
University of Glasgow
Oilthigh Ghlaschu

# Transparency and Explainable AI (XAI)

Prof Simone Stumpf

Simone.Stumpf@glasgow.ac.uk

University of Glasgow
Oilthigh Ghlaschu

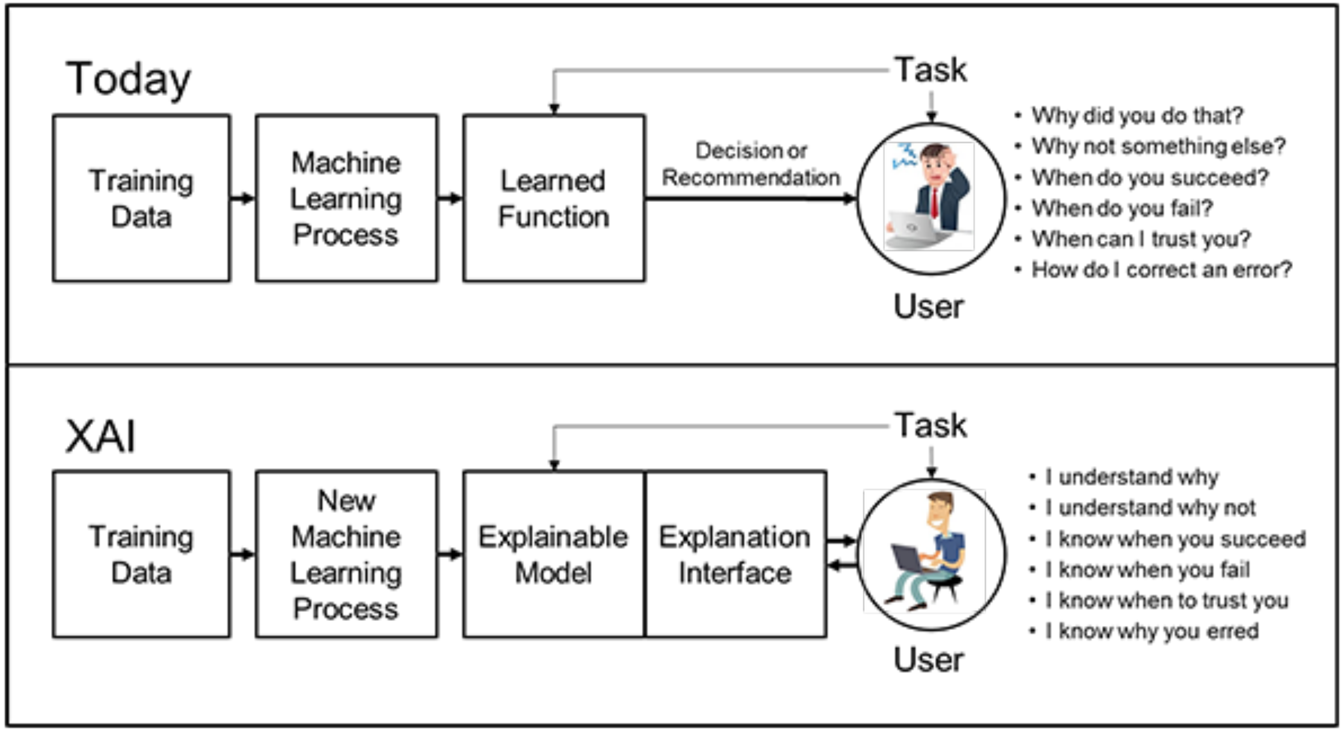
# History

# 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 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)

# Explainable AI (XAI) vision (2016)




Calibrated / appropriate trust

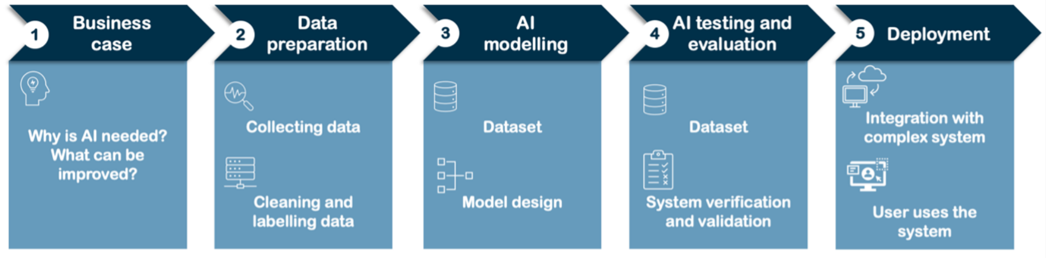
University of Glasgow
Oilthigh Ghlaschu

## 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 (ratified earlier this year)
  - High risk applications need to be assessed and **transparent**
  - Generative AI will have to be **transparent**

# 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)

# 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 AI ‘explainable’

[Molnar 2022]

- Global explanations:
  - Exposing the model

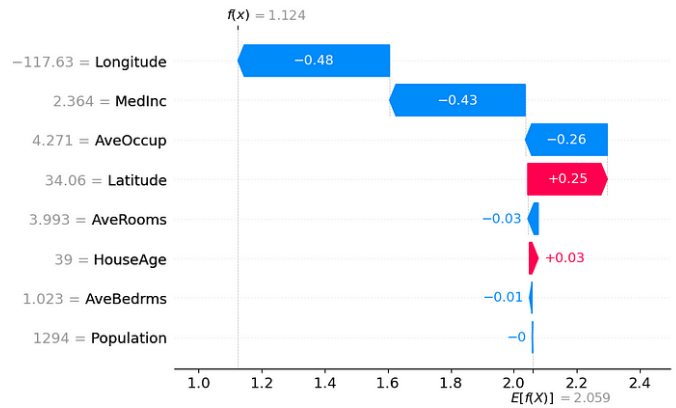
- Local explanations:
  - Exposing (combination of) features that contribute to a decision

# 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]



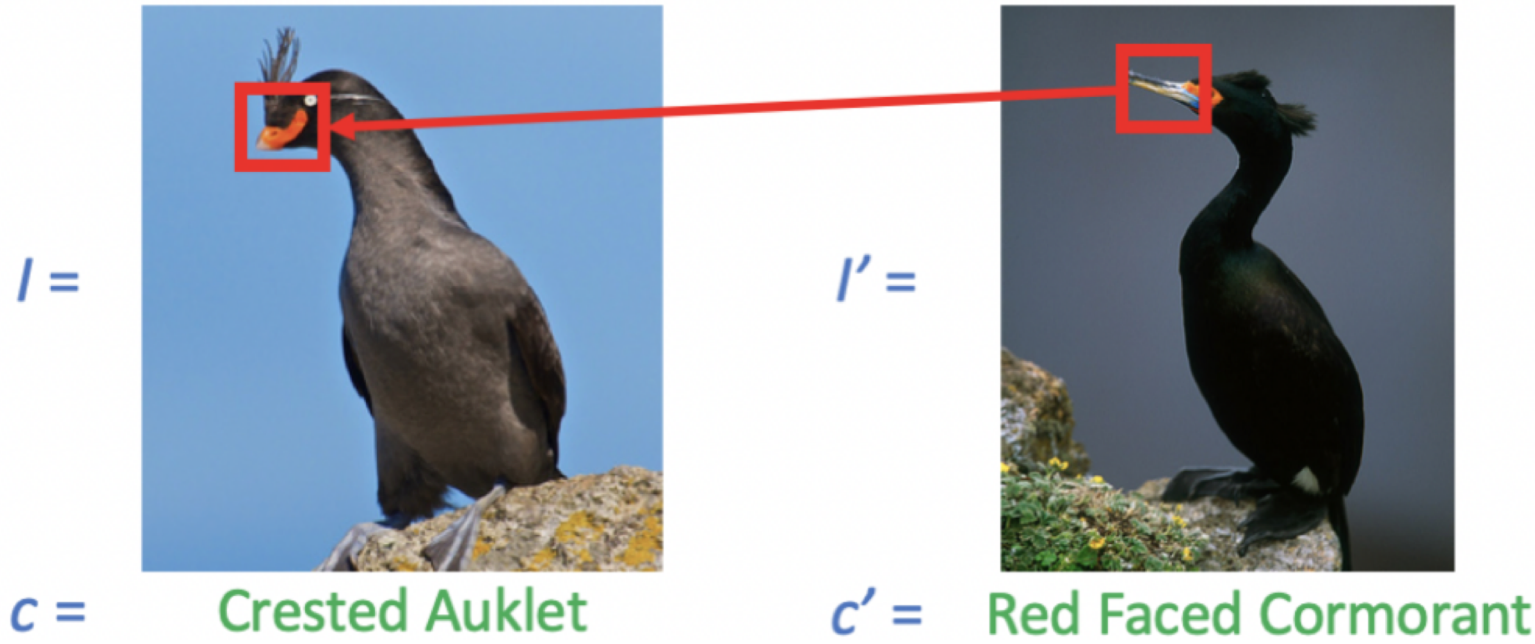
# Prototypes/Example

- 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

University of Glasgow
Oilthigh Ghlaschu

# Counterfactual Explanations

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

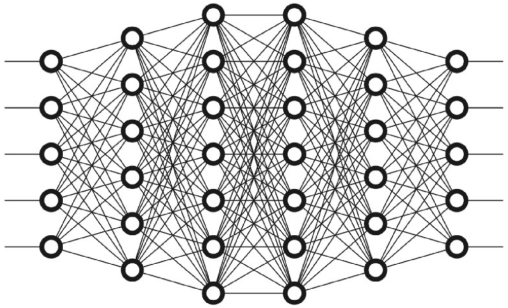



[Mothilal et al 2020]

University of Glasgow
Oilthigh Ghlaschu

# Saliency Maps

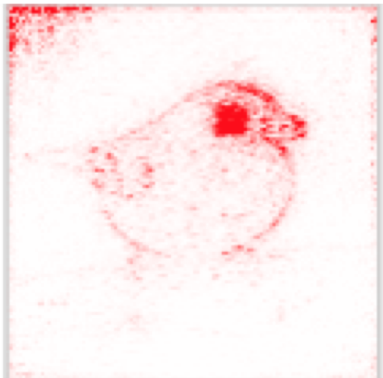
**Input**



**Prediction**

**Junco Bird**

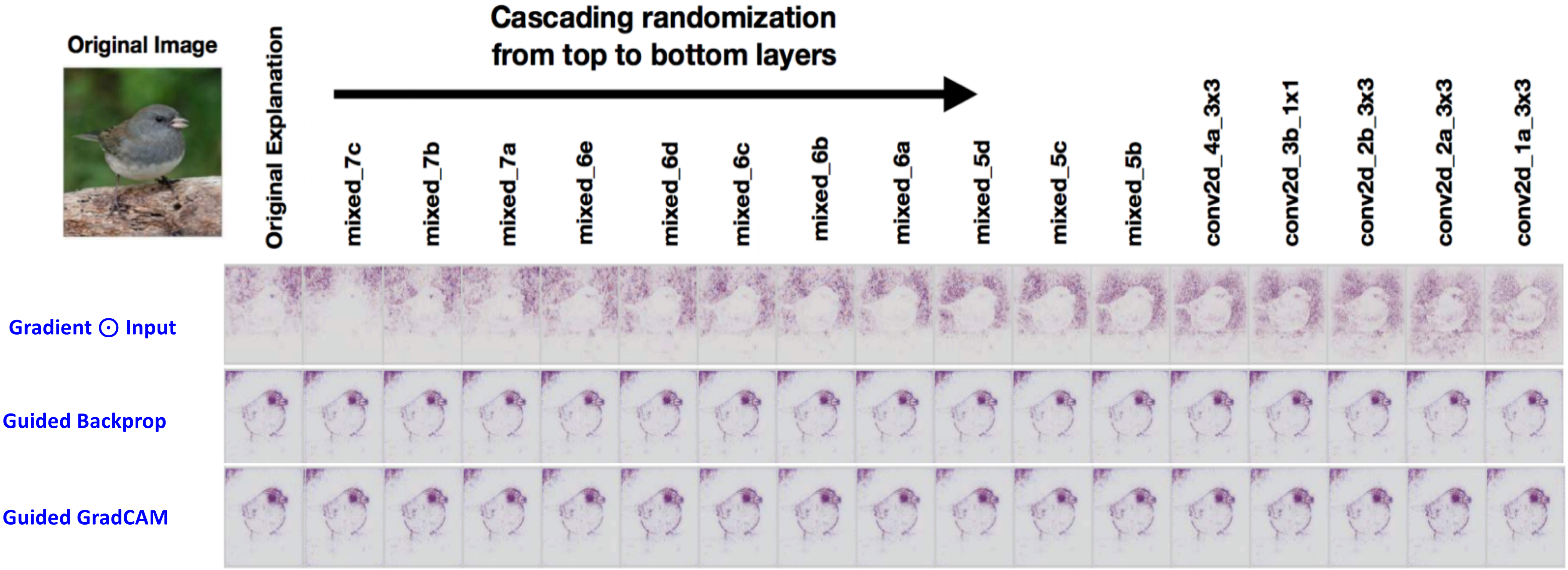
**What parts of the input are most relevant for the model's prediction: 'Junco Bird'?**



**Saliency Map**

# But beware: “explanation” might be misleading

## Model parameter randomization test

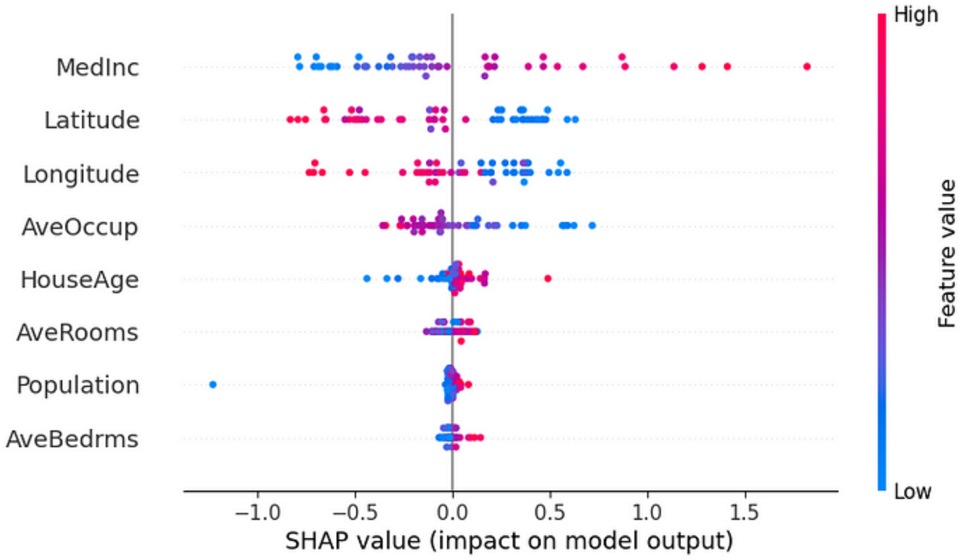



Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

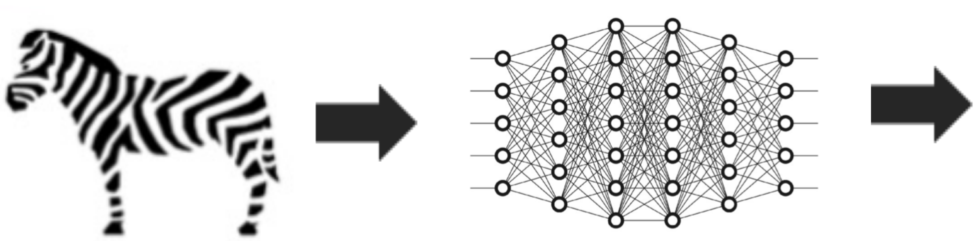
# Global explanations

# SHAP (SHapley Additive exPlanations)

- Explains model overall i.e. the importance of features globally and how feature values contribute to a decision



# Representation Based Explanations



Zebra
(0.97)

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

[Kim et. al., 2018]

# Human-Centric Explanations

University of Glasgow
Oilthigh Ghlaschu

# 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’

For mental model 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.

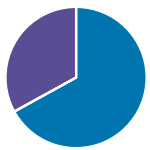
University of Glasgow
Oilthigh Ghlaschu

# 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.

University of Glasgow
Oilthigh Ghlaschu

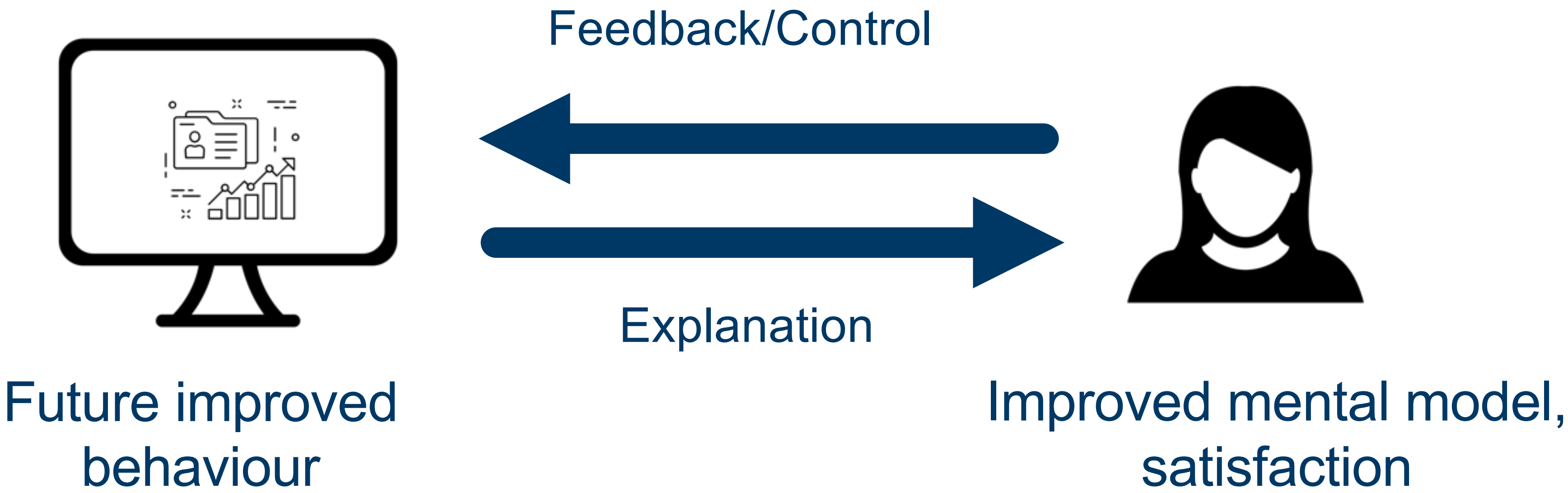
## Different stakeholders = different explanations?

- End users / lay users (e.g. loan applicants, patients)
- Decision makers / domain experts (e.g. doctors, judges, loan officers)
- Regulatory agencies (e.g. FDA, European commission)
- Researchers, developers and engineers

## Human-centric explainable AI (HCXAI) design

- Need to know who the user is and what they need to know
- What is the purpose of the explanations?
- Usually a combination of global and local explanations
- Measure explanations in terms of the purpose and other associated effects

# Explanatory debugging for interactive machine learning

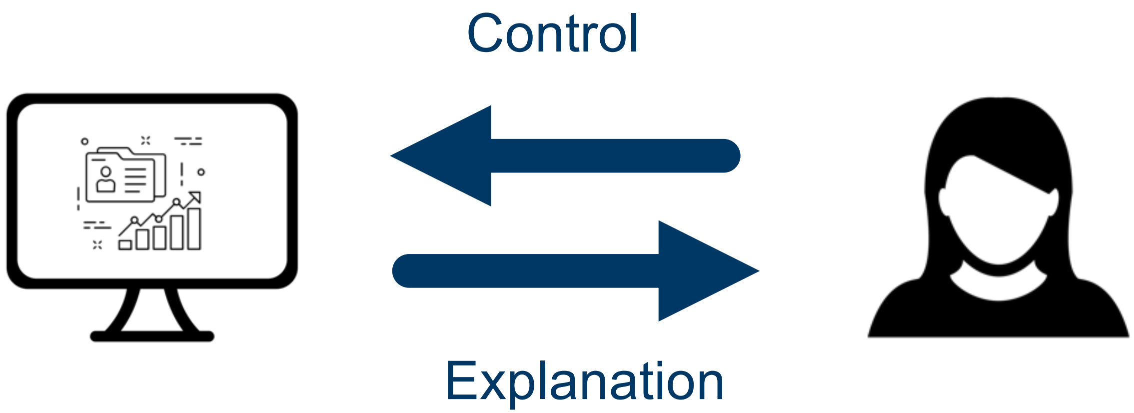



See:
- Stumpf et al. IJHCS 2009
- Kulesza et al. TiiS 2011
- Kulesza et al. CHI 2012
- Das et al. AI 2013
- Kulesza et al. IUI 2015

University of Glasgow
Oilthigh Ghlaschu

# Explanatory debugging principles

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

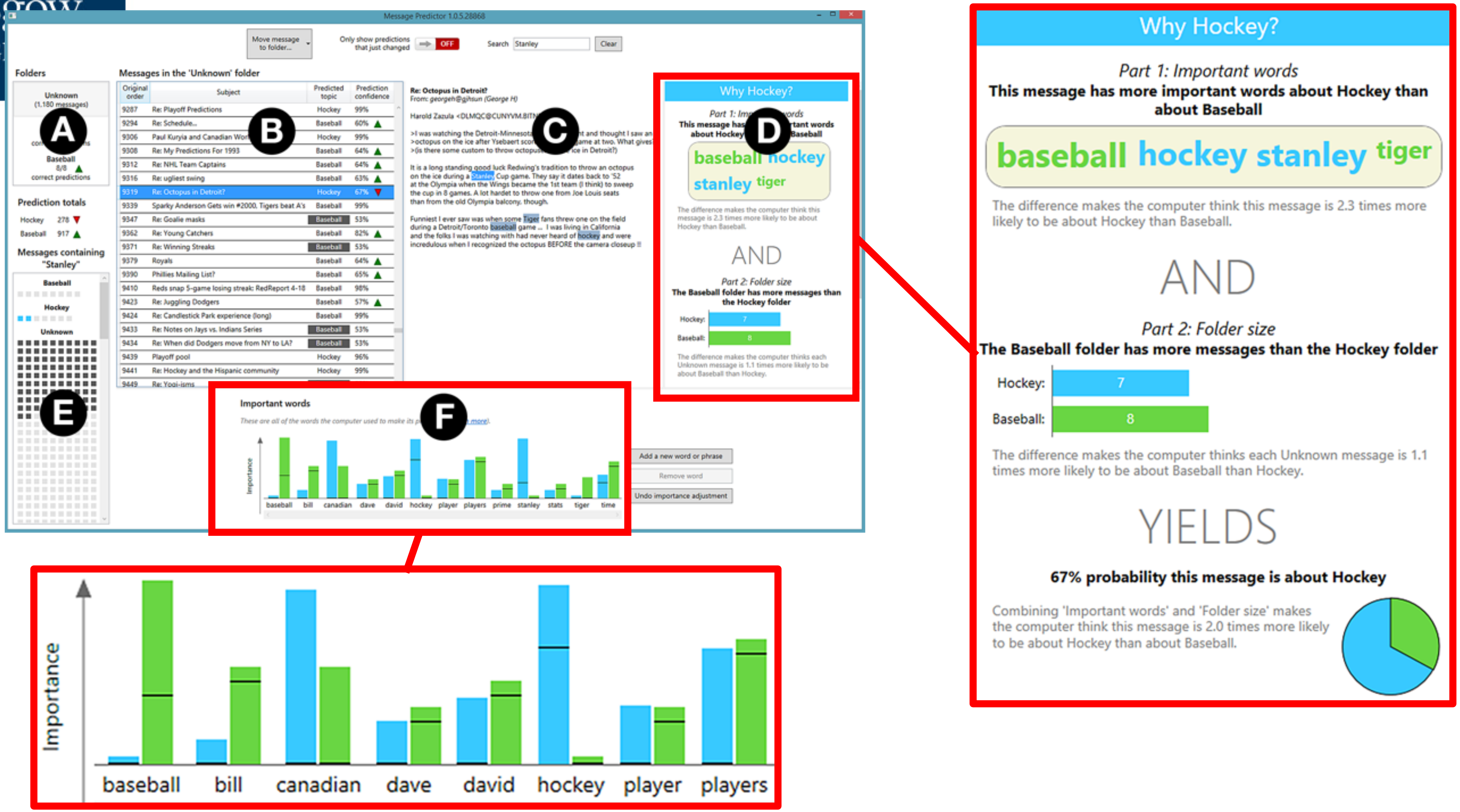


University of Glasgow
Oilthigh Ghlaschu

# Study design

- 2 versions
  - Elucidebug with explanations and ability to control with feature feedback
  - Control with no explanations and only ability to label instances to correct AI
- 20 newsgroups: Hockey and baseball
- 77 participants

- Measurements
  - NASA TLX
  - Mental model score
  - Amount of feedback
  - Accuracy F1 score



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

## Results

- More accurate system with less effort
  - 0.85 for our system versus 0.77% without explanations at end of study
  - Made adjustments to 47 messages while without explanations had to label 182 messages

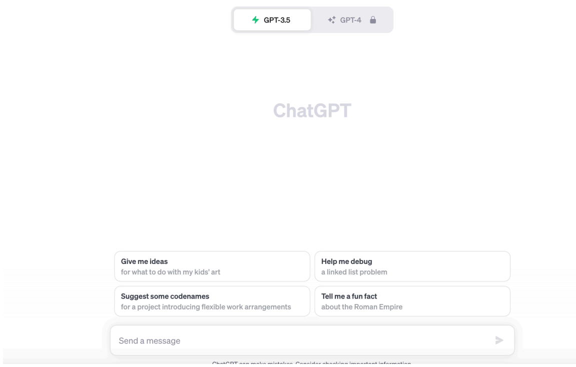
- With better understanding
  - 15.8 mental model score versus 10.4
  - The more you understand, the better you can make the system

- Does not overwhelm
  - No difference in NASA TLX 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 other kinds of AI

University of Glasgow
Oilthigh Ghlaschu

## 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

# Resources


- Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. interactions 4, 6: 42–61. https://doi.org/10.1145/267505.267514
- Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *Int. J. Hum.-Comput. Stud.* 67, 8: 639–662.
- Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 126–137. https://doi.org/10.1145/2678025.2701399
- Don Norman. 1983. Some observations on mental models. Lawrence Erlbaum Associates, Hillsdale,New Jersey,US.
- David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. Science Robotics 4, 37. https://doi.org/10.1126/scirobotics.aay7120
- Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. 2119–2128. https://doi.org/10.1145/1518701.1519023
- Christoph Molnar. Interpretable Machine Learning. Retrieved February 5, 2020 from https://christophm.github.io/interpretable-ml-book/
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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 (KDD '16), 1135–1144. https://doi.org/10.1145/2939672.2939778
- Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (Un)reliability of Saliency Methods. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen and Klaus-Robert Muller (eds.). Springer International Publishing, Cham, 267–280. https://doi.org/10.1007/978-3-030-28954-6_14
- Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT* '20), 607–617. https://doi.org/10.1145/3351095.3372850
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In International Conference on Machine Learning, 2668–2677. Retrieved December 11, 2018 from http://proceedings.mlr.press/v80/kim18d.html
- Thanks to Hima Lakkaraju and her tutorial on XAI!
- 1.
- Zijie J. Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E. Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. 2022. Interpretability, Then What? Editing Machine Learning Models to Reflect Human Knowledge and Values. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD '22), 4132–4142. https://doi.org/10.1145/3534678.3539074

# Resources


- Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 1–7. https://doi.org/10.23919/FAIRWARE.2018.8452913
- R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development 63, 4/5: 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287
- Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. ACM Transactions on Interactive Intelligent Systems 12, 3: 18:1-18:30. https://doi.org/10.1145/3514258
- Yuri Nakao, Lorenzo Strappelli, Simone Stumpf, Aisha Naseer, Daniele Regoli, and Giulia Del Gamba. 2022. Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. International Journal of Human–Computer Interaction 0, 0: 1–27. https://doi.org/10.1080/10447318.2022.2067936
- Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 9: 389–399. https://doi.org/10.1038/s42256-019-0088-2
- Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19), 1–13. https://doi.org/10.1145/3290605.3300233
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* ’19), 220–229. https://doi.org/10.1145/3287560.3287596