Not a new thing…

• Lots of work on explanations back in the 1980s and 1990s
  • neoMYCIN, etc to explain expert systems
• Ben Shneiderman versus Pattie Maes debate in the 1990s
  • Transparency and trust important
• Also work on autonomous agents and decision-making in the early 2000s
• But really ramped up since 2010s with rise in machine learning and black-box models

Mycin: Is Meningitis a hospital acquired illness?
Doctor: why?
Mycin: [2.0] this will aid in determining the organisms 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)
Regulation, standards and guidelines

• Industry-led by tech giants
  • Microsoft’s Guidelines for Human-AI Interactions
  • Google’s Responsible AI practices and People+AI Handbook
  • IBM’s Everyday Ethics for Artificial Intelligence
  • Fujitsu’s AI Ethics Impact Assessment Practice Guide
• EU Assessment List for Trustworthy Artificial Intelligence (ALTAI)
• EU AI Act (in draft)
  • High harm applications need to be assessed and **transparent**
  • Generative AI will have to be **transparent**
Transparency
So what is AI “transparency”? 

- How the AI model works
- Why a specific prediction was made by the AI … or not

- Currently somewhat overlooked:
  - Why was the model developed in the first place
  - What training data was used to develop the model
  - How was the model evaluated
  - How good is it
  - What biases or blind spots does it have
  - What decisions about the AI were made during its development
Explainable AI (XAI)
Explainable AI (XAI) vision (2016)

Calibrated / appropriate trust
Motivation for XAI

Model understanding is absolutely critical in several domains, particularly those involving *high potential for harm*, to support *debugging, bias detection* and *recourse*.
Lots of work to make ML ‘explainable’  [Molnar 2022]

- Global explanations:
  - Exposing the model
- Local explanations:
  - Exposing (combination of) features that contribute to a decision
Local explanations
LIME: Local Interpretable Model-Agnostic Explanations

- Explains important feature that led to a decision
- Uses a post-hoc explanation on a simplified model
- Another popular method which outputs feature importances: SHAP

[Ribeiro et al. KDD 2016]
Prototypes/Example

- Use examples (synthetic or natural) to explain individual predictions
  - Identify instances in the training set that are responsible for the prediction of a given test instance
  - Identify examples (synthetic or natural) that strongly activate a function (neuron) of interest
Counterfactual Explanations

What features need to be changed and by how much to flip a model’s prediction?

\[ l = \text{Crested Auklet} \quad l' = \quad c = \text{Red Faced Cormorant} \]

[Mothilal et al 2020]
Saliency Maps

What parts of the input are most relevant for the model’s prediction: ‘Junco Bird’?
But beware: “explanation” might be misleading

Model parameter randomization test

Global explanations
How important is the notion of “stripes” for this prediction?

[Kim et. al., 2018]
Model Distillation

Black Box Model

Data

v1, v2
v11, v12

Model Predictions

Label 1
Label 1

Label 2

Explainer

Simpler, interpretable model which is optimized to mimic the model predictions
Human-Centric Explanations
Explainability versus Interpretability

• Explainability = **system-centric** ability of an AI system to explain itself

• Interpretability = **human-centric** 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 a kind of internal representation in someone's thought process for how something works in the real world

• Users build mental models to guide how they interact, behave or fix things when they go wrong through
  • Extending and adapting existing mental models
  • Exploring and using a system
  • Being taught or having things explained

See:
- Norman 1983
- Johnson-Laird 1983
Lots of work to make explanations ‘useable’

• What should be explained?
  • Global/local explanations, intelligibility types, etc.

• How should we explain?
  • Natural language dialogue, textual explanations, visualisations, etc.
Intelligibility types [Lim and Dey CHI 2009]

• 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?
Explanation content versus explanation presentation/style

• What information is transmitted in an explanation versus its form and presentation
• E.g. decision confidence

0.67341
67% Accept / 33% Reject

I think it’s a little bit more likely that this application should be accepted.
Different stakeholders = different explanations?

- End users / lay users (e.g. loan applicants)
- Decision makers / domain experts (e.g. doctors, judges)
- Regulatory agencies (e.g. FDA, European commission)
- Researchers, developers and engineers
Human-centric explainable AI (HCXAI) design

- Need to know who the user is

- Global or local explanations or both?

- Global explanations
  - How the model works
  - The accuracy of the model
  - Important features

- Local explanations
  - Important features for this decision
  - Decision confidence
• What explanation styles do end-users prefer?

[Stumpf et al. IJHCS 2009]
Explanation styles

Keyword

From: buglove@houston.r.com
To: L.Farmer@enron.com
Subject: life in general

Good and — where do you find time for all of that? You should u...

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 pr... a few million 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 lov...

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

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 “proposals” and
  - The message does not contain the word “farm” and
  - The message does not contain the word “link”.

- Put the email in folder “Enron News” if:
  - No other rule applies.

Similarity

Message #2
From: lision@enron.com
To: All ENW employees
Subject: enron net works title policy

Please print and become familiar with the updated ENW TIE P... business-first travel, with supervisor approval, for international fl... 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 #1 in “Resume” because #2 and #3 have important words in common.

Message #3
From: lisa.coklo@enron.com, lisa.coklo@enron.com,
daren.farmer@enron.com,
daren.farmer@enron.com
To: rita.wynn@enron.com,
daren.farmer@enron.com
Subject: confirming requisitions

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

Lisa Coklo 104155, 104001
Rita Wynn 104364
Daren Farmer 104219
Mike Eiben 103523
Pat Clynes 104295

The posting dates have all been updated to reflect a current revision 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
Explanatory debugging for interactive machine learning

Future improved behaviour

Feedback/Control

Explanation

Improved mental model, satisfaction

See:
- Stumpf et al. IJHCS 2009
- Kulesza et al. TiiS 2011
- Kulesza et al. CHI 2012
- Das et al. AI 2013
- Kulesza et al. IUI 2015
Explanatory debugging principles

• Explanation
  • Iterative
  • Sound
  • Complete
  • Don’t overwhelm

• Control
  • Actionable
  • Incremental
  • Reversible
  • Honour feedback
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:**  
**Baseball:**

The difference makes the computer think each Baseball 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.
Results

• More accurate system 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

• Does not overwhelm
  • No difference in workload measures
HCXAI 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]; lay users neglected at the moment

• Explanations calibrate trust and reliance [Bussone et al. ICMI 2015, Holliday et al. IUI 2016, Nourani et al. HCOMP 2019]; “placebic” explanations [Eiband et al. CHI 2019]

• Explanations might come from outside of the ML [Ehsan et al. CHI 2021]

• Explanations, and then what? [Wang et al. 2022]
Transparency for Fairness
Bias and fairness in AI

• Biased humans produce biased data which gets trained into model or AI can also go ‘rogue’ and produce a biased model

• Tools to find and mitigate bias are emerging
  • 20+ different fairness metrics
  • IBM AI Fairness 360
  • FairML
  • Google’s What-if

• Fairness is a human value and can’t be necessarily reduced to metrics
  • Need transparency to understand if something is fair (or not)
  • Human-in-the-loop fairness tools such as FairVis, FairSight, etc
Towards Involving End-users in Interactive Human-in-the-loop AI Fairness [TiiS 2022]

- Leverages work from Explanatory Debugging -> find and fix fairness ‘bugs’ that do not meet users’ expectations
- Loan application domain, anonymized dataset from a partner bank
- 388 participants recruited through Prolific, no technical or domain expertise needed
- Logistic regression, 61.8% accuracy, failed DI metric (0.718) on Nationality attribute
- Using the average weight value for each attribute suggested by the participants on an application, recalculated model
Results

• Participants found UI useful and easy to use
• Used UI to find problematic decisions through sorting/filtering on Predicted decision, Confidence, Comparison
• 20% of assessed decisions judged unfair, mostly honing in on Nationality
• When looking at Nationality, unfair was applied to 57.6% of accepted citizens, 14.4% for rejected foreigners
• 230 participants made weight changes to 3.71 applications on average
• Suggested weight changes improved DI to 0.814
  • 50% of participants increased DI (M=0.91), other half decreased it (M=0.63)
Transparency for other kinds of AI
Problems with current explanations for generative or autonomous AI

• Explanations are delivered in visual form – no good for certain situations or people
• Explanations are meant to be pondered – not sure how to integrate into real-time settings for human-AI collaboration
• Currently we have a narrow view of explanations – what do we mean by ‘explanations’ and what should be explained
  • Why was the model developed in the first place
  • What decisions about the AI were made during its development
  • What training data was used to develop the model
  • How was the model evaluated
  • How good is it
  • What biases or blind spots does it have
Model Cards [Mitchell et al. 2019]

- Model Details. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- Intended Use. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- Factors. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- Metrics. Metrics should be chosen to reflect potential realworld impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- Evaluation Data. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
  - Training Data. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- Quantitative Analyses
  - Unitary results
  - Intersectional results
- Ethical Considerations
- Caveats and Recommendations
Summary

- Transparency is required and XAI has made some strides towards opening the black box
- However, ‘transparency’ is a very vague term and ‘explanations’ can come in different forms
- Need for a human-centred approach to transparency and explanations
- Consider what explanations are used/useful for


• Don Norman. 1983. Some observations on mental models. Lawrence Erlbaum Associates, Hillsdale, New Jersey, US.


• Thanks to Hima Lakkaraju and her tutorial on XAI!!

Resources
Resources


