Academic excellence for
business and the professions

Making Human-Centred Machine Intelligence Intelligible

Dr Simone Stumpf
Centre for Human-Computer Interaction Design
Simone.Stumpf.1@city.ac.uk
@DrSimoneStumpf

Dr Simone Stumpf

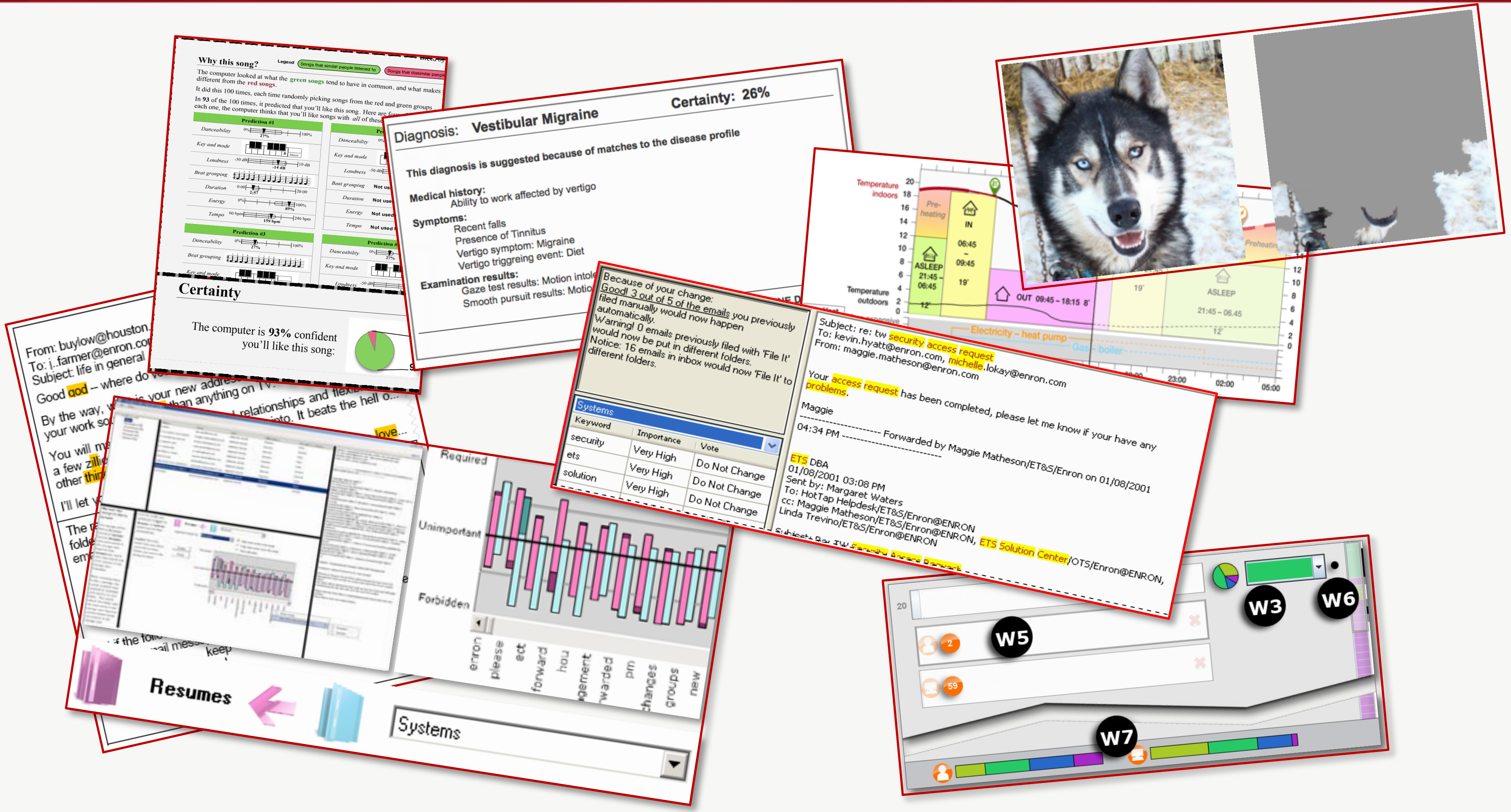
Cognitive science HCI design and evaluation methods HCI theory, concepts and models Human computer interaction Human-centered computing Interaction design Machine learning Organizing principles for web applications Peer-to-peer retrieval Spreadsheets Systems analysis and design Touch screens Web applications Web searching and information discovery Web services

The problem with machine intelligence for people

- Black boxes
- Limited user feedback
- Poor mental models

- **Intelligibility**
- **Controllability**
- **User Experience**

Explanations for intelligibility



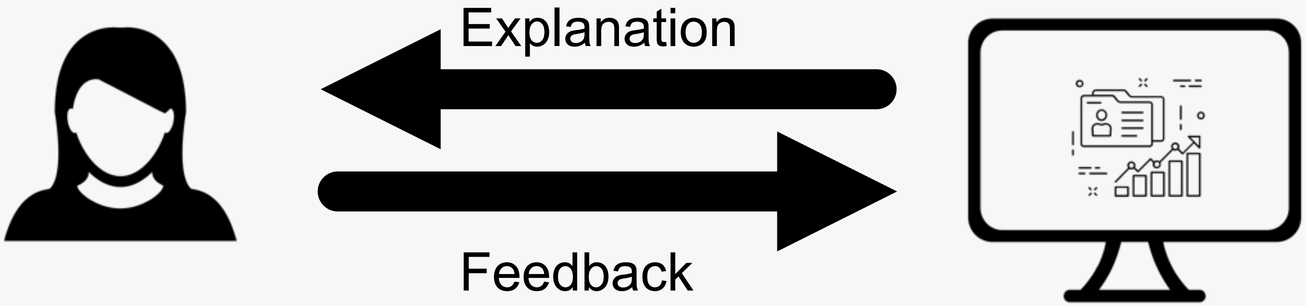
Current frameworks for creating explanations

Aspect	Interpretability	Explanatory Debugging
Main papers	Doshi-Velez and Been Kim 2017 Brian Y. Lim and Anind K. Dey 2011	Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf 2015
Context of Use	Incompleteness of AI system in optimization or evaluation	Interactive machine learning, personalization
Main Goals	Interpretability, users' understanding	Correct system "bugs"
Secondary Goals	Fairness, reliability, trust	Users' understanding, satisfaction
Explanation design – What to include	Explanations types, such as What, Certainty, Why, Why Not and Inputs	Interactive explanations including features, predictions, and model (e.g. weights, prediction confidence, class balance)
Explanation design – How to present	Communicate in "human-understandable" terms	Presented iteratively, as sound and complete as possible while not overwhelming the user

Improved **interactive machine learning** through better **intelligibility**

Explanatory debugging

[eg. Stumpf et al. IJHCS 2009, Kulesza et al. CHI 2012, Das et al. AI 2013, Kulesza et al. IUI 2015]

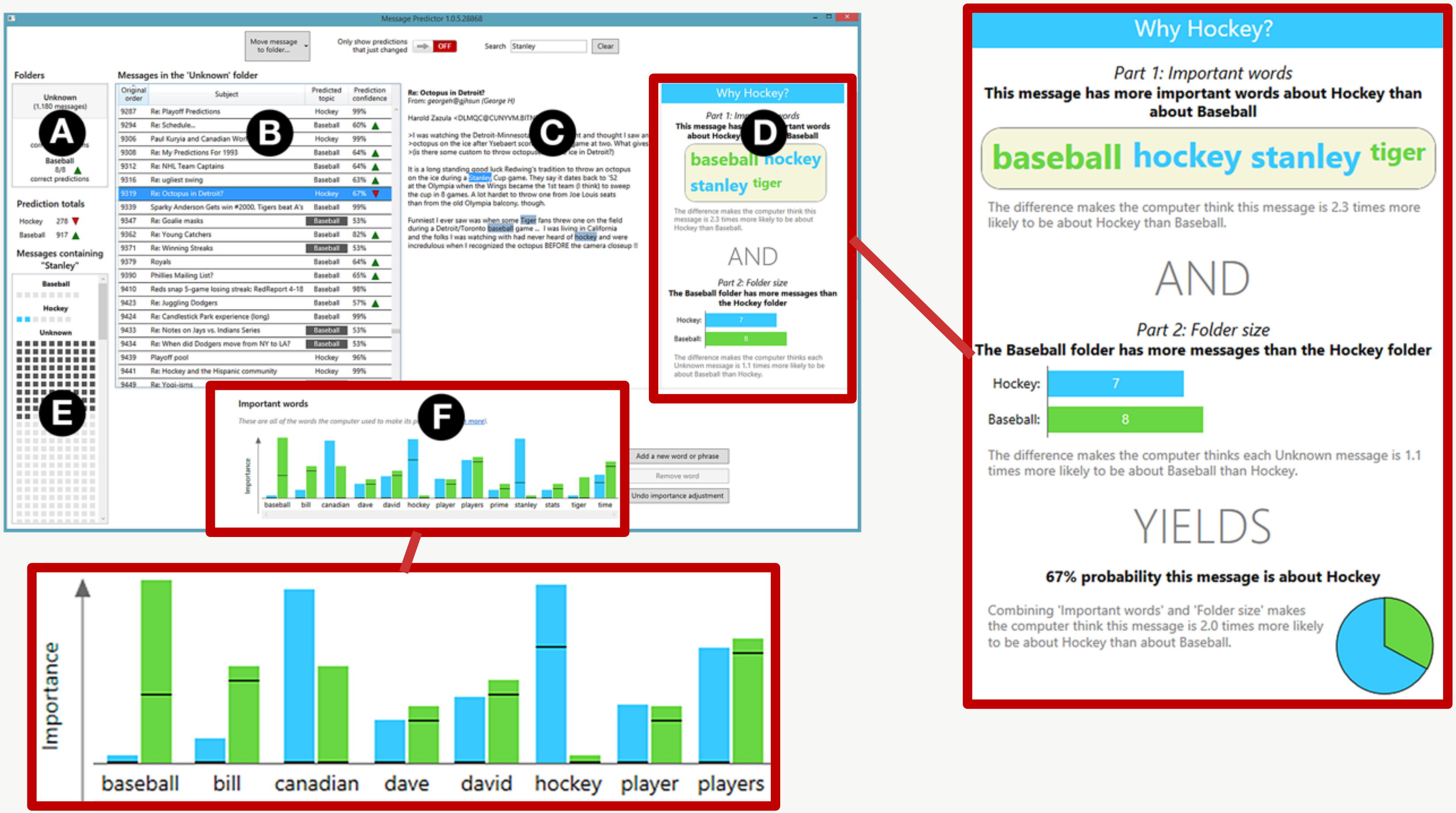


Improved mental model, satisfaction

Future improved behaviour

Intelligibility principles

Controllability principles



Message Predictor 1.0.5.28868
Move message to folder...
Only show predictions that just changed
OFF
Search
Stanley
Clear
Folders
Messages in the 'Unknown' folder
Unknown
(1,180 messages)
Baseball
8/8
correct predictions
Prediction totals
Hockey 278
Baseball 917
Messages containing "Stanley"
Baseball
Hockey
Unknown
Original order
Subject
Predicted topic
Prediction confidence
Re: Octopus in Detroit?
A
B
C
D
E
F
Important words
Add a new word or phrase
Remove word
Undo importance adjustment
Why Hockey?
Part 1: Important words
This message has more important words about Hockey than about Baseball
baseball hockey stanley tiger
The difference makes the computer think this message is 2.3 times more likely to be about Hockey than Baseball.
AND
Part 2: Folder size
The Baseball folder has more messages than the Hockey folder
Hockey: 7
Baseball: 8
The difference makes the computer thinks each Unknown message is 1.1 times more likely to be about Baseball than Hockey.
YIELDS
67% probability this message is about Hockey
Combining 'Important words' and 'Folder size' makes the computer think this message is 2.0 times more likely to be about Hockey than about Baseball.
Importance
baseball
bill
canadian
dave
david
hockey
player
players

What we know so far…

- Integrating user feedback:
  - No improvements in accuracy for all users
  - More accurate system accuracy (85% vs 77%)
  - With less effort (47 messages vs 182 messages)

- Explanations:
  - Rule-based ≥ Keyword-based (but beware of negative weights!)
  - Very individual preferences
  - Better understanding (15.8 MM score vs 10.4) → better system
  - No difference in workload

Transparency Design Process

[Eiband et al. IUI 2018]

WHAT to explain?			HOW to explain?	
Expert Mental Model	User Mental Model	Target Mental Model	Iterative Prototyping	Evaluation
What happens to the best of our knowledge? What *can* be explained? What does an *expert mental model* of the system look like?	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*?	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 algorithm?	How can the *target mental model* be reached through UI design? How and where can transparency be integrated into the UI of the system?	How has the *user mental model* developed? Has the *target mental model* been reached?

Smart heating

controls heat

makes heat

Persuasive Engagement

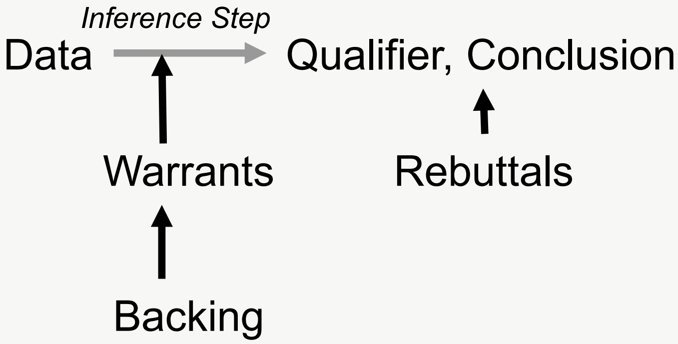
[eg. Stumpf IUI ExSS2019]

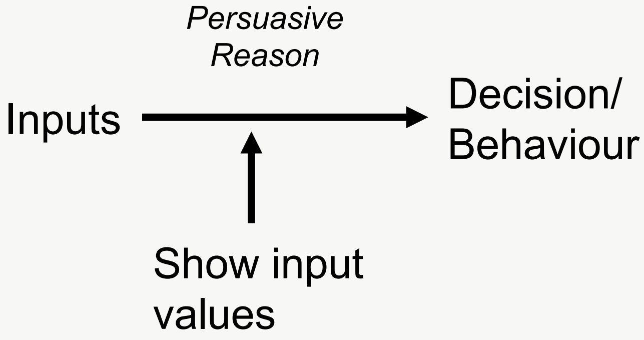
Aspect	Persuasive Engagement
Context of Use	Everyday low-risk systems, constrained engagement situations
Main Goals	User trust and satisfaction
Secondary Goals	Understanding
Explanation design – What to include	Inputs, Inference step, Decision/Behavior
Explanation design – How to present	Concise, lightweight, drill-down on demand

How to construct explanations within PE

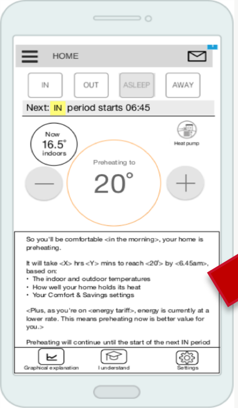
Argument structure	Persuasive engagement
Data/Facts	Inputs
Inference step	Persuasive reason for making the decision
Qualified Conclusion	Decision/Behavior
On ‘Why’: Show Warrants, Backing, Rebuttals	On request: show input values
Natural language	Present in easily understandable form

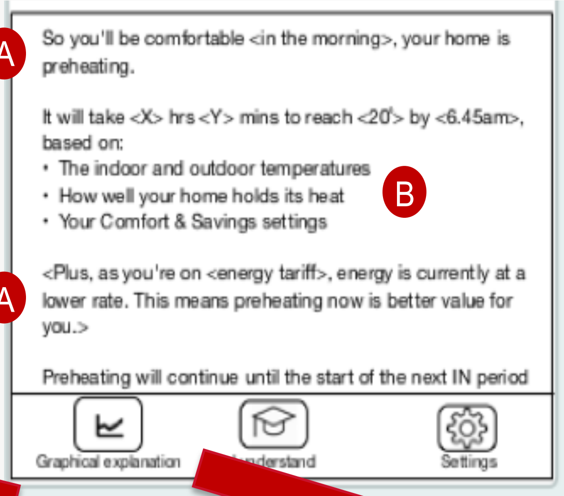
How to construct an explanation



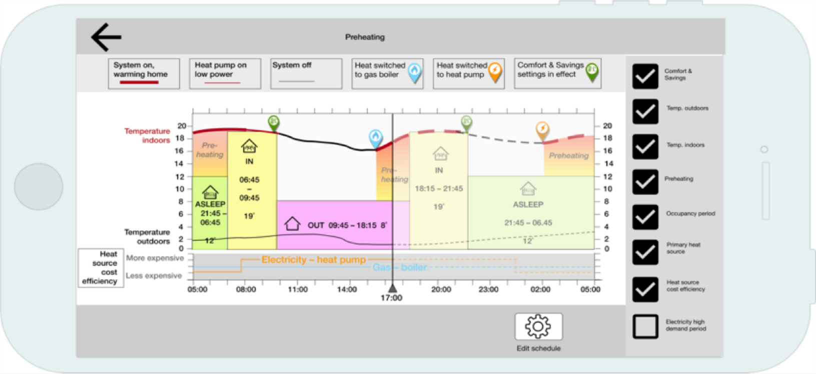


An example smart heating explanation





- 7 unexpected decision points
  - Provide persuasive reason for making decision (A)
  - Show inputs (B)
  - Values on request
  - Text first, graphs later



Conclusions … and yet more questions

- Intelligibility
  - How to make a system intelligible in different contexts and for different purposes?
  - How to extend and validate Persuasive Engagement framework?
  - How does a system become intelligible as the user interacts?
- Controllability
  - How can we empower the user to take control back over their data and over what the system does?
- User Experience
  - Complex relationship between understanding, trust, satisfaction, system performance, explanations, …

References

- S. Stumpf, “Horses For Courses: Making The Case For Persuasive Engagement In Smart Systems,” in *IUI Workshops’19, March 20, 2019, Los Angeles, USA*, 2019.
- T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, “Principles of Explanatory Debugging to Personalize Interactive Machine Learning,” in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, New York, NY, USA, 2015, pp. 126–137.
- T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, “Tell me more?: the effects of mental model soundness on personalizing an intelligent agent,” in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, New York, NY, USA, 2012, pp. 1–10.
- S. Stumpf *et al.*, “Interacting meaningfully with machine learning systems: Three experiments,” *Int. J. Hum.-Comput. Stud.*, vol. 67, no. 8, pp. 639–662, 2009.
- Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- Brian Y. Lim and Anind K. Dey. 2011. Investigating Intelligibility for Uncertain Context-aware Applications. In Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp ’11), 415–424.
- Shubhomoy Das, Travis Moore, Weng-Keen Wong, Simone Stumpf, Ian Oberst, Kevin McIntosh, and Margaret Burnett. 2013. End-user feature labeling: Supervised and semi-supervised approaches based on locally-weighted logistic regression. Artificial Intelligence 204: 56–74.
- Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In 23rd International Conference on Intelligent User Interfaces (IUI ’18), 211–223. https://doi.org/10.1145/3172944.3172961