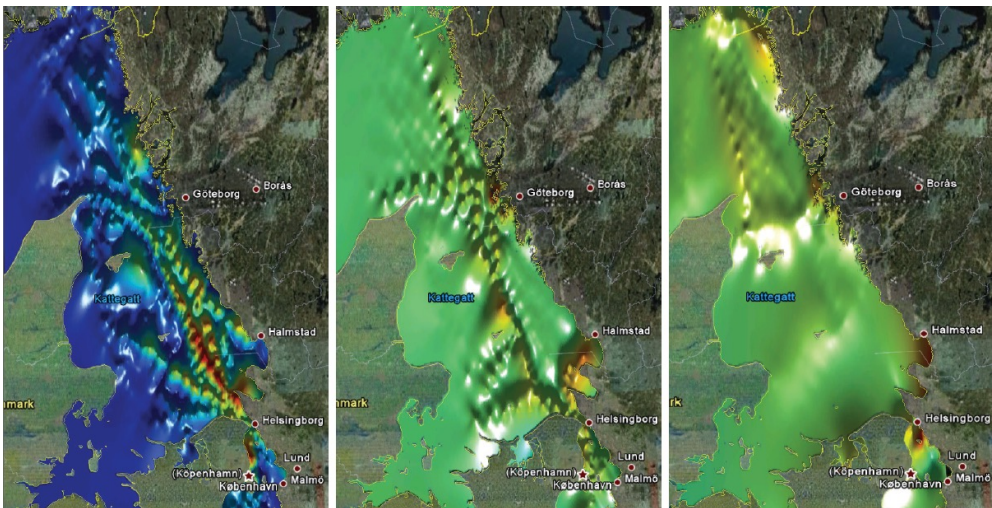
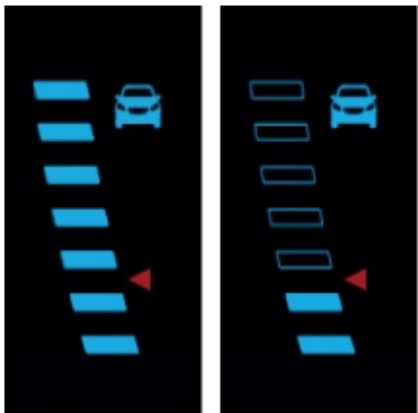
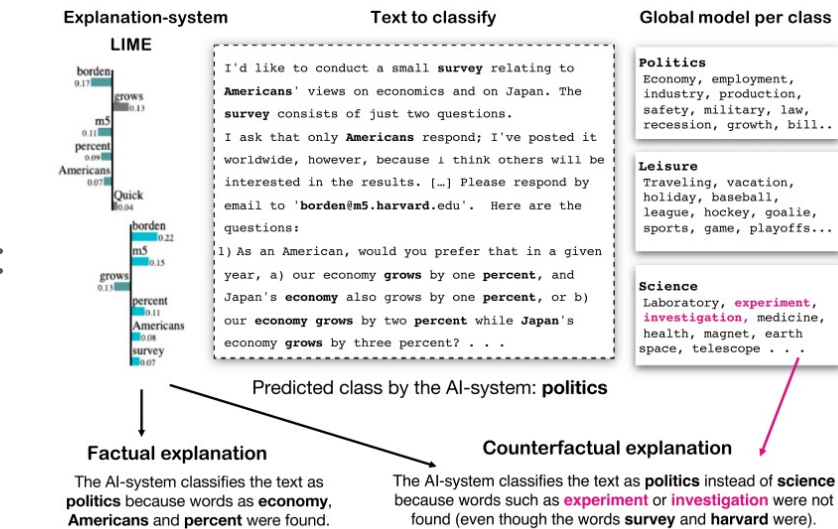
Human-Centered Technology Group

School of Engineering

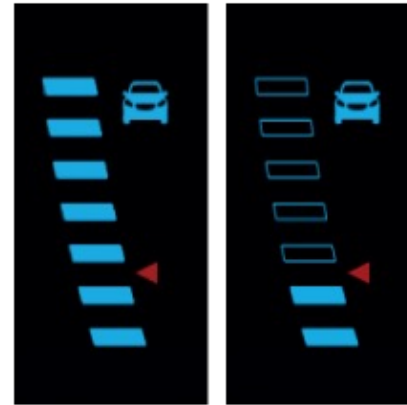


EXPLAINABLE AI

- Empirical evaluations
- Human-Centered AI
- Collaboration with several companies: Saab, Volvo, AstraZeneca, Husqvarna



Autonomous driving



Results:

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

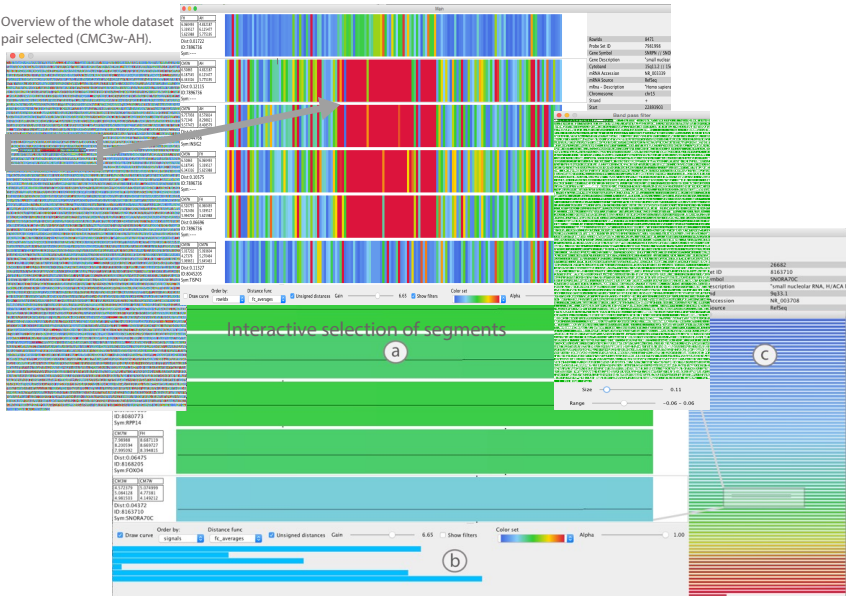
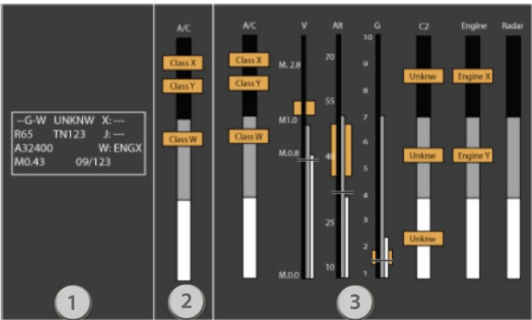
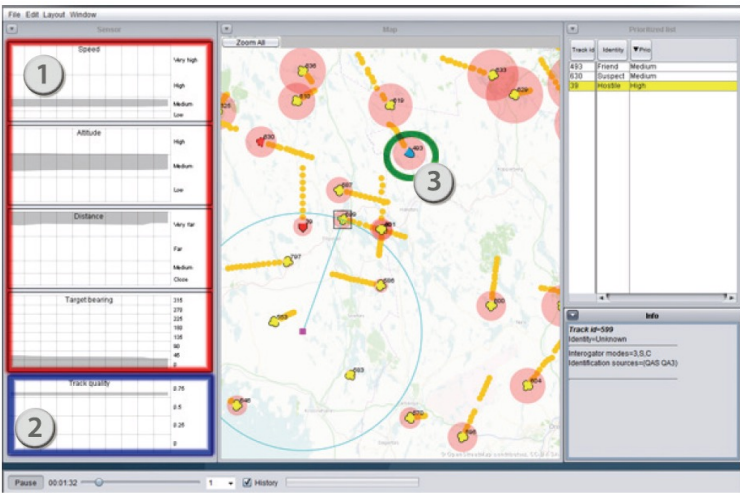
Eco-driving



Results:

More adherence to recommendations if they are explained

Similar results in other domains ...



Higher trust but more time...

Is it always good to show explanations?

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

negative effects or trade-offs

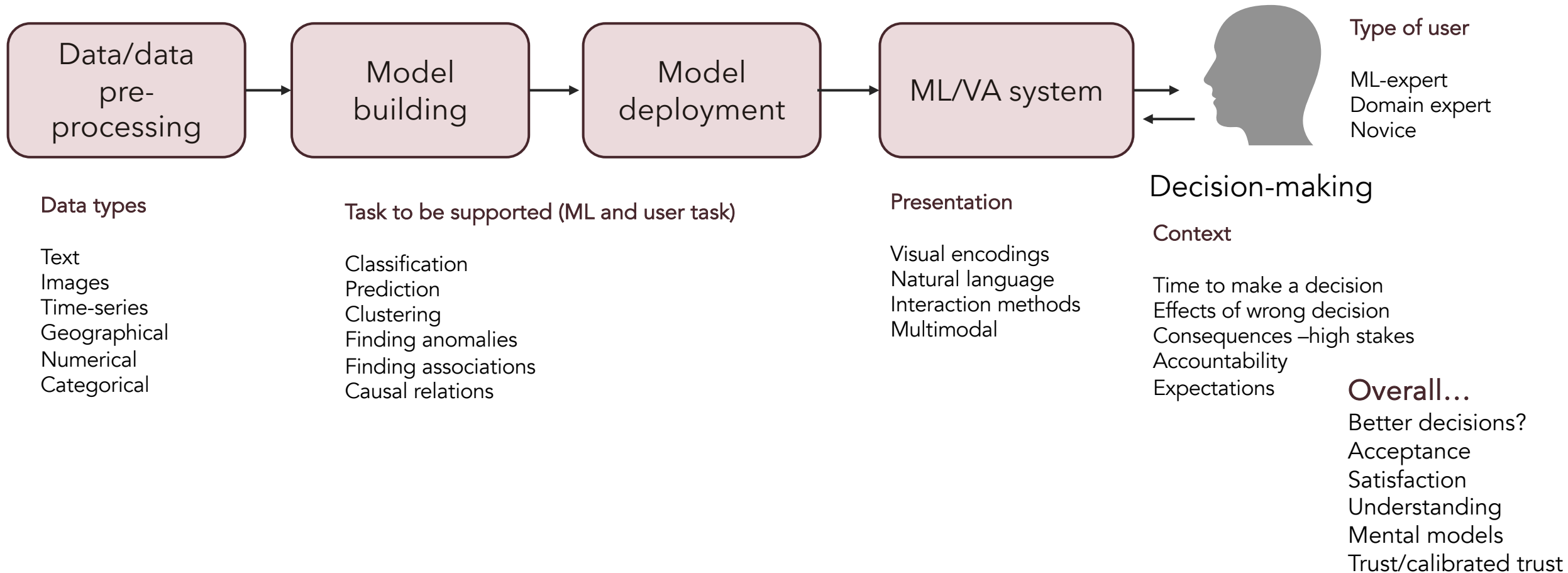
- revealing limitations led to negative heuristics, under reliance
- ... unnecessary explanations lead to higher cognitive load, information overload, more time
- ... confusion
- .. perceived accuracy is more important than explainability, no explanations needed
- .. persuasion (follow advice even if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

Trust. The relationship between explainability and trust is difficult to comprehend...

**.... principles for deriving a general theory
of explanations from AI-systems?**

XAI design space is complex ...

WHY, WHAT, WHEN, WHERE & HOW?



explaining something to someone is a complex cognitive process...



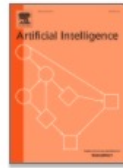
Expectations

- 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?

Role of expectations in building explanations from AI-systems



Artificial Intelligence
Volume 298, September 2021, 103507



“That's (not) the output I expected!” On the role of end user expectations in creating explanations of AI systems ☆

Maria Riveiro ^a  , Serge Thill ^b

> *Proceedings* > UMAP '22 > *The challenges of providing explanations of AI systems when they do not behave like users expect*

RESEARCH-ARTICLE OPEN ACCESS

The challenges of providing explanations of AI systems when they do not behave like users expect

Authors: [Maria Riveiro](#), [Serge Thill](#) [Authors Info & Claims](#)

UMAP '22: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization • July 2022 • Pages 110–120 • <https://doi.org/10.1145/3503252.3531306>

Online: 04 July 2022 [Publication History](#)

Motivation and aim (paper I)

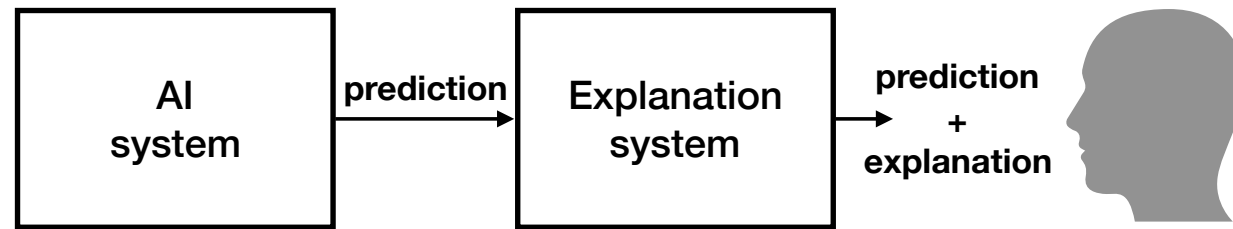
- Do expectations determine explanation content (what) and when?
- Are counterfactuals preferred when outcomes from AI-system are unexpected?

Type of explanations (what)

- Local and global
- Mechanistic
- Functional
- Factual
- Counterfactual

Role of expectations in explanations

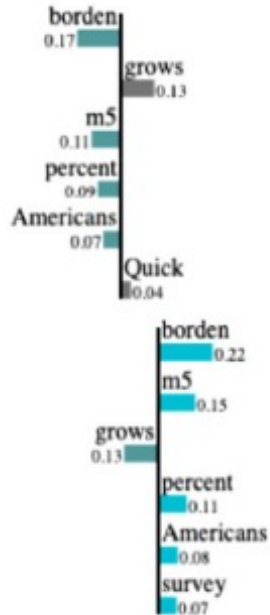
- Do expectations determine explanation content?
- Are counterfactuals preferred when outcomes from AI-system are unexpected?



- 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

Explanation-system

LIME



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 . . .

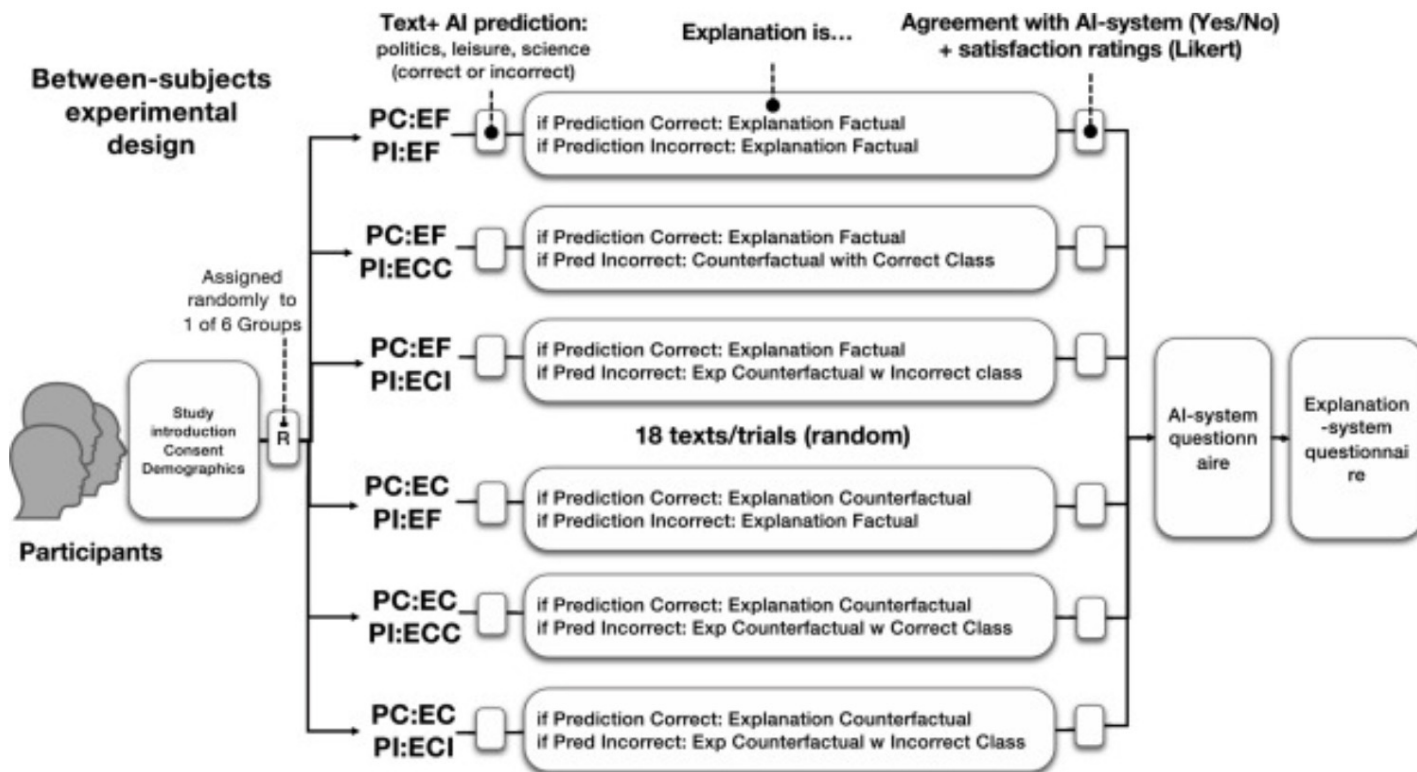
Predicted class by the AI-system: **politics**

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).



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

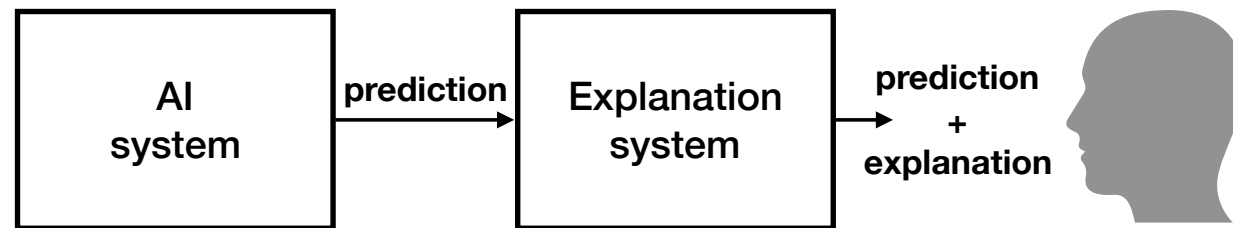
- Explanation Factual (EF): The AI-system classifies the text as *politics* because words as *economy*, *Americans* and *percent* were found.
- Explanation Counterfactual (EC): The AI-system classifies the text as *politics* instead of *science* because words such as *experiment* or *investigation* were 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).

Measures/metrics

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

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 we don't agree with the system/when it does
something that we don't expect?

Motivation and aim (paper II)

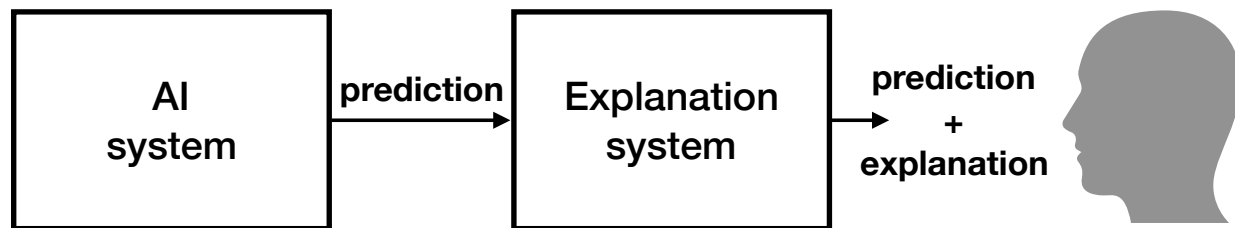
- Expectations that users may have about the system behaviour play a role since they co-determine appropriate content of the explanations
- **We investigate user-desired content of explanations when the system behaves in unexpected ways**

Method

STUDY I
(multiple-choice)

STUDY II
(open questions)

- 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)



0%  100%

EXAMPLE 2 of 2


Text to classify

No, if you put a conductor in a changing magnetic field, it produces a voltage. The two ways you can do that with a permanent magnet is to move the magnet or move the conductor. The slow shifting of the Earth's magnetic field isn't really significant, especially when you consider how weak the Earth's magnetic field is to begin with.

Well, it would require generating an incredibly large magnetic field to repel the Earth's magnetic field (as a magnet can repel another magnet). Of course, this force only works in one direction, and the magnetic field generated has to be unimaginably powerful. Magnetic repulsion drops off as $1/r^3$, and the earth's magnetic field on the surface is already very weak. It would require some sort of unknown superconductor, and special nonmagnetic construction. And seriously hardened electronics (optical computers, perhaps). And the physiological danger would be significant (due to the iron content in our blood, among other things). In other words, forget it.

I missed out on the "dragless satellite" thread, but it sounds totally bogus, from this little bit.

Predicted class by the AI-system: **leisure**.

0%  100%

EXAMPLE 2 of 2

Text to classify

No, if you put a conductor in a changing magnetic field, it produces a voltage. The two ways you can do that with a permanent magnet is to move the magnet or move the conductor. The slow shifting of the Earth's magnetic field isn't really significant, especially when you consider how weak the Earth's magnetic field is to begin with.

Well, it would require generating an incredibly large magnetic field to repel the Earth's magnetic field (as a magnet can repel another magnet). Of course, this force only works in one direction, and the magnetic field generated has to be unimaginably powerful. Magnetic repulsion drops off as $1/r^3$, and the earth's magnetic field on the surface is already very weak. It would require some sort of unknown superconductor, and special nonmagnetic construction. And seriously hardened electronics (optical computers, perhaps). And the physiological danger would be significant (due to the iron content in our blood, among other things). In other words, forget it.

I missed out on the "dragless satellite" thread, but it sounds totally bogus, from this little bit.

Predicted class by the AI-system: **leisure**.

Please, select the explanation that you would prefer to see in terms of understanding the system's output (examples in blue will not be shown during the actual test):

- ☐ An explanation that explains why the predicted class was selected. Example: The AI-system classifies the text as **politics** because words as **economy**, **Americans** and **percent** were found.
- ☐ An explanation that explains why another class was not selected. Example: The AI-system classifies the text as **politics** instead of **science** because words such as **experiment** or **investigation** were not found.
- ☐ An explanation that contains information about why the predicted class was selected and why another class was not. Example: The AI-system classifies the text as **politics** because words as **economy** and **percent** were found and not **science** since words as **experiment** or **investigation** were not found.
- ☐ I do not need an explanation.
- ☐ I need another type of information.

Factual

Counterfactual

Hybrid

No explanation

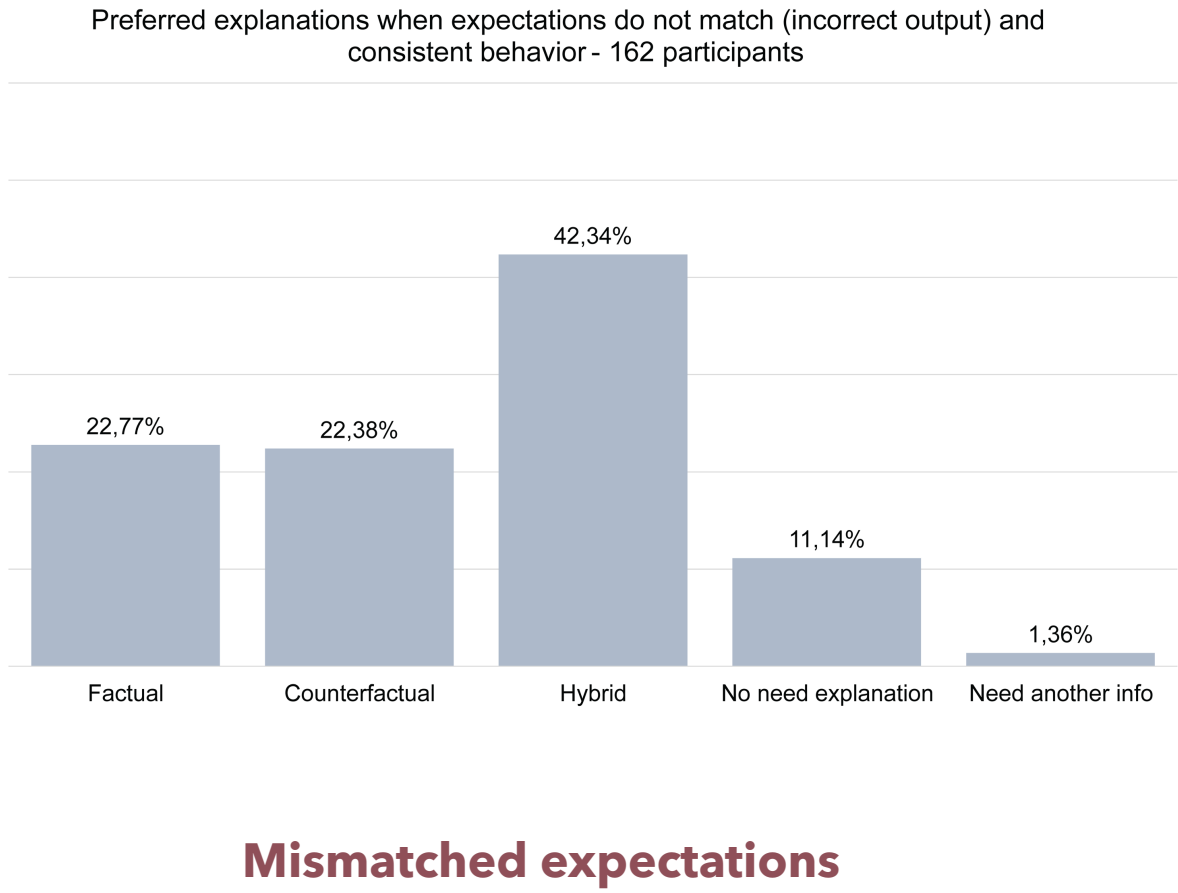
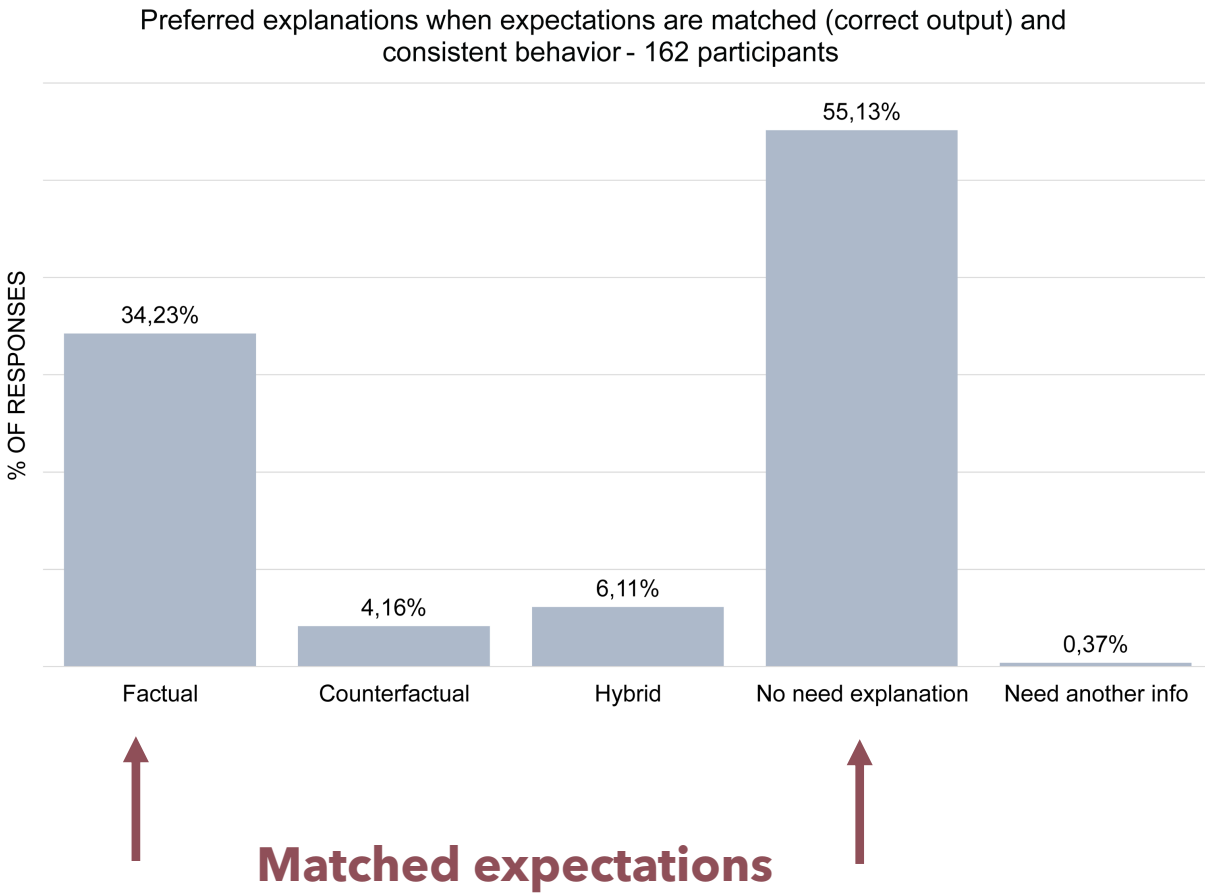
Other info

Do you agree with the classification made by the AI-system? If not, please, indicate which type of text you think it is.

- ☐ Yes, I agree
- ☐ No, it is science
- ☐ No, it is leisure

162
15 (+ 2)

STUDY I





STUDY II

0%  100%

EXAMPLE 2 of 2
Text to classify

No, if you put a conductor in a changing magnetic field, it produces a voltage. The two ways you can do that with a permanent magnet is to move the magnet or move the conductor. The slow shifting of the Earth's magnetic field isn't really significant, especially when you consider how weak the Earth's magnetic field is to begin with.

Well, it would require generating an incredibly large magnetic field to repel the Earth's magnetic field (as a magnet can repel another magnet). Of course, this force only works in one direction, and the magnetic field generated has to be unimaginably powerful. Magnetic repulsion drops off as $1/r^3$, and the earth's magnetic field on the surface is already very weak. It would require some sort of unknown superconductor, and special nonmagnetic construction. And seriously hardened electronics (optical computers, perhaps). And the physiological danger would be significant (due to the iron content in our blood, among other things). In other words, forget it.

I missed out on the "dragless satellite" thread, but it sounds totally bogus, from this little bit.

Predicted class by the AI-system: **leisure**.

What kind of information you would like to see in the explanations from the AI-system in order to understand this prediction?

Do you agree with the classification made by the AI-system? If not, please, indicate which type of text you think it is.

☐ Yes, I agree

☐ No, it is politics

☐ No, it is science

Table 1. “Factual”, “counterfactual”, “hybrid”, “no need explanation” and “other information” responses 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 (n=229) and did not match (n=265) user expectations.

SYSTEM OUTPUT MATCHED			SYSTEM OUTPUT DID NOT MATCH	
Category	Count	Example evidence from responses	Count	Example evidence from responses
Factual	55	<i>“The keywords used to come to this conclusion.” “Key words from the text which have been selected.”</i>	31	<i>“What keywords has it used.” “Key words, percentage of key words matched to the winning category.”</i>
Counterfactual	1	<i>“Although the passage is clearly about sport, is there anything else the AI system picked up to eliminate that it wasnt science or politics?”</i>	9	<i>“Why was science disregarded when there are a lot of technical jargon employed.”</i>
Hybrid	0		4	<i>“What has led the AI to classify as leisure and why politics was discounted.” “How the text relates to this category more than any other.”</i>
No need explanation	47	<i>“I don’t think anything is needed as it is clearly politics.”</i>	23	<i>“I dont think this needs more information.” “Its fine.”</i>
Other information	126	see Table 2	198	see Table 3

55 participants

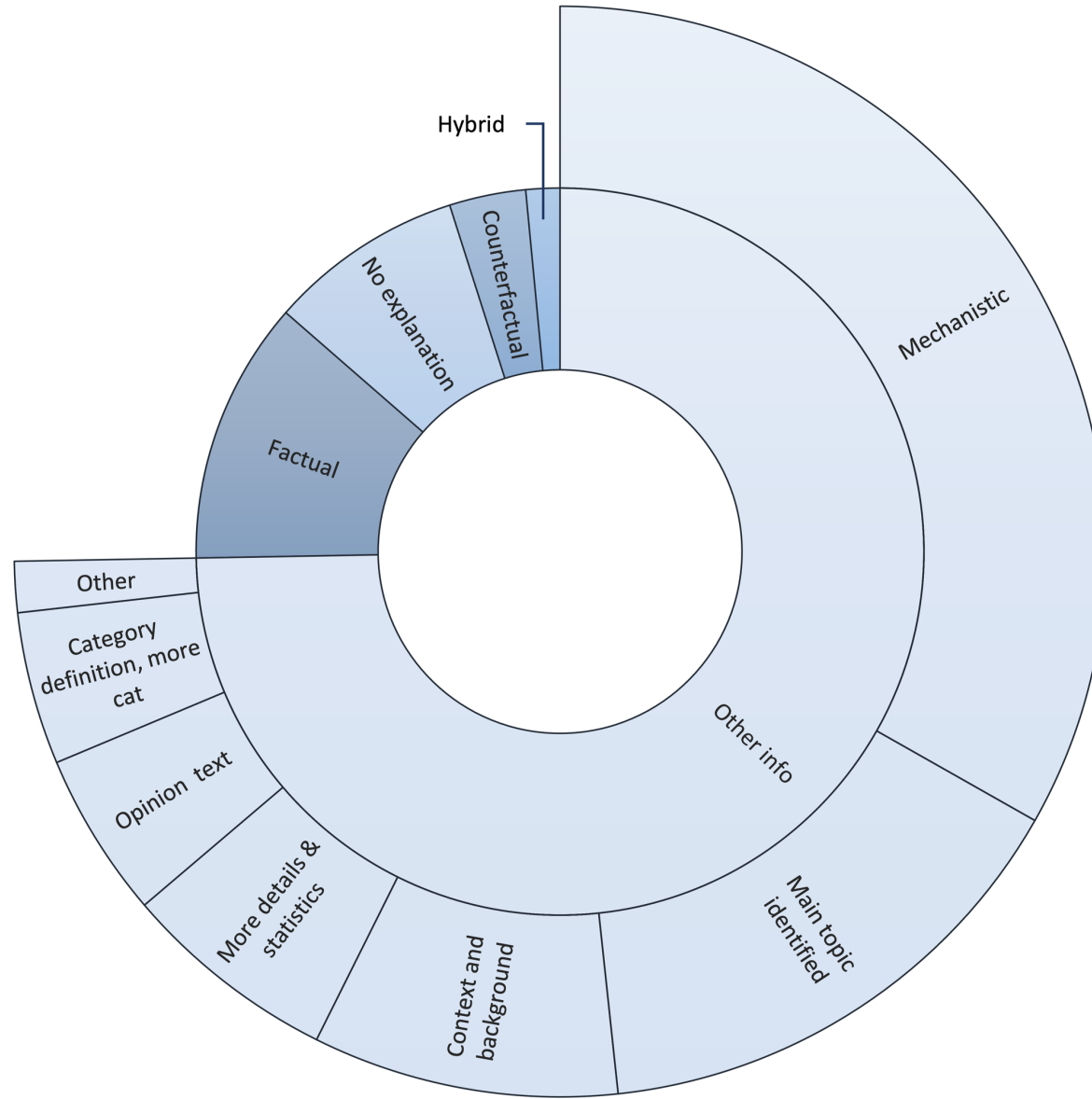
Table 2. Main categories found under class “other information” in response to “What kind of information you would like to see in the explanations from the AI system in order to understand this prediction?” when the AI system output matched user expectations. n=126.

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

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

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



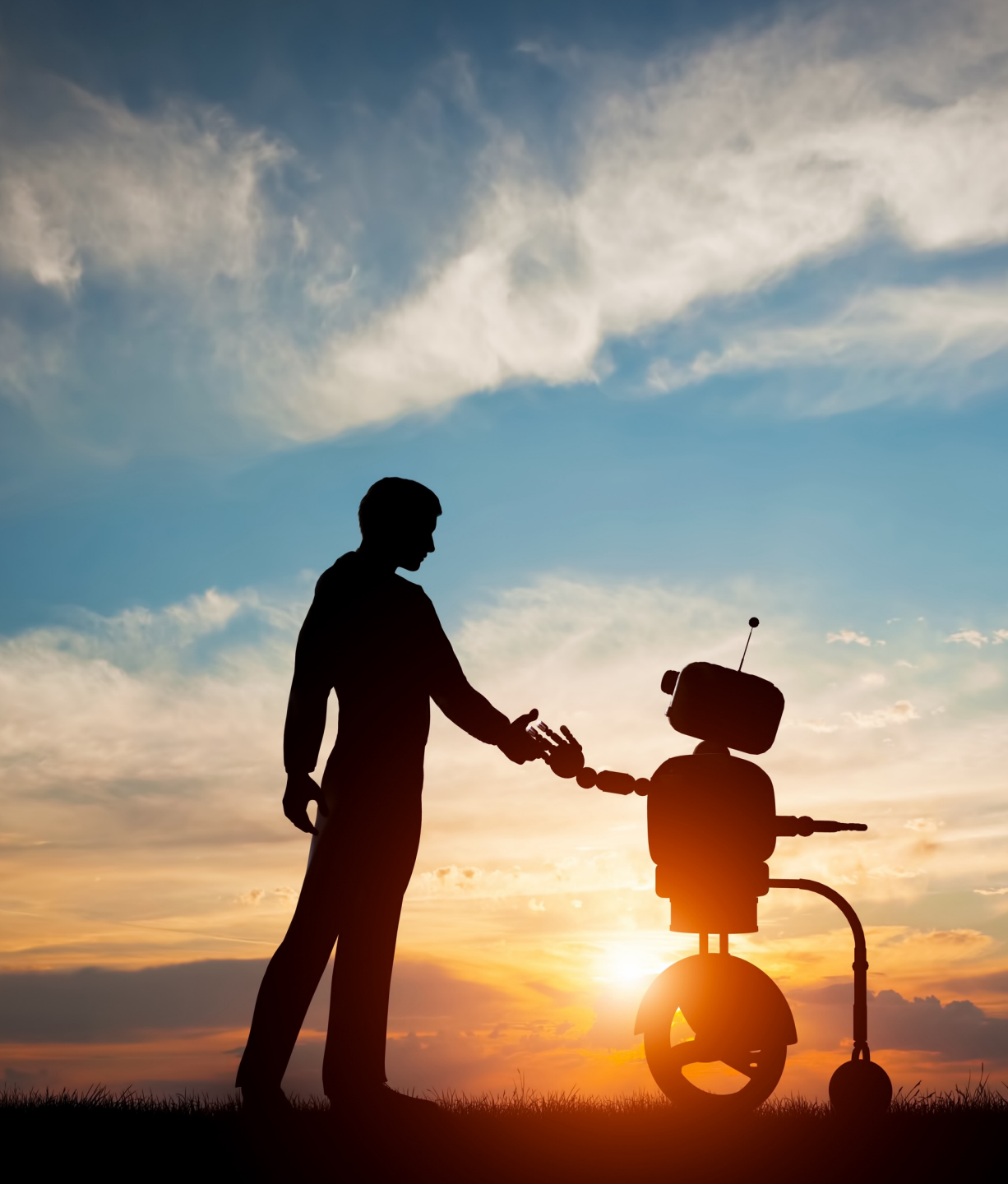
Mismatched expectations

Conclusions

- 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)

... brainstorming

- Mechanistic, how does it work? Learning at the beginning ... once the basics are covered, I could use something like counterfactuals...
- ... causality, cause and effect...
- ... we build stories...
- ... we are predictive machines....
- *I am probably biased too! Need good use cases!*



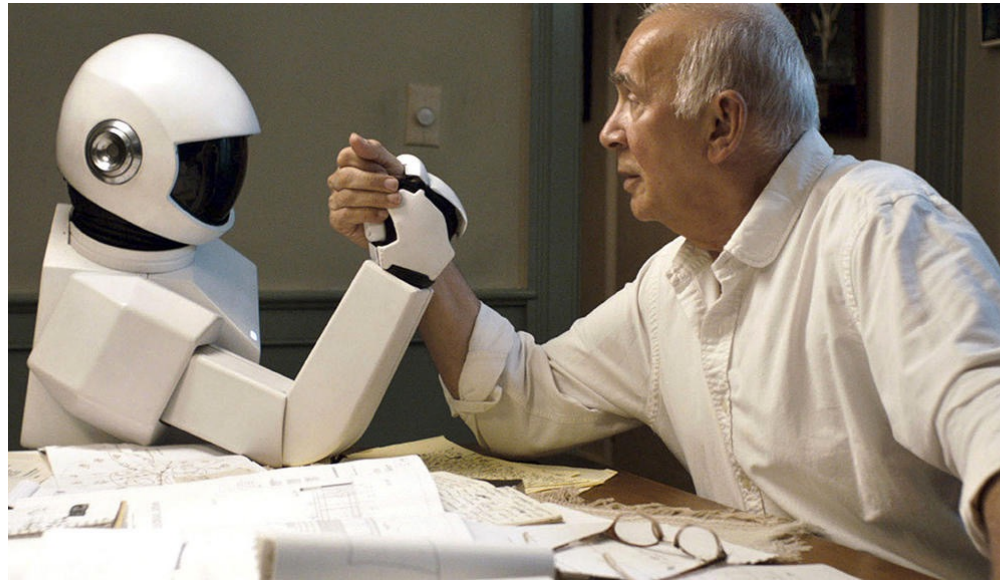
Human-machine collaboration

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

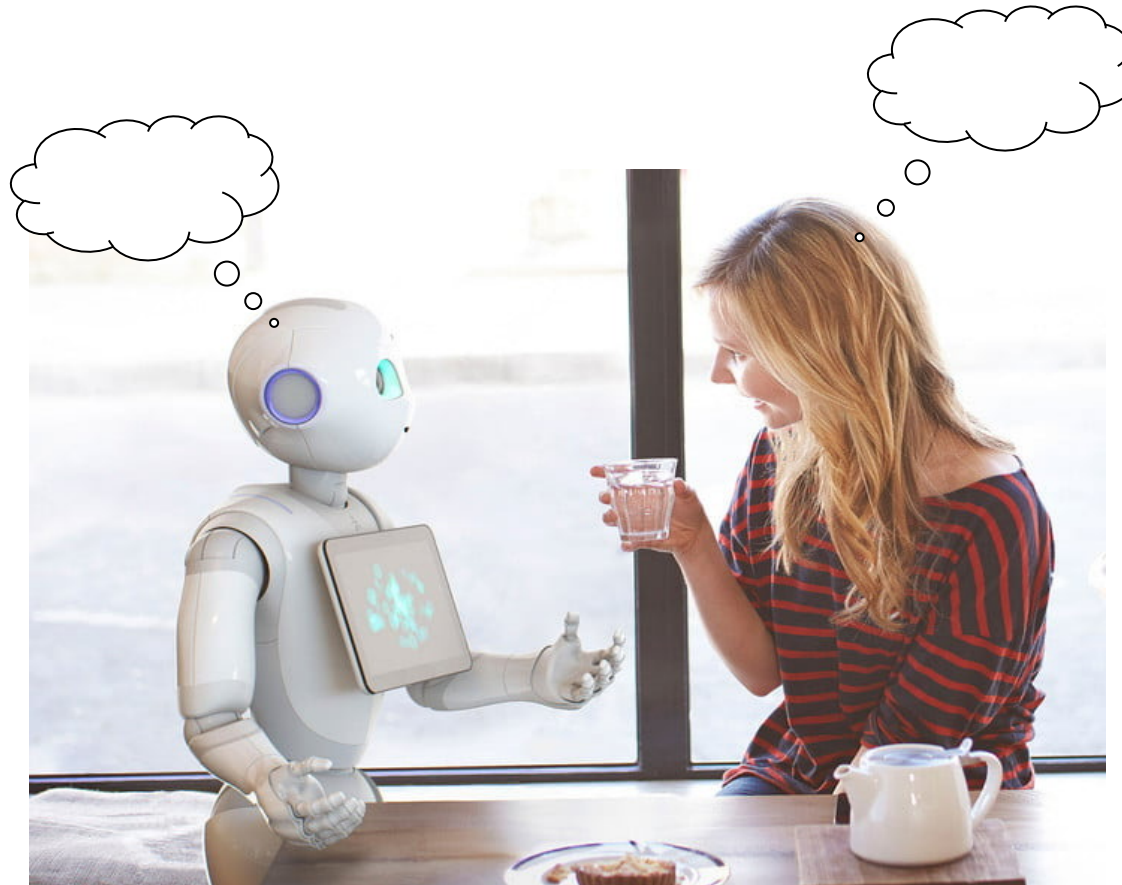
2. AI-systems need to be able to understand humans

Interaction

As humans, we interact with machines mirroring the way we interact with people



AI-systems need to be able to understand humans



Personalization and adaptation

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

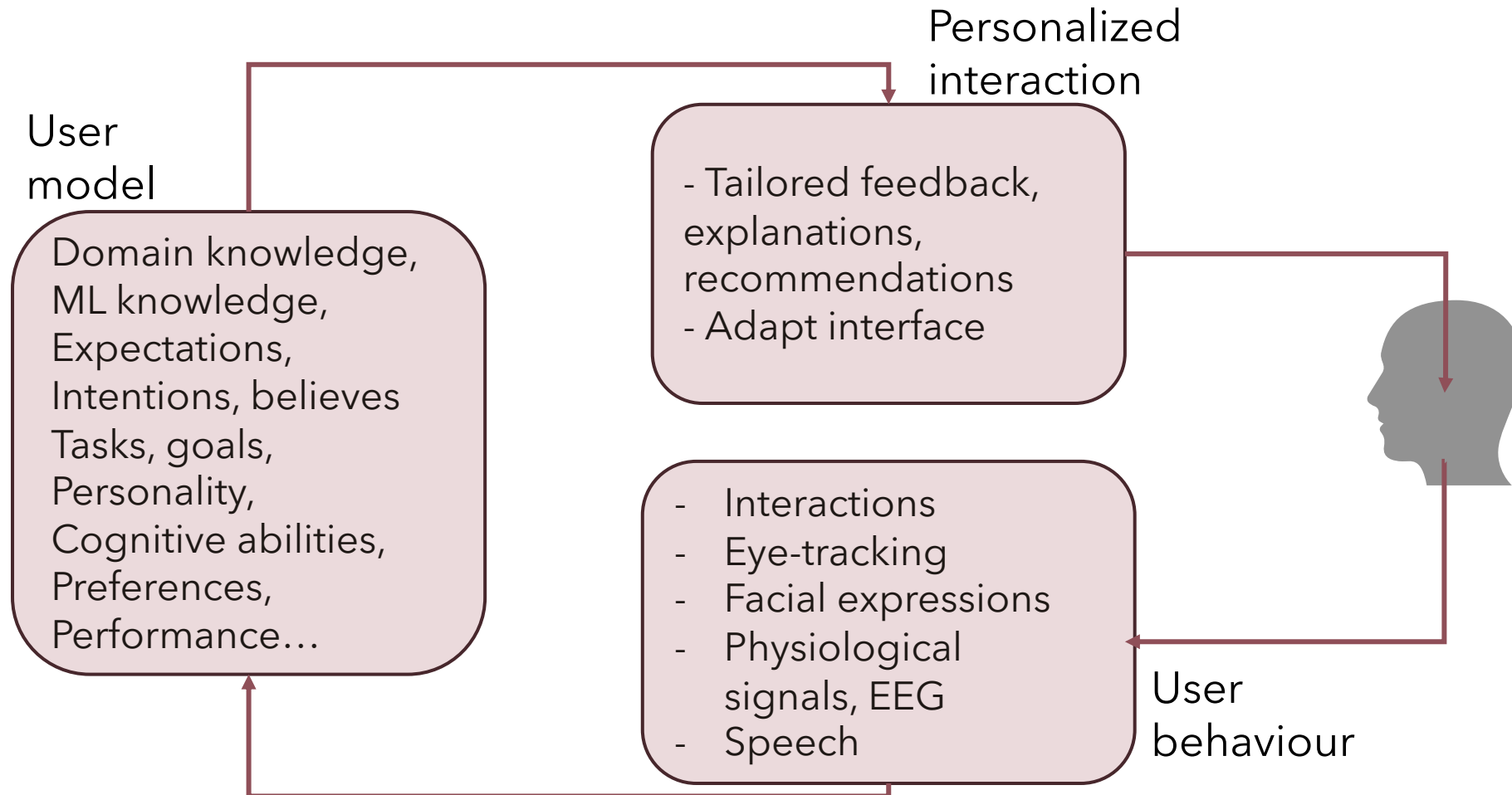


Figure inspired by
Cristina Conati's work

New studies and projects

- User modelling for predicting expectations: robots and card game with Donders, NL
- Eye-tracking to predict the foil in counterfactuals
- XPECT - VR how to build explanations from AI systems that are tailored to user expectations, mimicking human-human interactions.
- XPECT contributes to providing improved human-AI communication to support human-AI collaboration





Interests- future

- User models
- Theory of Mind
- Expectations (beliefs, intentions)
- Evaluation

Thanks!

Maria Riveiro

Visit us at Jönköping!

References

My own work:

- ❖ Riveiro, M., 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, 298.
- ❖ Riveiro, M., 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.
- Thill, S., Riveiro, M., Lagerstedt, E., Lebram, M., Hemeren, P., Habibovic, A., Klingegård, M. (2018). Driver adherence to recommendations from support systems improves if the systems explain why they are given: A simulator study Transportation Research Part F: Traffic Psychology and Behaviour, 56, 420-435.
- Thill, S., Riveiro, M. (2019). Memento hominibus: on the fundamental role of end users in real-world interactions with neuromorphic systems. Robust Artificial Intelligence for Neurorobotics, 26 – 28 August 2019, Edinburgh, Scotland.
- Gleicher, M., Riveiro, M., Von Landesberger, T., Deussen, O., Chang, R., Gillman, C. (2023). A Problem Space for Designing Visualizations IEEE Computer Graphics and Applications, 43(4), 111-120.
- Helldin, T., Falkman, G., Riveiro, M., Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. New York: Association for Computing Machinery (ACM), 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'13), 28-30 October, 2013, Eindhoven, The Netherlands.
- Pettersson, T., Riveiro, M., Löfström, T. (2023). Explainable local and global models for fine-grained multimodal product recognition. Multimodal KDD 2023, International Workshop on Multimodal Learning, in conjunction with 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023), Long Beach, CA, USA. More information

References

- Reviews in XAI
 - Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5). <https://doi.org/10.1145/3236009>
 - Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
 - Vilone, G., & Longo, L. (2020). *Explainable Artificial Intelligence: a Systematic Review*.
 - Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- LIME
 - Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " 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).

References

- Evaluation methods and metrics:
 - Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects Metrics for Explainable AI: Challenges and Prospects I. 1–50.
 - Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Survey of Evaluation Methods and Measures for Interpretable Machine Learning.
- Others
 - Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31-57.
 - Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.