# Explainability & Explainable AI (XAI)

Dr Simone Stumpf
Reader in Responsible and Interactive AI
University of Glasgow

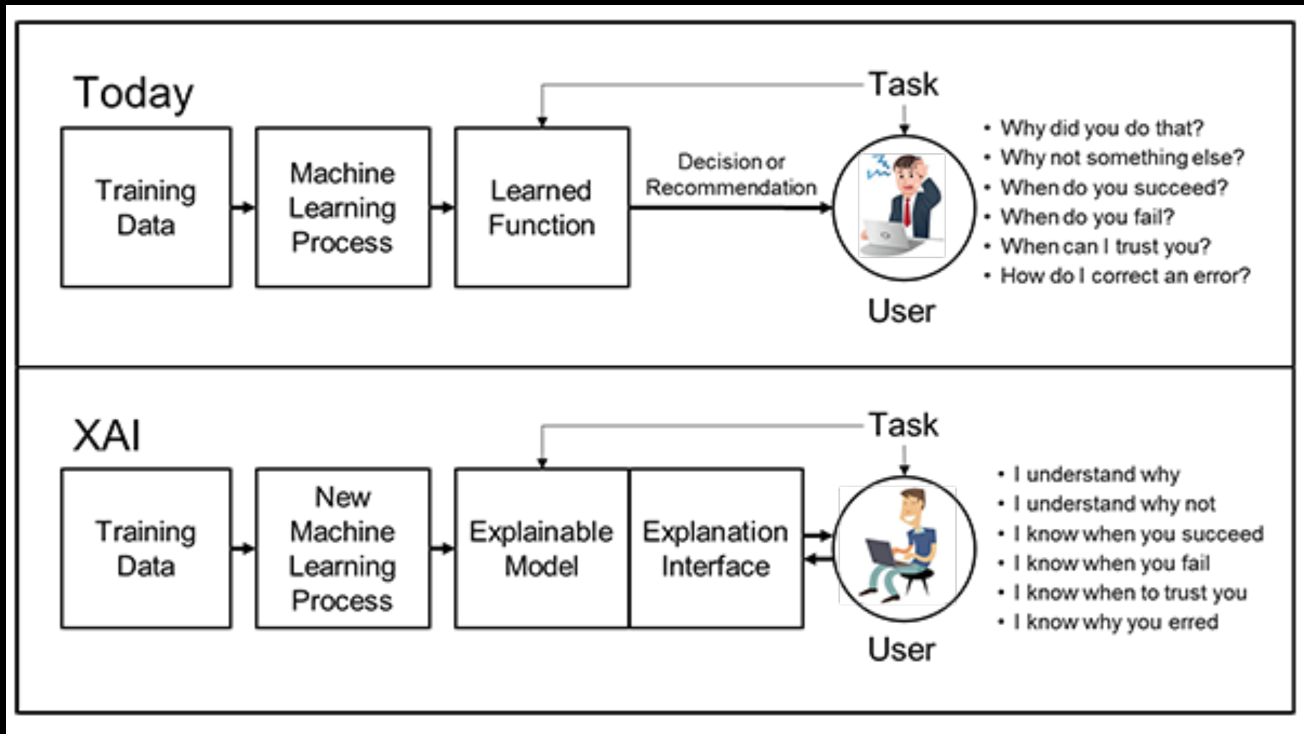
Simone.Stumpf@glasgow.ac.uk
@DrSimoneStumpf

# XAI's roots

- (Way back in humanities & social sciences)
- 1970s/1980s: Expert system explanations
- 1990s/2000s: Growth of machine learning
- 2016: DARPA XAI programme

Mycin: Is Meningitis a hospital acquired illness?
Doctor: why?
Mycin: [2.0] this will aid in determining the orgamisms that might be causing the infection
It has already established
[2.1] the infection that requires therapy is meningitis
[2.2] organisms were not seen on the stain of this pending csf culture
[2.3] the type of infection is bacterial
[2.4] neurosurgery is not one of the relevant items of patient 100
[2.5] the age of patient 100 is greater than 25
therefore if [2.6] the infection was acquired while the patient was in hospital then there is evidence that the organisms that might be causing the infection are E coli(0.75) staphylococcus-coag-pos(0.3) pseudomonas-aeruginosa(0.3) klebsiella-pneumoniae(0.5)
Doctor: How 2.3
Mycin: The following rules concluded about the likelihood that the type of the infection is bacterial:
[2.7] RULE148 (0.19)
[2.8] RULE500 (0.51)
[2.9] RULE501 (0.95)
[3.0] RULE502 (0.97)
[3.1] RULE526 (0.98)
[3.2] RULE504 (0.97)
[3.3] RULE524 (0.98)

# XAI vision


Today
Training Data
Machine Learning Process
Learned Function
Decision or Recommendation
Task
User
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
XAI
Training Data
New Machine Learning Process
Explainable Model
Explanation Interface
Task
User
• I understand why
• I understand why not
• I know when you succeed
• I know when you fail
• I know when to trust you
• I know why you erred

# Lots of work to make ML explainable

[Molnar 2022]

- Simplest: I give you the source code of the model

- Next simplest: I give you a representation of the model

    - Exposing the model (global explanation)
    - Exposing (combination of) features that contribute to a decision (local explanation)

[Ribeiro et al. KDD 2016]

# A quick aside on explanations

- Is it an explanation or a justification?
- Explanation content versus explanation presentation
- Some models are naturally interpretable. Discuss.

# Explainability versus Intelligibility

- *Explainability* = ability of an AI system to explain itself
- *Intelligibility* = ability of a user to build an appropriate mental model that guides interaction with the AI system
    - Understanding of how the system works
    - Being able to use the system successfully
    - Being able to 'trouble-shoot' system and fix 'mistakes'

# Mental Models

- A mental model is kind of internal representation in someone's thought process for how something works in the real world
- Based on meaning, understanding and experience
- Users build mental models to guide how they interact, behave or fix things when they go wrong

[Norman 1983, Johnson-Laird 1983]

# Intelligibility types

- What did the system do?
- Why did the system do W?
- Why did the system not do X?
- What would the system do if Y happens?
- How can I get the system to do Z, given the current context?

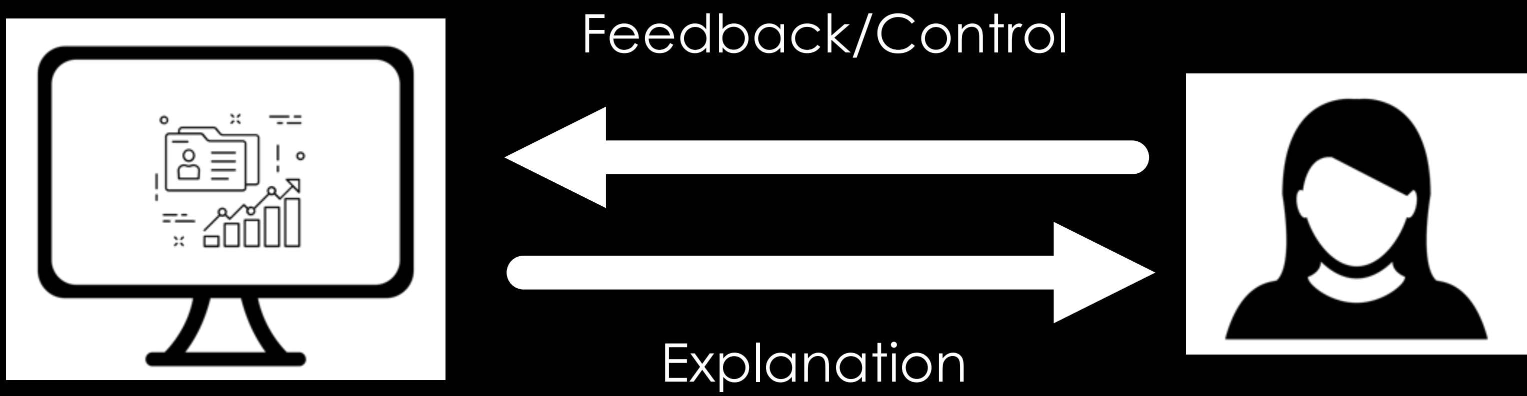
[Lim and Dey CHI 2009]

- What can you do?
- What am I doing and what have I done?
- Who is doing what, and what have they done?
- What will happen when I do this?
- Stop that!

[Bellotti and Edwards HCI 2001]

Btw, AI ≠ automation

# Explanatory debugging for interactive machine learning



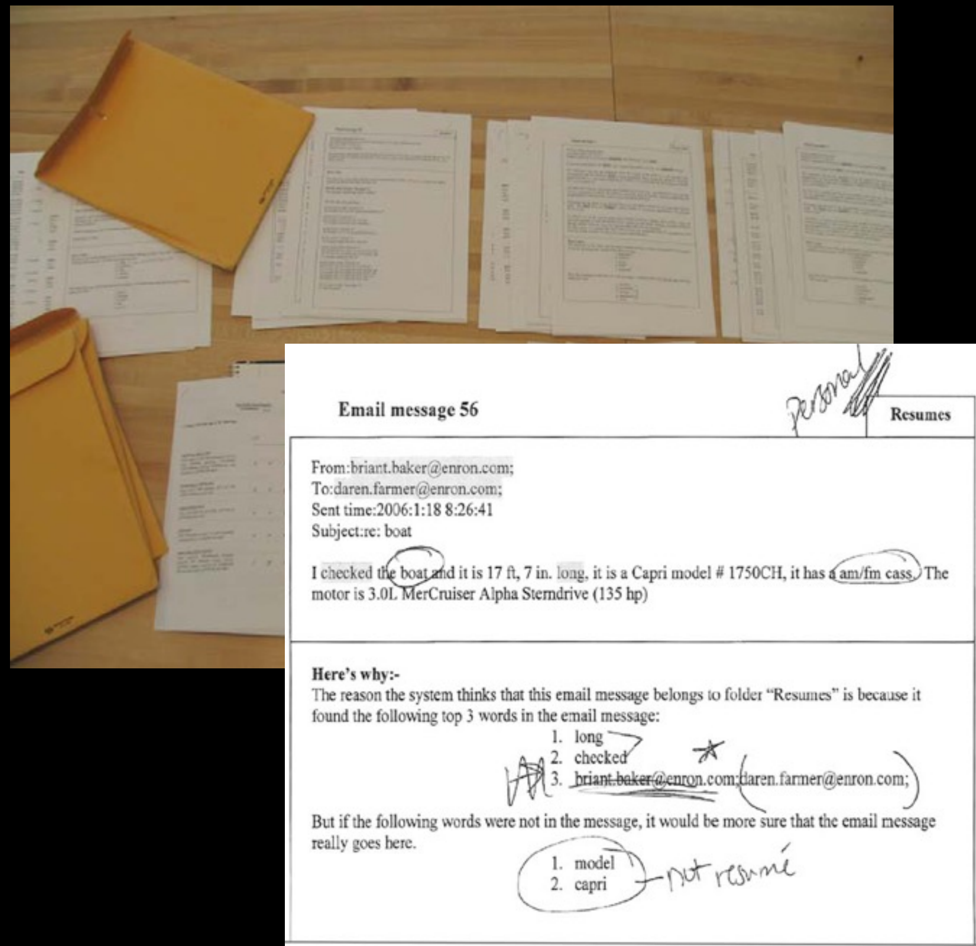

Future improved behaviour

Improved mental model, satisfaction

[Stumpf et al. IJHCS 2009, Kulesza et al. TiiS11, Kulesza et al. CHI 2012, Das et al. AI 2013, Kulesza et al. IUI 2015]

# Explanation styles and feedback

- Enron email dataset folders (farmer-d): Personal, Resume, Bankrupt, Enron News (122 messages)
- Lo-fi prototypes with 3 explanation styles of 3 different algorithms
- 13 participants
- Think-aloud




[Stumpf et al. IJHCS 2009]

# Explanation styles

## Keyword

Personal

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

Good god -- where do you find time for all of that? You should w...

By the way, what is your new address? I may want to come by ... your work sounds better than anything on TV.

You will make a good trader. Good relationships and flexible pri... a few zillion other intangibles you will run into. It beats the hell o... other things.

I'll let you be for now, but do keep those stories coming we love...

The reason the system thinks that this email message belongs to folder "Personal" is because it found the following top 5 words in the email message:

1. ill
2. love
3. better
4. things
5. god

But if the following words were not in the message, it would be more sure the email message really goes here.

1. keep
2. find
3. trader
4. book
5. general

## Rule

Resume

From: toni.graham@enron.com
To: daren.farmer@enron.com
Subject: re: job posting

Daren, is this position budgeted and who does it report to?
Thanks,
Toni Graham

The reason the system thinks that this email message belongs to folder "Resume" is because the highest priority rule that fits this email message was:

- **Put the email in folder "Resume" if:**
  **It's from toni.graham@enron.com.**

The other rules in the system are:

...

- Put the email in folder "Personal" if:
  The message does not contain the word "Enron" and
  The message does not contain the word "process" and
  The message does not contain the word "term" and
  The message does not contain the word "link".
- Put the email in folder "Enron News" if:
  No other rule applies.

## Similarity

Resume

Message #2
From: 40enron@enron.com
To: All ENW employees
Subject:enron net works t&e policy
From: Greg Piper and Mark Pickering

Please print and become familiar with the updated ENW T&E P... business-first travel, with supervisor approval, for international fli... Mexico). Supervisors will be responsible for making the decision...

If you have any questions about the policy or an expense not co... Costello.

Wow! The message is really similar to the message #3 in "**Resume**" because #2 and #3 have important words in common.

Message #3
From: toni.graham@enron.com
To: lisa.csikos@enron.com, rita.wynne@enron.com, daren.farmer@enron.com
CC: renda.herod@enron.com
Subject: confirming requisitions

Confirming the open requisitions for your group. If your records indicate otherwise, please let me know.

Lisa Csikos 104355, 104001
Rita Wynne 104354
Daren Farmer 104210
Mike Eiben 104323
Pat Clynes 104285

The posting dates have all been **updated** to reflect a current posting date

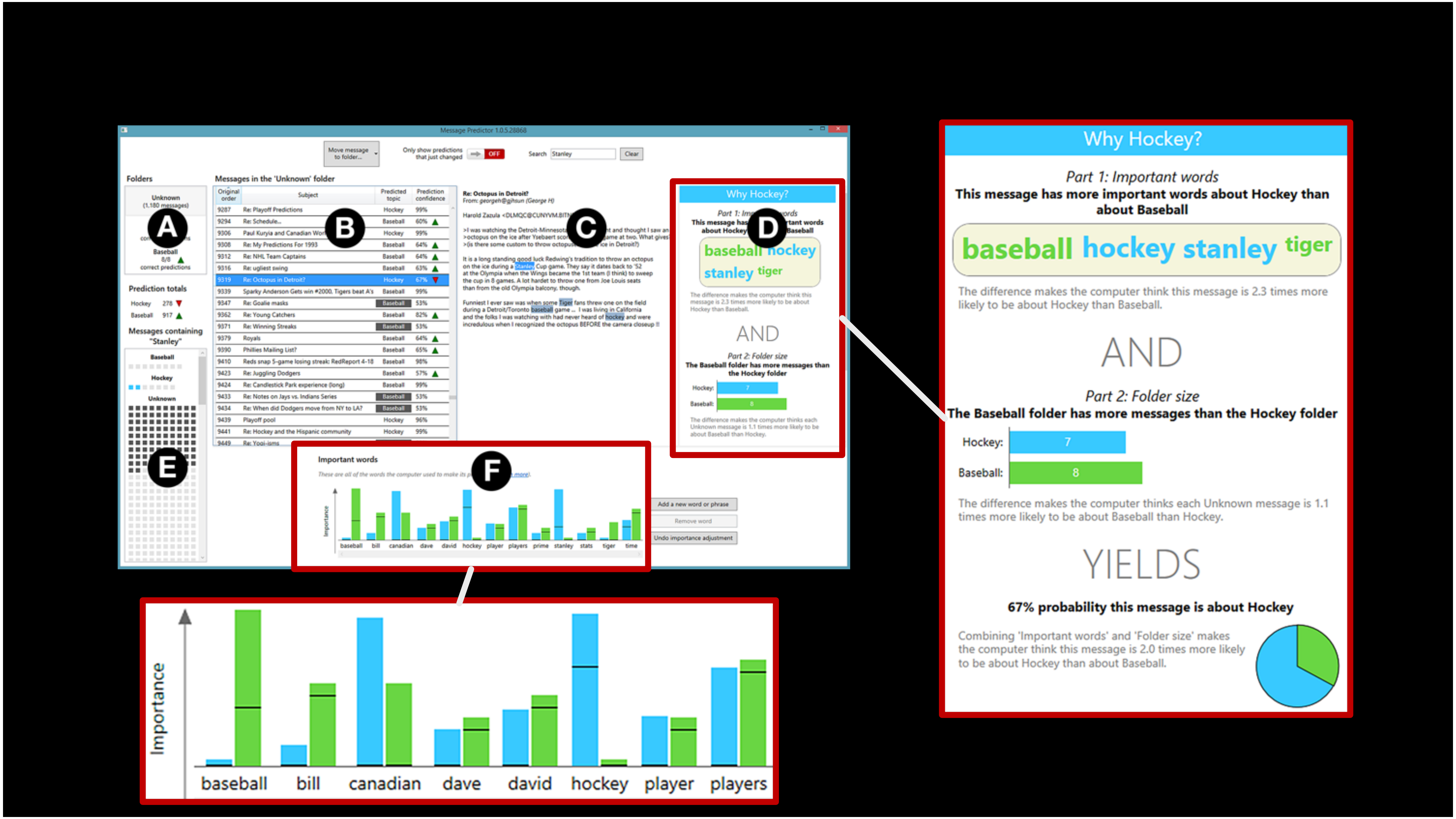
# Results

- Explanation styles:
  - Rule-based best understood
  - Keyword-based also good but negative weights problematic (absence of features)
  - Serious understandability problems with Similarity-based
  - No clear overall preference, very individual
- Potential control by users:
  - 65% feature adjustments
  - 12% feature extraction/new features
  - 5% n-grams

# Explanatory debugging principles

- Explanation
  - Iterative
  - Sound
  - Complete
  - Don’t overwhelm
- Control
  - Actionable
  - Incremental
  - Reversible
  - Honour feedback

[Kulesza et al. IUI 2015]



Message Predictor 1.0.5.28868
Move message to folder...
Only show predictions that just changed
OFF
Search
Stanley
Clear
Folders
Unknown
(1,180 messages)
Baseball
8/8
correct predictions
Prediction totals
Hockey 278
Baseball 917
Messages containing "Stanley"
Baseball
Hockey
Unknown
Messages in the 'Unknown' folder
Original order
Subject
Predicted topic
Prediction confidence
Re: Octopus in Detroit?
Important words
Add a new word or phrase
Remove word
Undo importance adjustment
A
B
C
D
E
F
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

# Study setup

- 77 participants split into two groups: 40 using EluciDebug, 37 using a version without explanations and advanced feedback
- 20 Newsgroup data set (Hockey and Baseball): initial system training on 5 messages for each subject, 1850 unlabeled messages to sort
- 30 minutes to “make the system as accurate as possible”
- Measures: accuracy, amount of feedback given, mental model scores, perceived workload
- Multinomial Naïve Bayes, retrained after every feedback

# Results

- More accurate system accuracy with less effort
  - 85% for our system versus 77% without explanations at end of study
  - Made adjustments to 47 messages while without explanations had to label 182 messages
- With better understanding
  - 15.8 mental model score versus 10.4
  - The more you understand, the better you can make the system
- Do not overwhelm
  - No difference in workload measures

# Intelligibility revisited

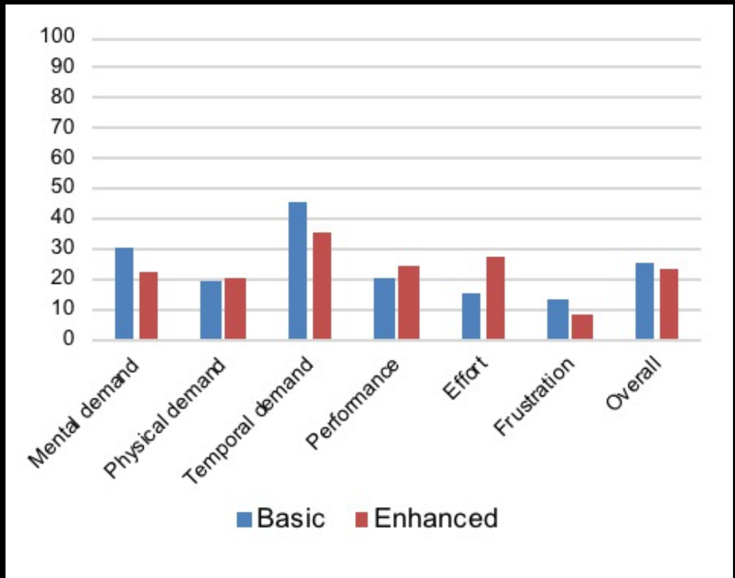
- Wearable system for blind users to identify people
  - Information by system is provided in a continuous stream
  - Blind users can't use visual explanations, spoken explanations would interfere with system use

[Ahmed et al. IUI 2020]


# Methods

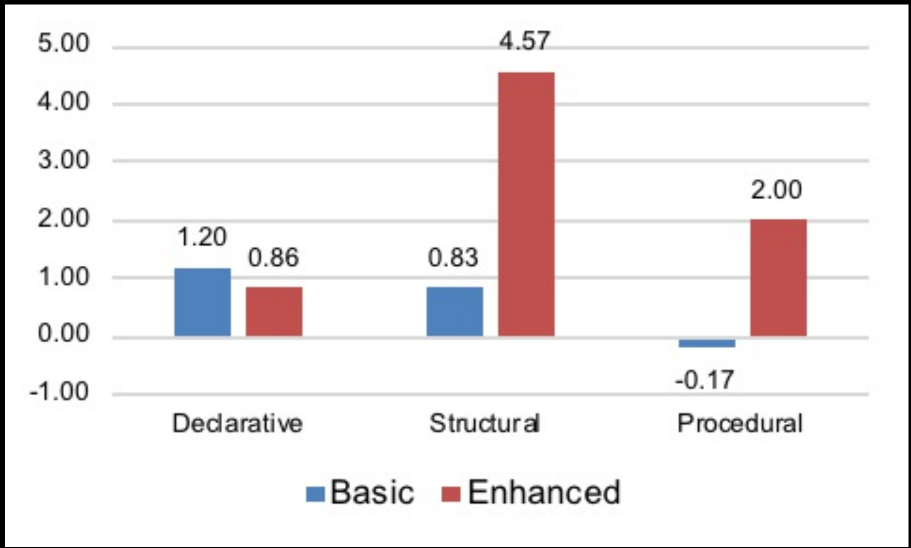
- 13 participants (12 male, 1 female), average age 20.85
- All registered blind but with varying visual abilities (e.g. some light perception, some can see objects from 3-6 metres away), many had been blind since birth
- Instructions: Basic (what sounds means) and Enhanced (how system works e.g. to detect a person it will need to see the head or the torso)
- Measures
    - NASA-TLX involving a tactile scale
    - Task success: Time to locate the recruiter, accuracy of ID (percentage of NEW or UNDETECTED instances until the correct ID )
    - Knowledge levels: declarative, structural, procedural
    - Behaviour strategies: Gaze, Walking

# Results – User Experience




- No diff between groups for NASA-TLX
  - Headset was quite heavy
  - Duration of task really short
  - Difficulties with misidentifications and direction of sound
- No diff between groups for system accuracy or time to locate

# Results – Knowledge Gained



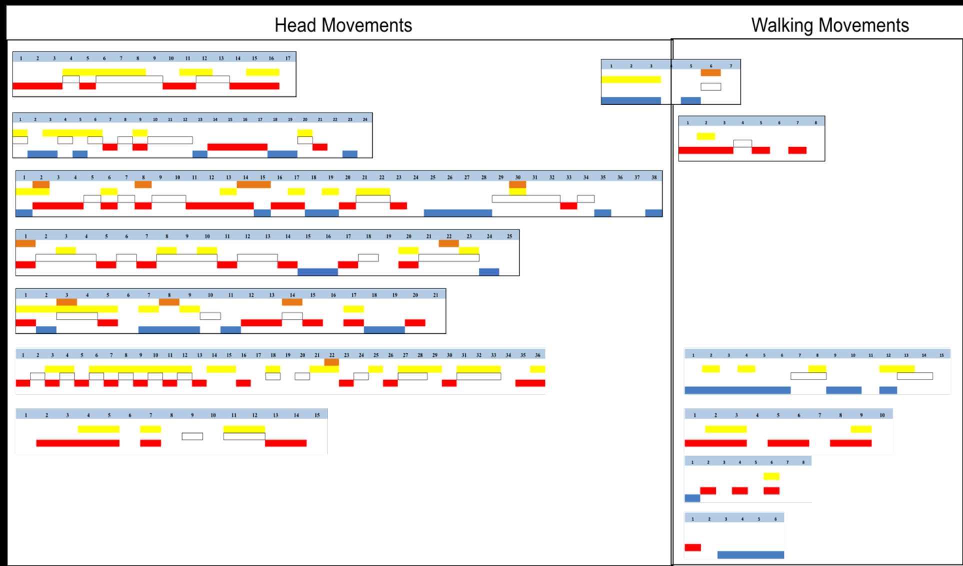

- No diff between groups for declarative
- Enhanced had better structural and procedural knowledge
  - Structural crucial for knowing **the cause** if something goes wrong
  - Procedural is needed to know **what to do** if it goes wrong
- Structural difficult to learn from basic instructions
- Nobody got taught Procedural. Enhanced used structural knowledge to build procedural knowledge

# Results – Strategies used

- More participants in the Enhanced group than the Basic group used horizontal head movements to explore their environment

- Basic group mainly used walking to explore the space

- Enhanced strategies better suited to technology

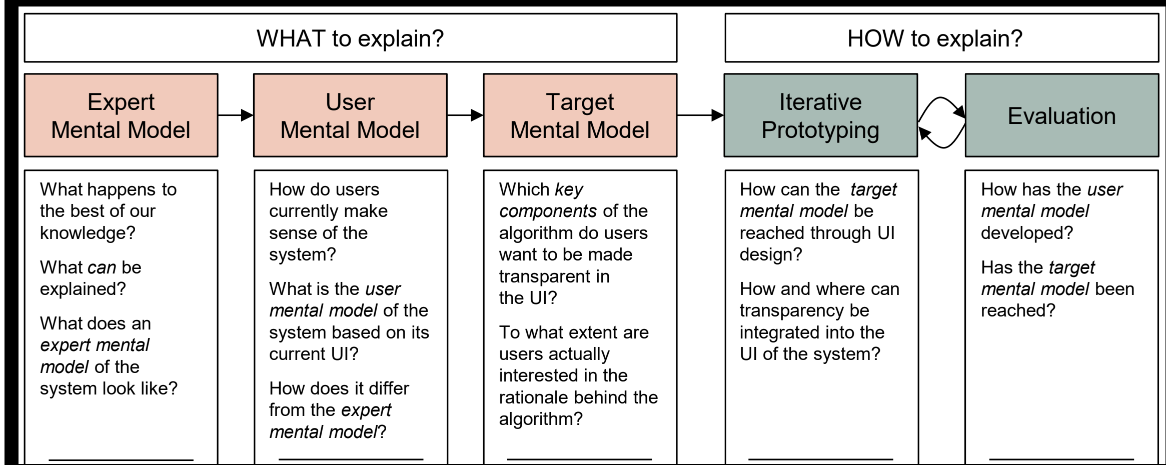



**Horizontal (yellow) and vertical (orange) head movements, stopping (white), and walking slowly (red) and at a normal pace (blue). Enhanced group participants' journeys are outlined in black.**

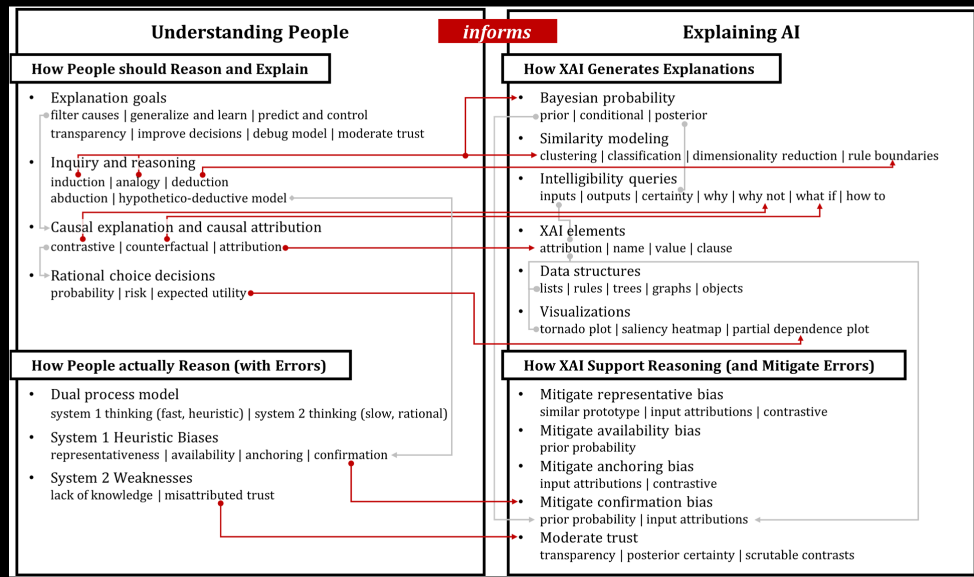
# What do we know so far?

- We can help users build better mental models by making information available how of how ML works
- Not enough to just explain a decision, need to know a bit about how system works
- Better mental models help to spot when system goes wrong and to use these interactive systems better

# Designing for Intelligibility



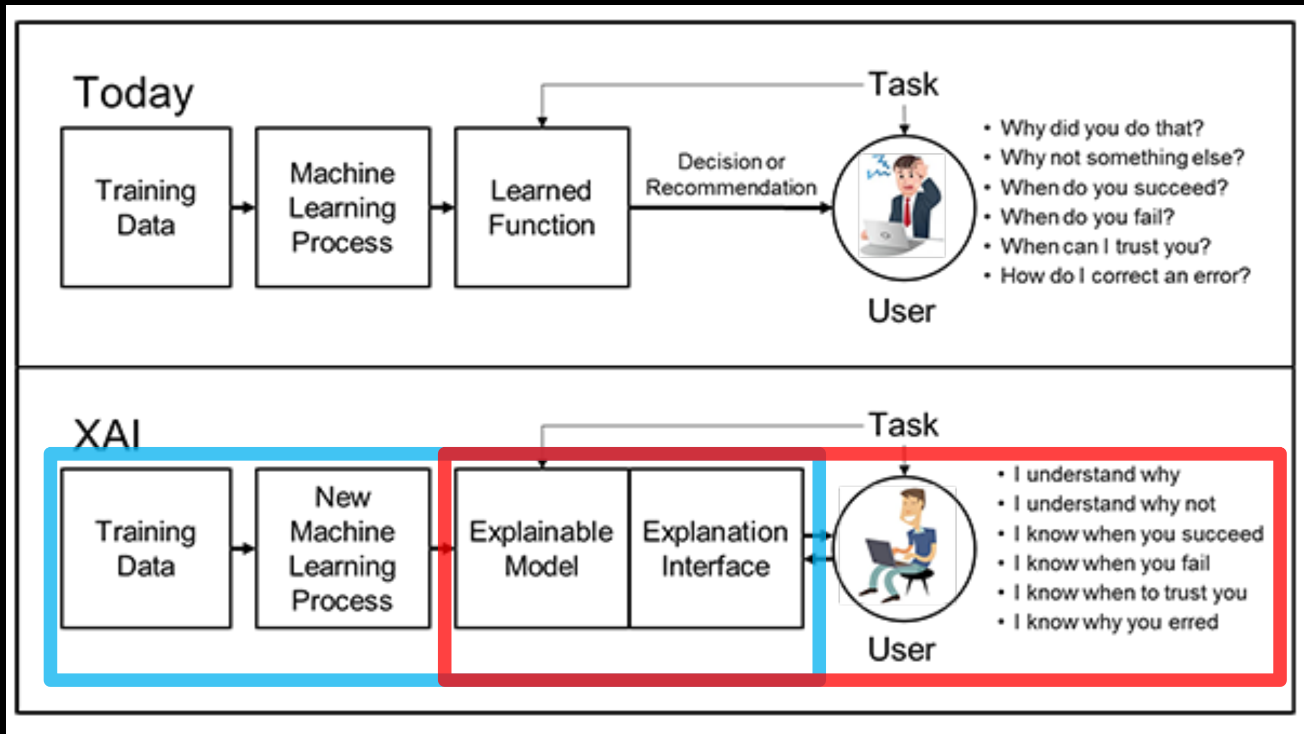

[Eiband et al. IUI 2018]




[Wang et al. CHI 2019]

- Essentially hand-crafted for each user group and each AI system

# XAI vision reprised




"Appropriate trust"

**Technical** **Human**

# Complex socio-technical system

How does it work?

**Physical system**

What is the purpose?

**Structure**

**Task**

What does it do?

**People**

Who is the user?

# Structure

- Why explain?
  - Increased adoption / trust / satisfaction
  - Better use / appropriate trust
  - Spot the mistakes / identify biases
  - Learn from user

# Physical systems

- How does it work?
  - Models
  - Interfaces
  - Interactions

# People

- Who are we explaining to?
    - Expectations and attitudes
    - Capabilities
    - Mental models

# Tasks

- What decisions/ recommendations/actions are we trying to explain?
  - High stake versus low stake
  - Level of automation
  - Situational context

# XAI Research challenges

- No explanations desired for certain tasks and contexts

[Bunt et al. IUI 2012]

- Different people need different explanations

[Gunning et al. Science Robotics 2019]

- “Placebic” explanations and persuasive force

[Eiband et al. CHI 2019, Bussone et al. ICMI 2015]

- Trust and Reliance

[Holliday et al. IUI 2016, Nourani et al. HCOMP 2019]

- Perceived control increases user satisfaction

[Smith-Renner et al. CHI 2020]

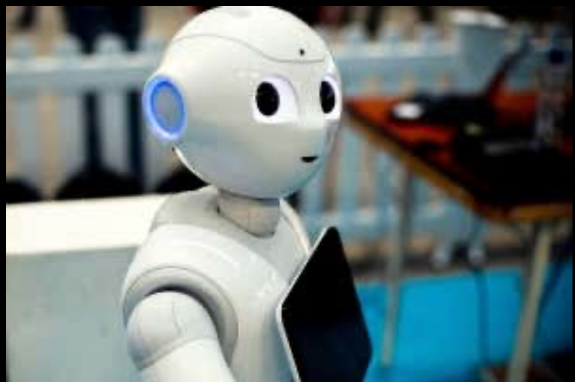
- Explanations might be outside of the ML

[Ehsan et al. CHI 2021]

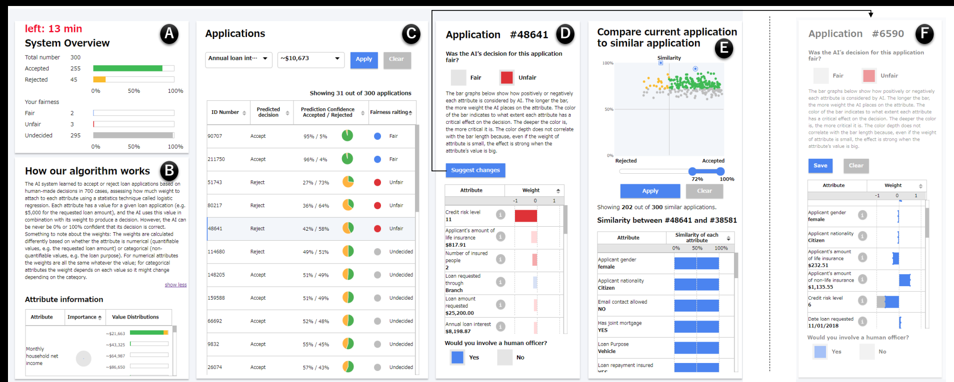
# New frontiers for XAI

- Making more complex ML intelligible
  - Reinforcement learning, Deep learning
  - Structured explanations

- Apply XAI to new areas



[Sawal New Scientist 21/04/2021]




[Stumpf et al. TiiS forthcoming]

# Five take-aways

- Explain with humans in mind
- Know why you are explaining and what you are explaining
- Think about different ways of explaining best suited to users and situation
- Be aware of unintended effects
- Plenty of work left to do!