



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](mailto:Simone.Stumpf.1@city.ac.uk)

[@DrSimoneStumpf](https://twitter.com/DrSimoneStumpf)

[www.city.ac.uk](http://www.city.ac.uk)



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

### Why this song?

The computer looked at what the green songs tend to have in common, and what makes them different from the red songs. It did this 100 times, each time randomly picking songs from the red and green groups. In 93 of the 100 times, it predicted that you'll like this song. Here are some examples. In each one, the computer thinks that you'll like songs with all of these:

Preference 73	Preference 43
Discrepancy	Discrepancy
Key and mode	Key and mode
Leadlines	Leadlines
Beat grouping	Beat grouping
Duration	Duration
Energy	Energy
Tempo	Tempo

**Certainty**

The computer is 93% confident you'll like this song:



### Diagnosis: Vestibular Migraine

**Certainty: 26%**

This diagnosis is suggested because of matches to the disease profile

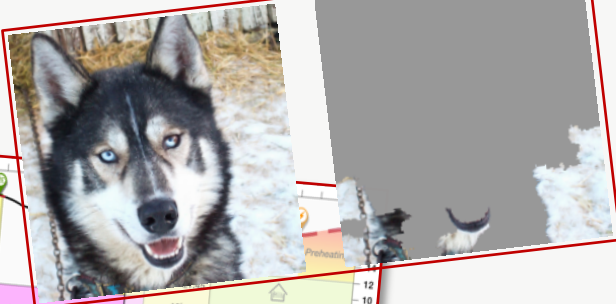
**Medical history:**  
Ability to work affected by vertigo

**Symptoms:**  
Recent falls  
Presence of Tinnitus  
Vertigo symptom: Migraine  
Vertigo triggering event: Diet

**Examination results:**  
Gaze test results: Motion into  
Smooth pursuit results: Motion into

Because of your change, I moved 3 out of 5 of the emails you previously filed manually would now happen automatically. Warning! 0 emails previously filed with 'File It!' would now be put in different folders. Notice: 16 emails in inbox would now 'File It' to different folders.

Systems	Importance	Vote
security	Very High	Do Not Change
ets	Very High	Do Not Change
solution	Very High	Do Not Change



From: buylow@houston.com  
To: j.farmer@enron.com  
Subject: life in general

Good god -- where do you find your new address? It's more than anything on TV. By the way, you're doing better than relationships and flexing your work skills. You will miss a few zillion other things. I'll let you know when I'll let you know. The folders are empty.

Required

Unimportant

Forbidden

Resumes

Systems

Subject: re: tw security access request  
To: kevin.fyatt@enron.com, michelle.lokay@enron.com  
From: maggie.matheson@enron.com

Your access request has been completed, please let me know if you have any problems.

Maggie

04:34 PM Forwarded by Maggie Matheson/ET&S/Enron on 01/08/2001

ETS DBA  
01/08/2001 03:08 PM  
Sent by: Margaret Waters  
To: HotTap Helpdesk/ET&S/Enron@ENRON  
cc: Maggie Matheson/ET&S/Enron@ENRON, Linda Trevino/ET&S/Enron@ENRON

W3

W5

W6

W7

# Current frameworks for creating explanations

Aspect	Interpretability	Explanatory Debugging
<b>Main papers</b>	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
<b>Context of Use</b>	Incompleteness of AI system in optimization or evaluation	Interactive machine learning, personalization
<b>Main Goals</b>	Interpretability, users' understanding	Correct system "bugs"
<b>Secondary Goals</b>	Fairness, reliability, trust	Users' understanding, satisfaction
<b>Explanation design – What to include</b>	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)
<b>Explanation design – How to present</b>	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**

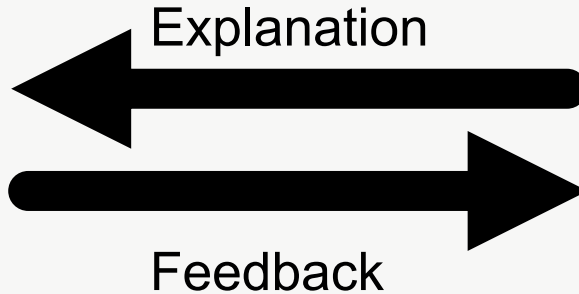


# Explanatory debugging

[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 1.0.5.210584

Move message to folder... Only show predictions that just changed  OFF Search Stanley Clear

**Folders**

- Unknown (1,180 messages)
- Baseball (18 correct predictions)
- Hockey 278
- Baseball 917

**Messages in the 'Unknown' folder**

Original order	Subject	Predicted topic	Prediction confidence
9287	Re: Playoff Predictions	Hockey	99%
9284	Re: Schedule...	Baseball	60%
9206	Re: Paul Kuryla and Canadian Wings	Hockey	99%
9208	Re: My Predictions For 1993	Baseball	64%
9212	Re: NHL Team Captains	Baseball	64%
9216	Re: ugliest swing	Baseball	63%
9219	Re: Octopus in Detroit?	Hockey	67%
9239	Sparky Anderson Gets win #2000. Tigers beat A's	Baseball	99%
9247	Re: Goalie masks	Baseball	82%
9262	Re: Young Catchers	Baseball	82%
9271	Re: Winning Streaks	Baseball	53%
9279	Royals	Baseball	64%
9290	Phillies Mailing List?	Baseball	68%
9410	Reds snap 5-game losing streak: RedReport 4-18	Baseball	98%
9423	Re: Juggling Dodgers	Baseball	57%
9424	Re: Candlestick Park experience (song)	Baseball	99%
9433	Re: Notes on Jays vs. Indians Series	Baseball	53%
9434	Re: When did Dodgers move from NY to LA?	Baseball	53%
9439	Playoff pool	Hockey	96%
9441	Re: Hockey and the Hispanic community	Hockey	99%
9449	Re: Yoo-hims		

**Re: Octopus in Detroit?**  
From: georgh@ghum (George H)  
Harold Zula - <DLMQC@CUNYVM.BIT>  
>I was watching the Detroit-Minnesota game and thought I saw an octopus on the ice after Ysebaert scored the game at two. What gives  
>is there some custom to throw octopus on the ice in Detroit?  
It is a long standing good luck Redwings' tradition to throw an octopus on the ice during a playoff Cup game. They say it dates back to 52 at the Olympia when the Wings became the 1st team (I think) to sweep the cup in 8 games. A lot harder to throw one from Joe Louis seats than from the old Olympia balcony, though.  
Funniest I ever saw was when some local fans threw one on the field during a Detroit/Toronto baseball game... I was living in California and the folks I was watching with had never heard of hockey and were incredulous when I recognized the octopus BEFORE the camera closeup!!

**Important words**  
These are all of the words the computer used to make its prediction.

**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 think 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.

## 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

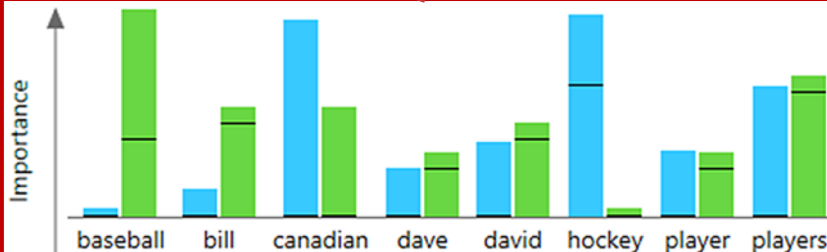
Baseball: 8

The difference makes the computer think 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.

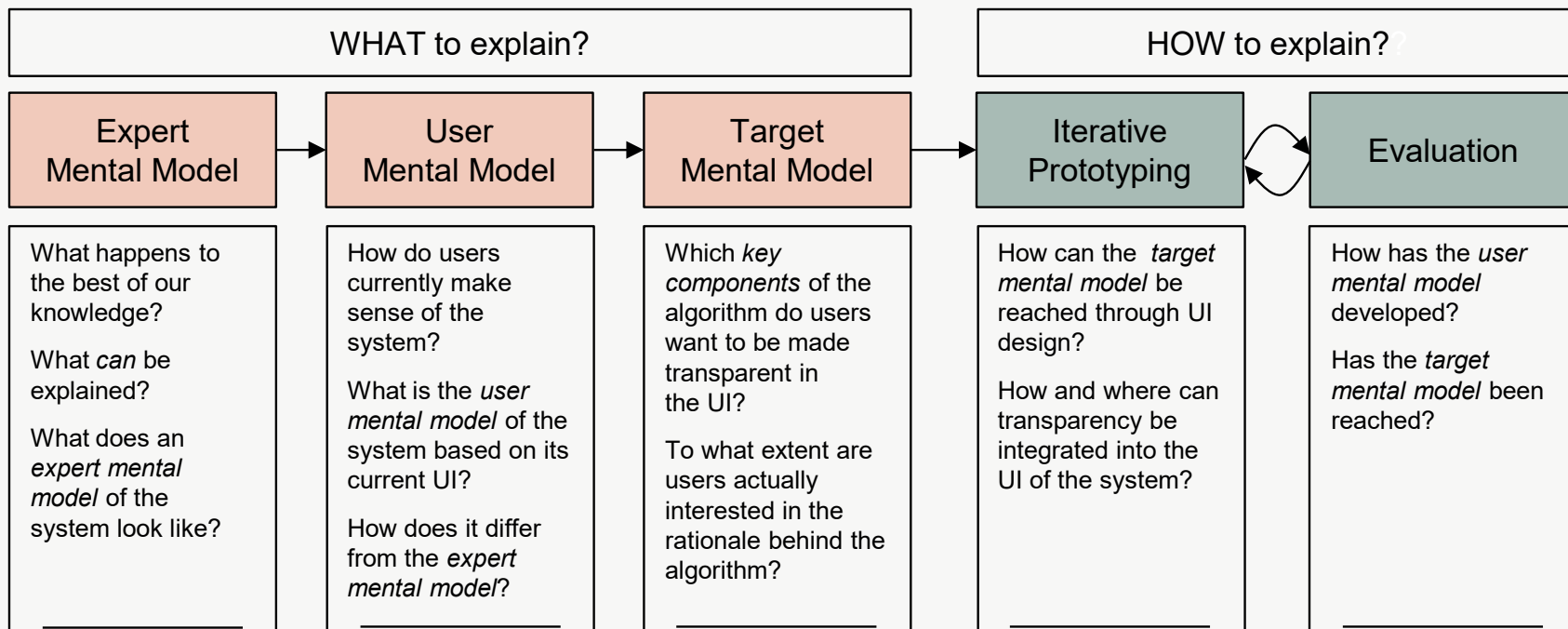


# What we know so far...

- 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  $\geq$  Keyword-based (but beware of negative weights!)
  - Very individual preferences
  - Better understanding (15.8 MM score vs 10.4)  $\rightarrow$  better system
  - No difference in workload

# Transparency Design Process

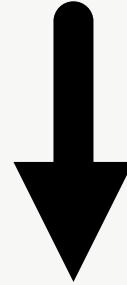
[Eiband et al. IUI 2018]



# Smart heating



**controls heat**

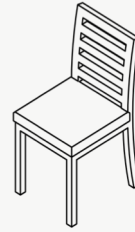


**makes heat**

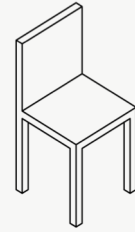


# Persuasive Engagement

[eg. Stumpf IUI ExSS2019]



Interpretability



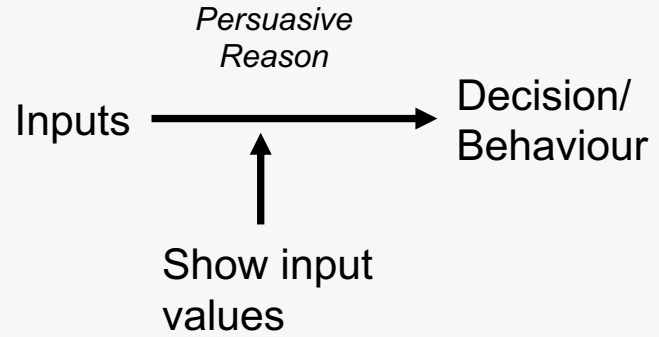
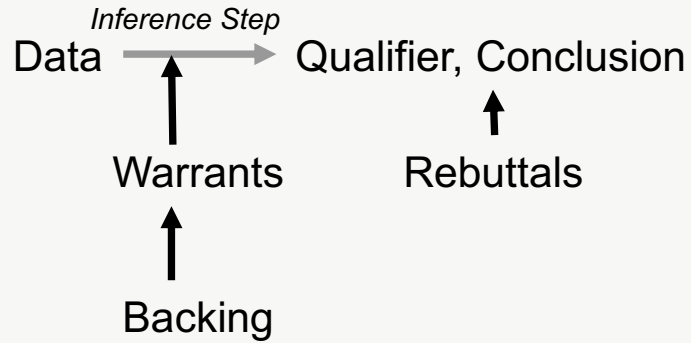
Explanatory Debugging

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

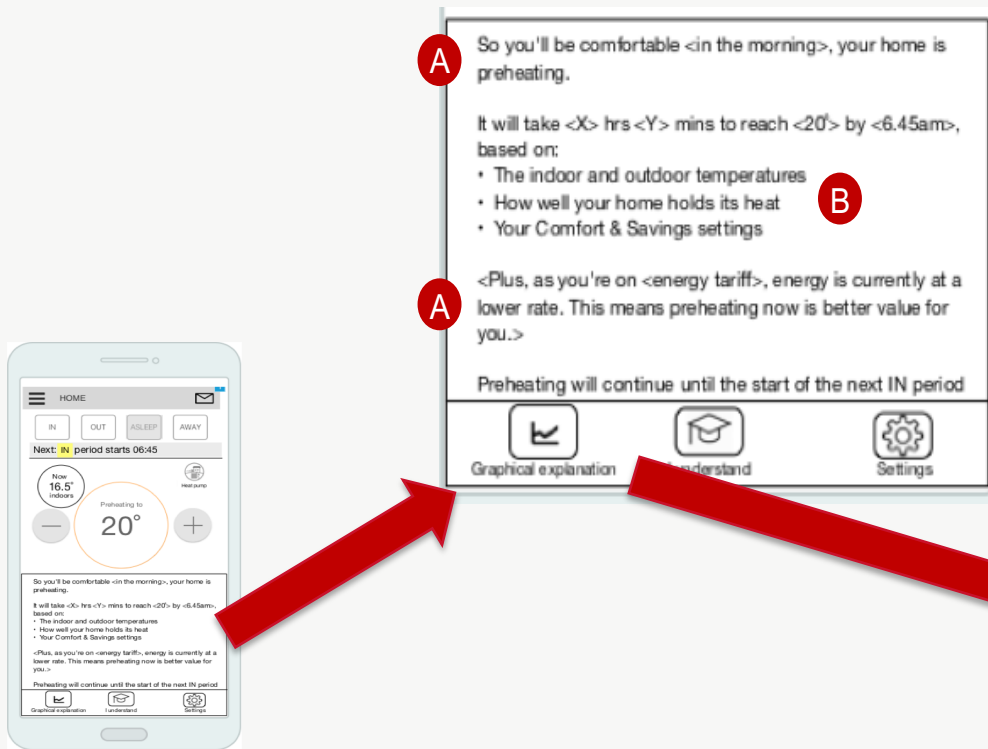
# How to construct explanations within PE

<b>Argument structure</b>	<b>Persuasive engagement</b>
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

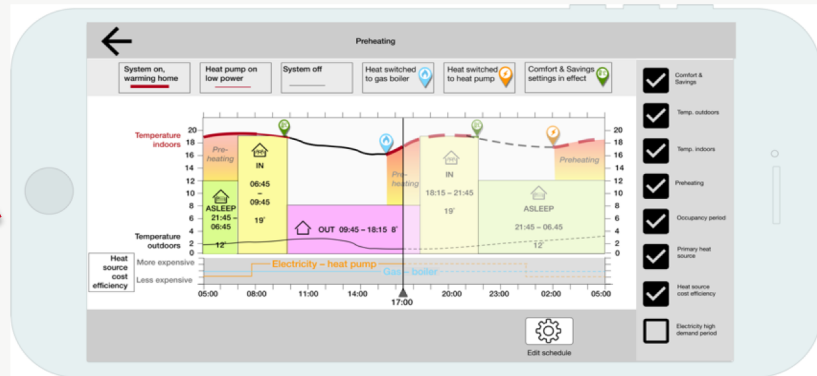
# How to construct an explanation



# 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>