Tutorial on Human-Centred Explainable AI (HCXAI)
Principles, Guidelines, and Practical Examples

THE 22ND AUSTRALASIAN DATA SCIENCE AND MACHINE LEARNING CONFERENCE (AUSDM'24) | Melbourne, Australia | 25-27 November 2024

Dr Ronal Singh

Research Scientist

Human-Centric Security & CINTEL FSP,

CSIRO's Data61

Prof Tim Miller

Professor of AI

School of Electrical Engineering and Computer Science

The University of Queensland

Acknowledgement of Traditional Owners

I would like to begin by acknowledging the Wurundjeri people of the Kulin Nations as the Traditional Owners of the land that we're meeting on today, and pay my respect to their Elders past and present.

Who are we?

Dr Ronal Singh

Research Scientist

Human Centric Security & CINTEL FSP
CSIRO's Data61

Prof Tim Miller

Professor of Artificial Intelligence

The School of Electrical Engineering and Computer Science
The University of Queensland

Agenda

Item	Description	Time (mins)
Introduction	Welcome and Objectives. Overview of XAI	20
Taxonomy of Explanation Types	Taxonomy of Explanation: Explain local vs. global, model-agnostic vs. model-specific Foundational XAI Methods: feature-based, example-based, and concept-based.	30
Overview of HCXAI	Key HCXAI principles, such as understanding, trust, and user centred design.	30
BREAK		15-20
Designing for User Engagement and Control	Techniques to improve engagement, control and enhance decision-making	30
User Studies in XAI	Evaluation metrics in HCXAI, focusing measure such as understanding, trust, and task performance.	30
Conclusion and Q&A	Summary of Key Concepts and Q&A	15-20

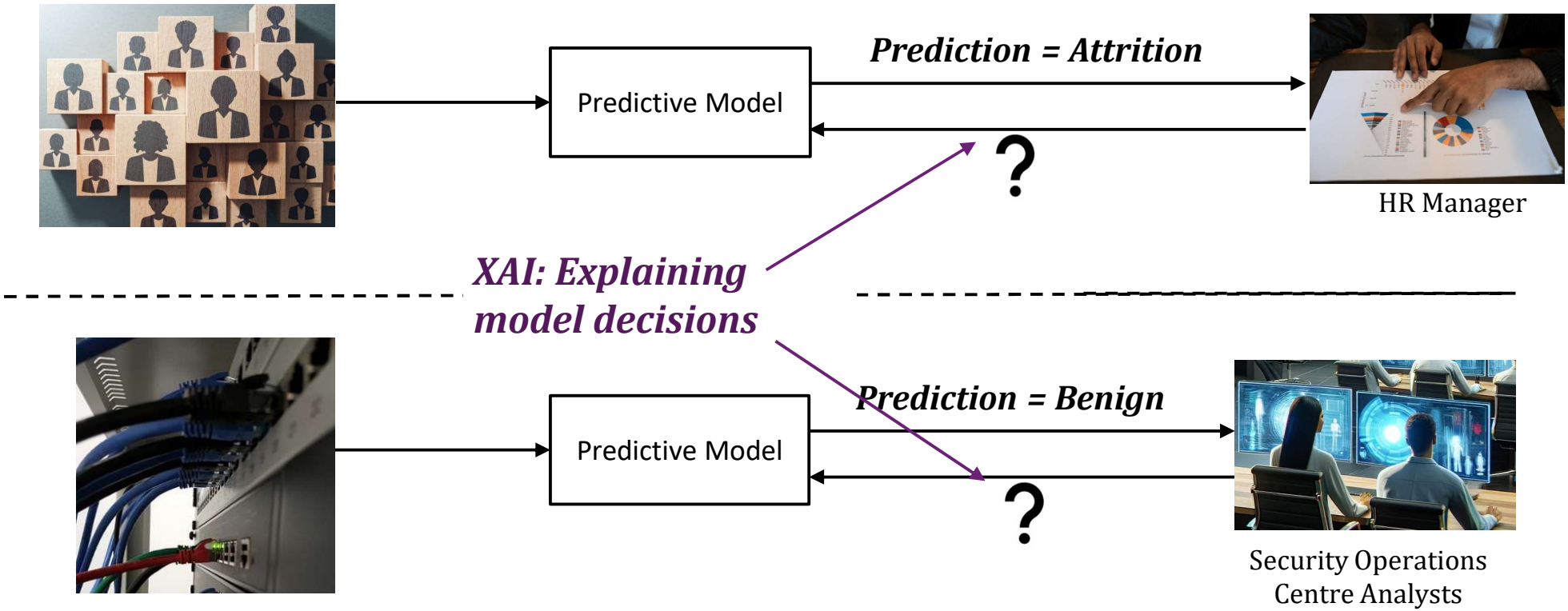
Tutorial Materials Link:

What you get:
1) Slides
2) Examples of Code

Part 1: Welcome and Overview

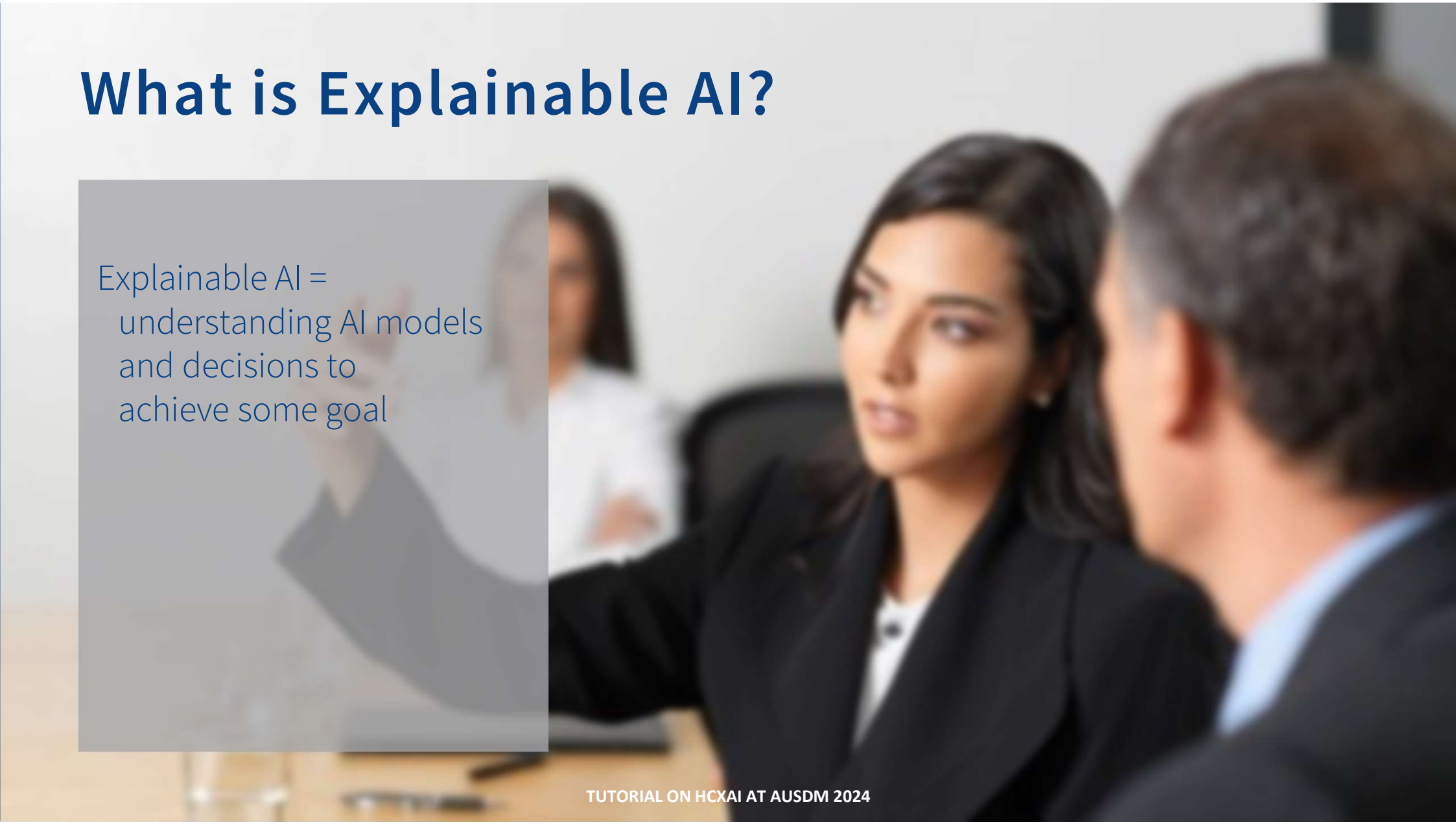
What do you think are the most important factors that make Explainable AI effective and impactful for its users?

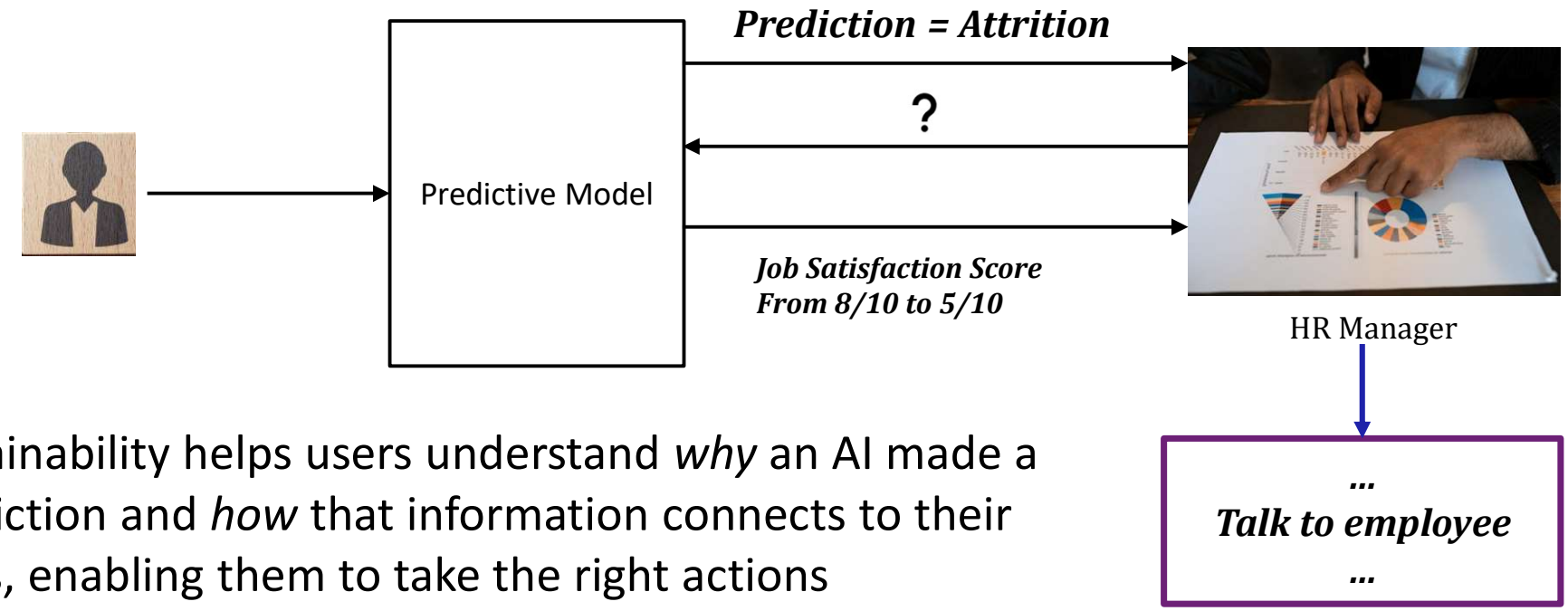
AI/ML applied in many domains, including healthcare, law, cybersecurity


Predictive Model
Prediction = Attrition
?
HR Manager
XAI: Explaining model decisions
Predictive Model
Prediction = Benign
?
Security Operations Centre Analysts

What is Explainable AI?

Explainable AI =
understanding AI models
and decisions to
achieve some goal





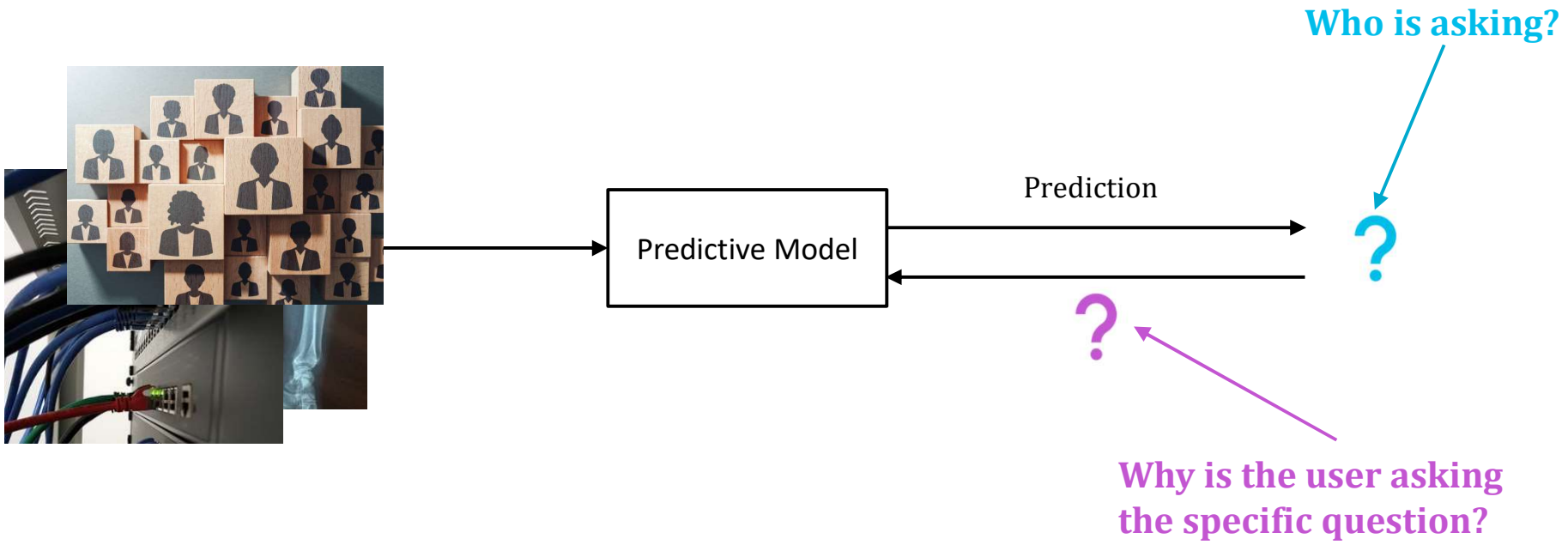
Explainability helps users understand *why* an AI made a prediction and *how* that information connects to their goals, enabling them to take the right actions

What is Explainable AI?

Explainable AI =
understanding AI models
and decisions to
achieve some goal

Explanation answers a
why question

Who Cares About Explainable AI? Why? When?



Who Cares About Explainable AI? Why? When?

How does a model work?

What is driving decisions?

Can I trust the model?

Data Scientist

Understand the model

Debug the model

Improve model performance

Business Owner

Understand the model

Evaluate model suitability

Accept model use

Risk Modeller

Challenge the model

Ensure model robustness

Approve model

Regulator

Check impact of model on consumers

Verify model reliability

Consumer

"What is the impact on me?"

"What actions can I take?"

Source: V. Belle and I. Papantonis: Principles and practice of explainable machine learning, *arXiv*, 2020. https://arxiv.org/abs/2009.11698

Who Cares About Explainable AI? Why? When?

RESOLVE CURIOSITY

UNDERSTAND CAUSES

FACILITATE LEARNING

ENHANCE TRUST AND RELIABILITY

Explainability
Generally, refers to making the decisions of AI systems **understandable** to humans, focusing on explanations that clarify why a decision was made.

Interpretability
Involves **understanding** the internal workings of the model and how it arrives at specific predictions

Transparency
Ensures that the AI system's processes are open and accessible, making the system's operations clear and **understandable** to users, thus building trust.

1. Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.
2. Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.” An International Journal on Information Fusion 58 (June): 82–115.
3. Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. 2021. “Explainable Artificial Intelligence: An Analytical Review.” Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery 11 (5).
4. Ali, Sajid, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. “Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence.” An International Journal on Information Fusion 99 (November): 101805.
5. Haque, Akm Bahalul, A. K. M. Najmul Islam, and Patrick Mikalef. 2023. “Explainable Artificial Intelligence (XAI) from a User Perspective: A Synthesis of Prior Literature and Problematizing Avenues for Future Research.” Technological Forecasting and Social Change 186 (January): 122120.

Goals of this tutorial

1. What is Human-Centered Explainable AI (HCXAI), and how does it differ from and complement traditional XAI approaches?
2. What are different explanation methods (e.g., feature-based, example-based) available to us?
3. How to design explanations that promote user engagement and control, and potentially prevent over-reliance or under-reliance on AI?
4. What are essential considerations and metrics for evaluating XAI systems?

Part 2: Taxonomy of Explanation Types

MODEL-AGNOSTIC VS. MODEL-SPECIFIC

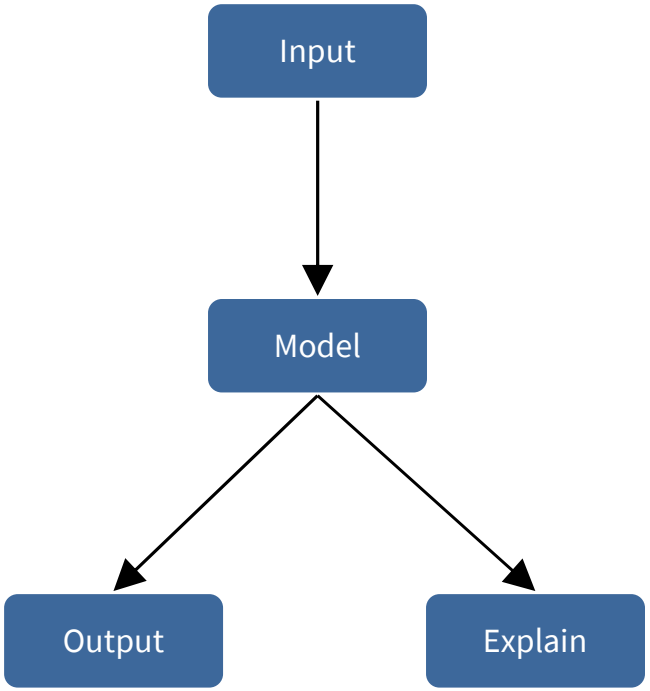
MODEL SPECIFIC

Inner workings of model used for XAI

MODEL AGNOSTIC

Uses only inputs & outputs for XAI

Applicable to any model with that interface



MODEL-AGNOSTIC VS. MODEL-SPECIFIC

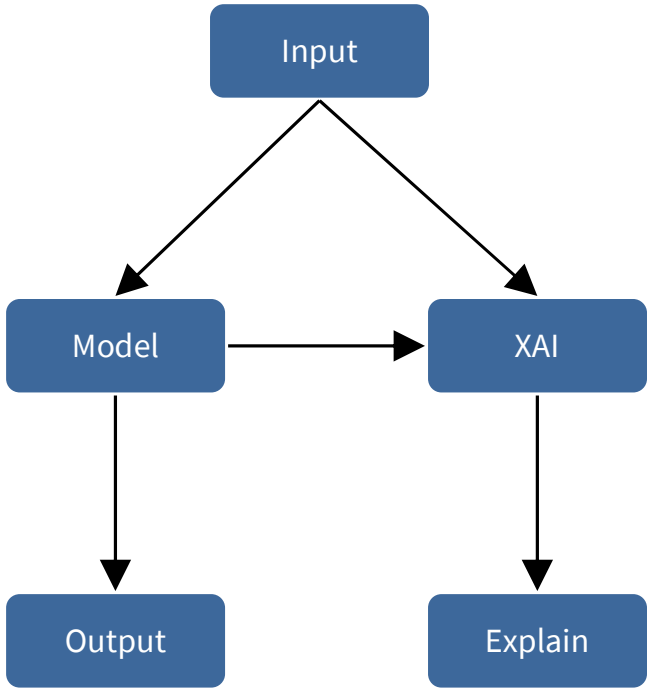
MODEL SPECIFIC

Inner workings of model used for XAI

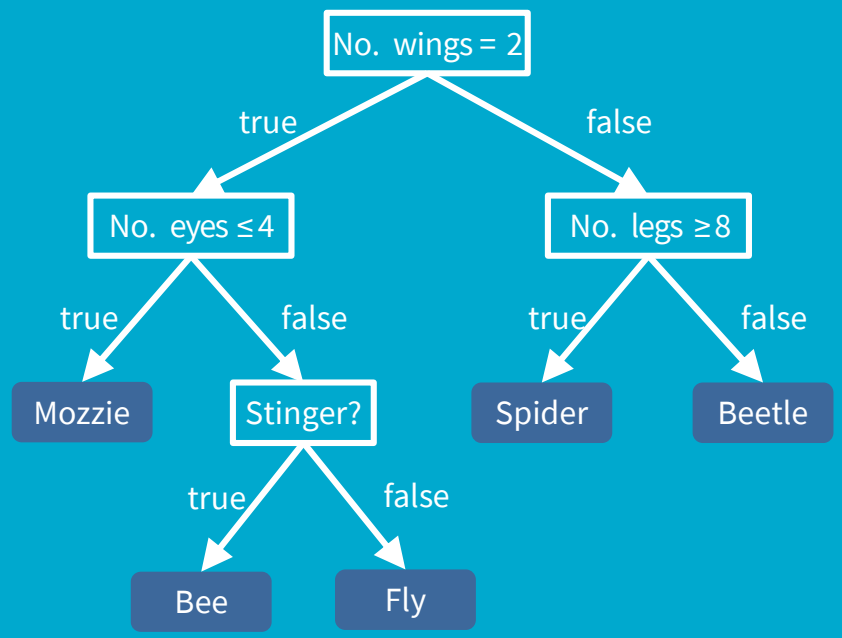
MODEL AGNOSTIC

Uses only inputs and outputs for XAI

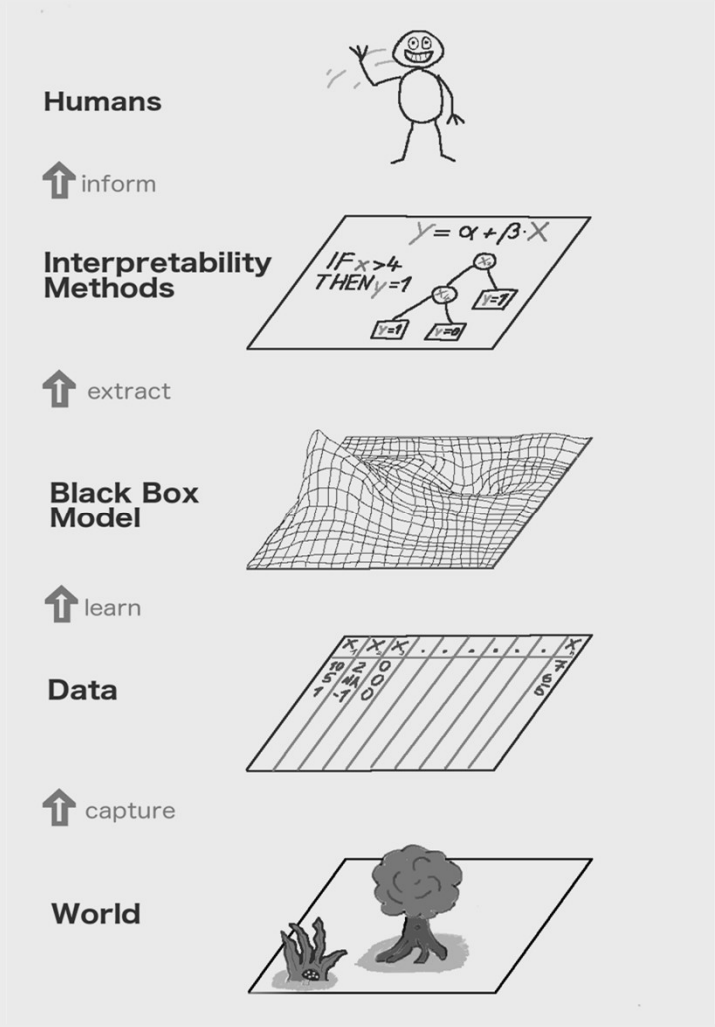
Applicable to any model with that interface



INTRINSIC VS. POST-HOC

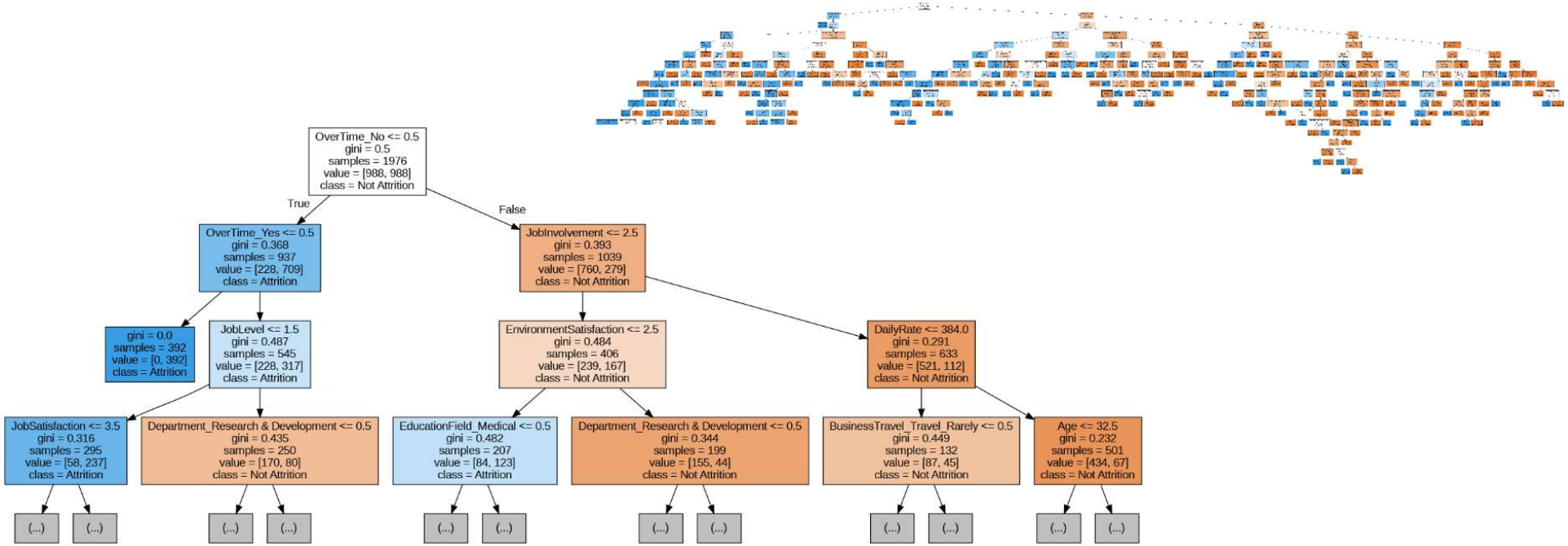


vs.



Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.

Employee Attrition Example: Training a Decision Tree



FOUNDATIONAL METHODS IN EXPLAINABLE AI

Exercise

Suppose you applied for a job via a company's job application portal. A few hours after your application, you get an email that says that your application was analysed by an automated system that decided not to offer you an interview.

What would you want the explanation to contain?

What would make it a good explanation for you?

Tools

Alibi (SeldonIO):
Alibi is a Python library that offers methods for instance-based explanations, such as counterfactuals and anchors

Captum (PyTorch):
Captum is an open-source, model interpretability library built on PyTorch.

DALEX (ModelOriented):
DALEX is a model-agnostic explainability package designed for both R and Python.

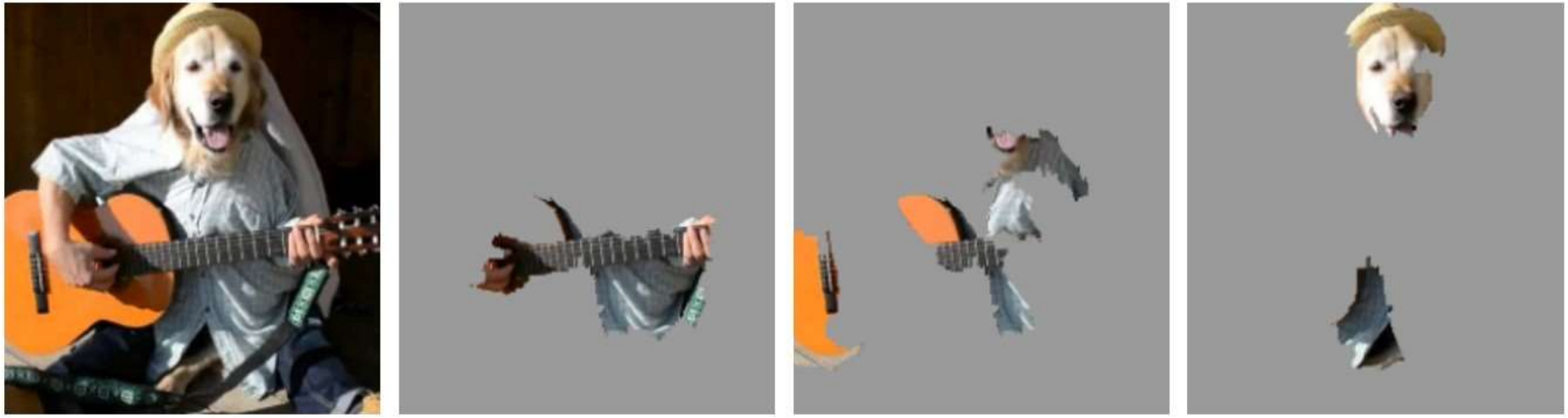
AI Explainability 360 (AIX360) (Trusted-AI):
AIX360 is an open-source toolkit developed by IBM for explainability across a wide range of models.
It provides both interpretable models (like rule-based models) and post-hoc explanations for black-box models

InterpretML (interpretml):
InterpretML is a Microsoft-developed. It offers a combination of glass-box (interpretable-by-design) and black-box (post-hoc explainability) models

ATTRIBUTION-BASED EXPLANATIONS

Source: T. Miller: "But why?" Understanding explainable artificial intelligence. *XRDS 25, 3* (Spring 2019), 20–25. https://doi.org/10.1145/3313107

ATTRIBUTION-BASED EXPLANATIONS



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Source: Ribeiro et al.: Why should I trust you?: Explaining the predictions of any classifier. In SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

EXAMPLE-BASED EXPLANATION: PROTOYPES

Source: Kim et al.: Examples are not enough, learn to criticize! Criticism for interpretability. In NeurIPS. 2016.

EXAMPLE-BASED: CONTRASTIVE EXPLANATION

THE DIFFERENCE CONDITION

"The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case"

DENIS HILTON: CONVERSATIONAL PROCESSES AND CAUSAL EXPLANATION, PSYCHOLOGICAL BULLETIN. 107(11):65-81, (1990)

EXAMPLE-BASED: CONTRASTIVE EXPLANATION

THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY RATHER THAN A BEETLE?

EXAMPLE-BASED: CONTRASTIVE EXPLANATION

THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY RATHER THAN A BEETLE?

EXAMPLE-BASED: CONTRASTIVE EXPLANATION

THE DIFFERENCE CONDITION

Type	Legs	Stinger	Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

WHY IS IT A FLY RATHER THAN A BEETLE?

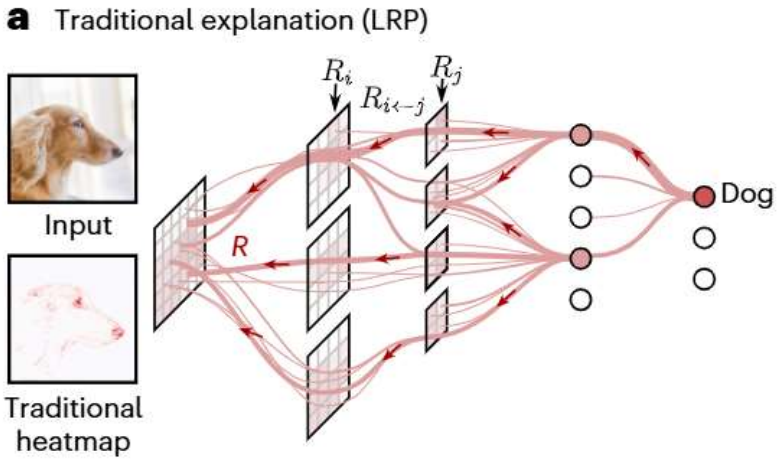
RULED-BASED EXPLANATION

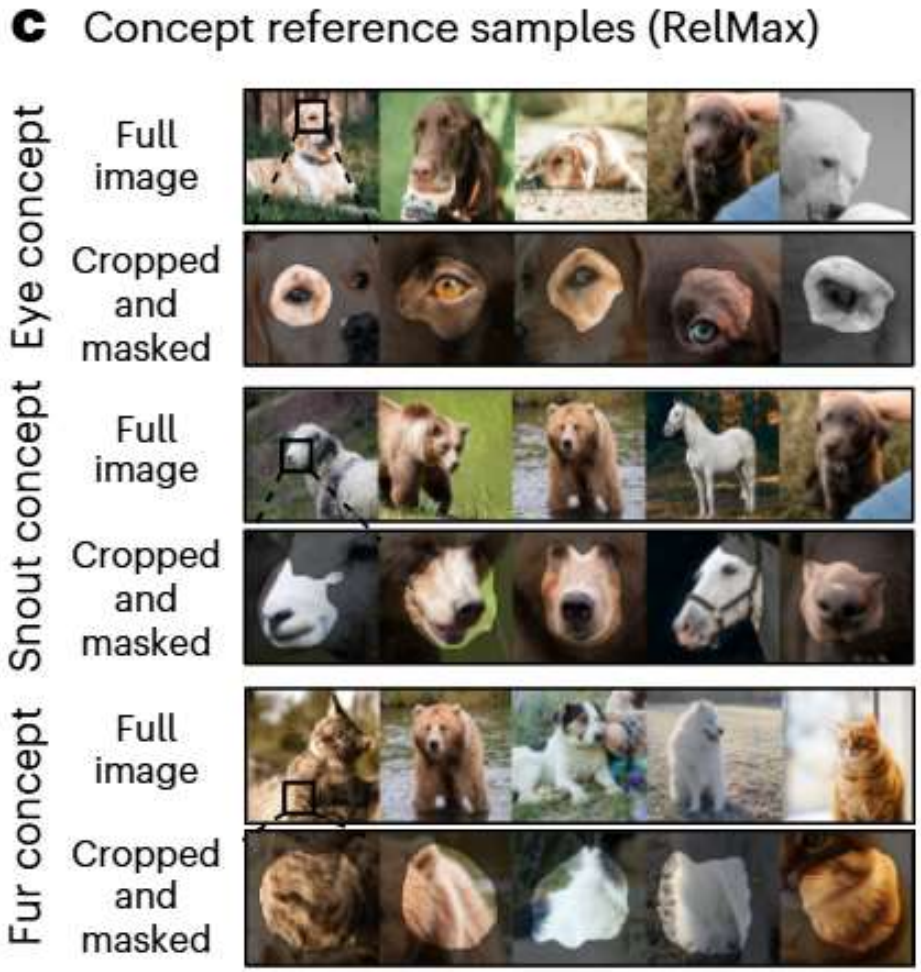
EXTRACT RULES POST-HOC

LEARN INTERPRETABLE RULES DIRECTLY

```
if respiratory_illness = Yes
    and smoker = Yes
    and age >= 50
then lung_cancer
elif risk_lung_cancer = Yes
    and blood_pressure >= 0.3
then lung_cancer
elif risk_depression = Yes
    and past_depression = Yes
then depression
elif BMI >= 0.3
    and insurance = None
    and blood_pressure >= 0.2
then depression
elif smoker = Yes
    and BMI >= 0.2
    and age >= 60
then diabetes
elif risk_diabetes = Yes
    and BMI >= 0.4
    and prob_infections >= 0.2
then diabetes
Else Null
```

CONCEPT-BASED





Poeta, Eleonora, et al. "Concept-based explainable artificial intelligence: A survey." *arXiv preprint arXiv:2312.12936* (2023).

Recap

What do you recall about the taxonomy of explanations?

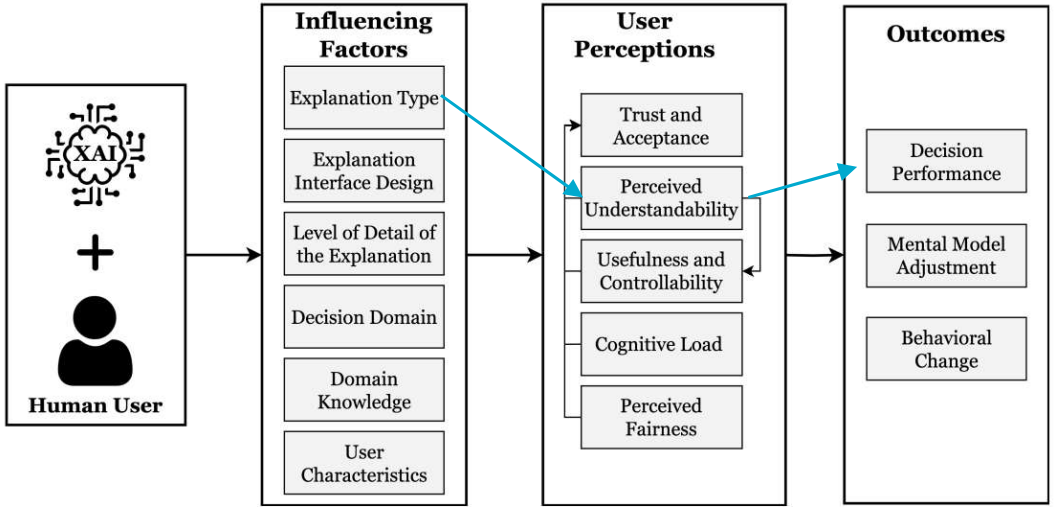
What are some types of explanations that you know (now ☺)?

Part 3: Towards HCXAI
A few principles

Understanding What Influences Human-XAI Team Outcomes

Perceptions are shaped by multiple, interdependent factors.

These perceptions influence user behaviour and the effectiveness of AI systems in real-world tasks.



Source: Schauer, Andreas. 2024. "In the Eyes of the User: A Systematic Literature Review of User Perceptions in Human-XAI Interaction." In ICIS 2024 Proceedings. aisel.aisnet.org.

Alignment with Human Cognitive Processes

Claim: without explicit support for how humans logically approach problem-solving, XAI can lead to poor outcomes.

E.g. **user's rationalising the AI's predictions** with critically evaluating these; **System 1 thinking**, **relying on intuition or heuristics**, which are fast but prone to biases like confirmation bias

Hypothetico-deductive reasoning (H-D reasoning)

Takeaways:

Foster independent thinking, encouraging users to critically evaluate AI suggestions.

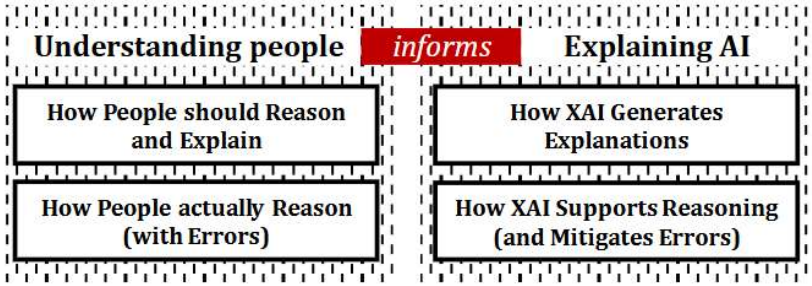
Use different types of explanations to **support iterative reasoning processes.**

Designing Theory-Driven User-Centric Explainable AI

Authors: Danding Wang, Qian Yang, Ashraf Abdul, Brian Y. Lim Authors Info & Claims

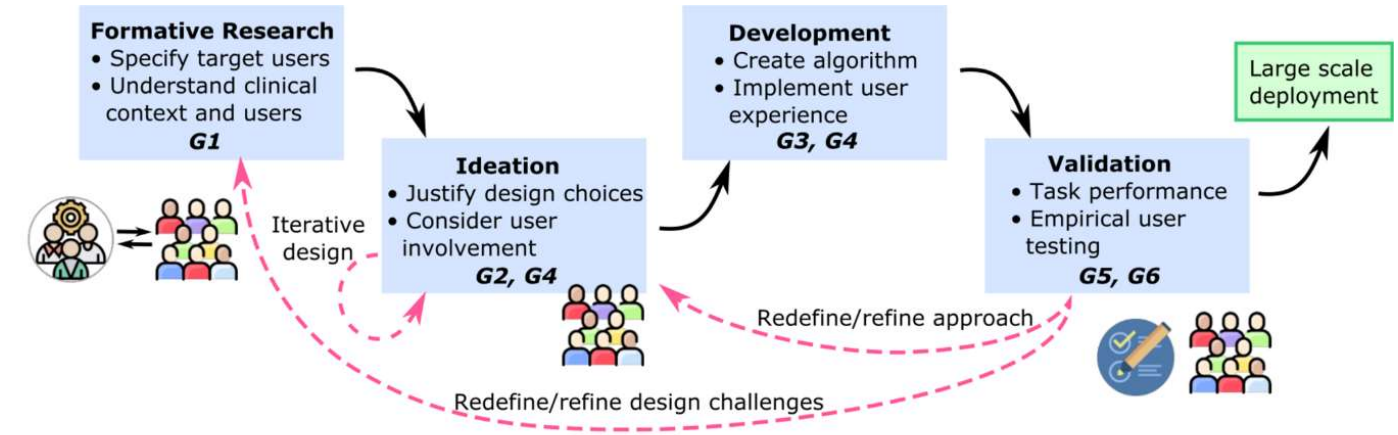
CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems • Paper No.: 601, Pages 1 - 15
https://doi.org/10.1145/3290605.3300831

Published: 02 May 2019 Publication History Check for updates



Wang, Danding, et al. "Designing theory-driven user-centric explainable AI." *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019.

Iterative and User-Centred Design



Incorporation, Task, Target ***G1***: Identify target users within the clinical context, their constraints and requirements given the task.
Prior, Interpretability ***G2***: Specify design choices and justify the assumptions that motivate them.
Interpretability ***G3***: Ensure that the developed computational technique is consistent with the assumptions in G2.
Reporting ***G4***: Determine efficacy of the presentation and interaction paradigms of the user interface.
Reporting ***G5***: Report metrics and experimental design to quantify on-task performance.
Reporting ***G6***: Evaluate the correctness and effectiveness in achieving the desired human factors engineering goals.

Chen, Haomin, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. 2022. “Explainable Medical Imaging AI Needs Human-Centered Design: Guidelines and Evidence from a Systematic Review.” Npj Digital Medicine 5 (1): 156.

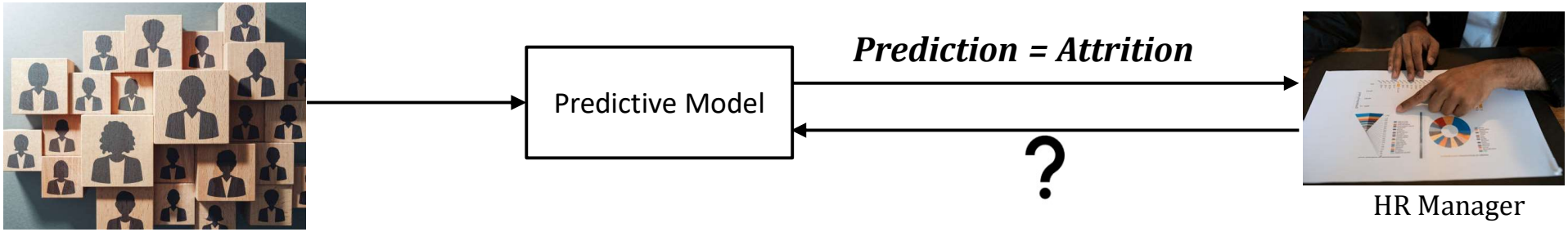
Exercise [3 mins]:

Employee attrition is a major challenge for organisations. Understanding why employees leave and taking proactive steps to retain them is crucial for building a stable and engaged workforce.

Suppose you work as a HR specialists and a predictive model produced by the organisation's platform is used for identifying employees at risk of leaving the organisation.

What questions would you like to ask the predictive model?

Get into pairs, one playing Manager and another Designer. The designer's goal is to identify the questions that the manager wants to ask the model.



Does these somewhat capture what you were thinking?

Why is this specific employee predicted to leave?

Why this prediction instead of another?

What would change the model's prediction?

How reliable are these predictions?

What data is the model using to make this decision?

What comparisons or contrasts are available?

How accurate is the model overall, and what is its error rate?

What are the common patterns in predictions?

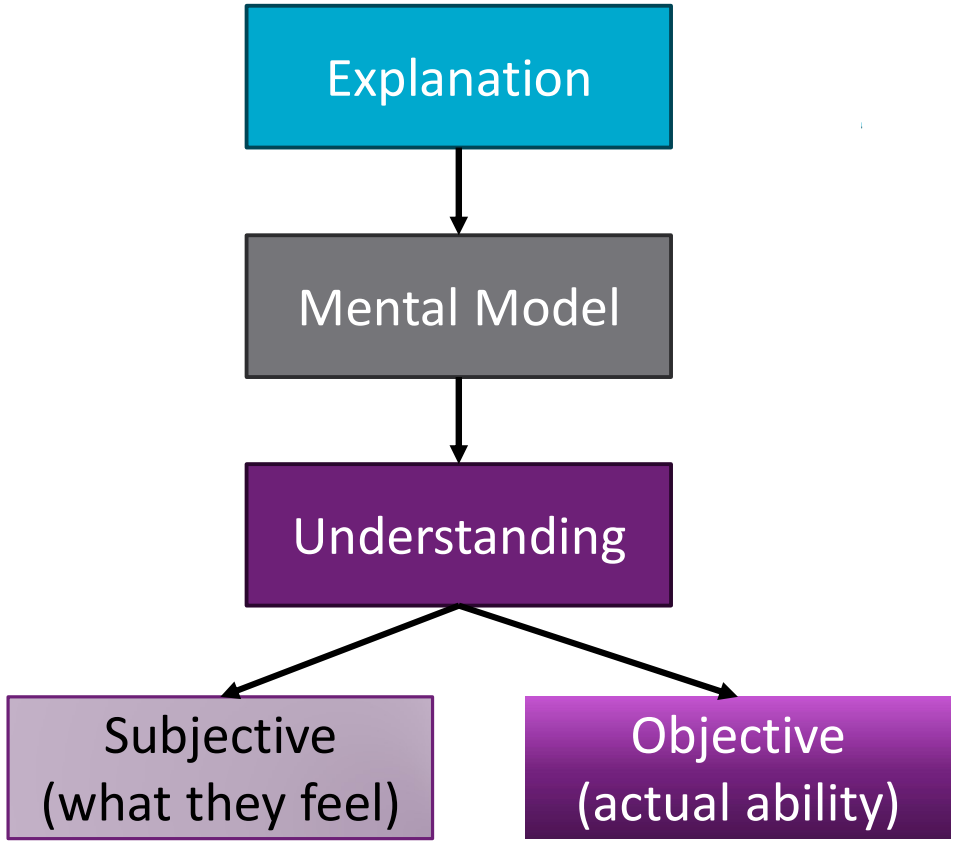
Understandability

Understandable explanations fosters trust, accountability, and informed user interaction.

I Think I Get Your Point, AI!
The Illusion of Explanatory Depth in Explainable AI

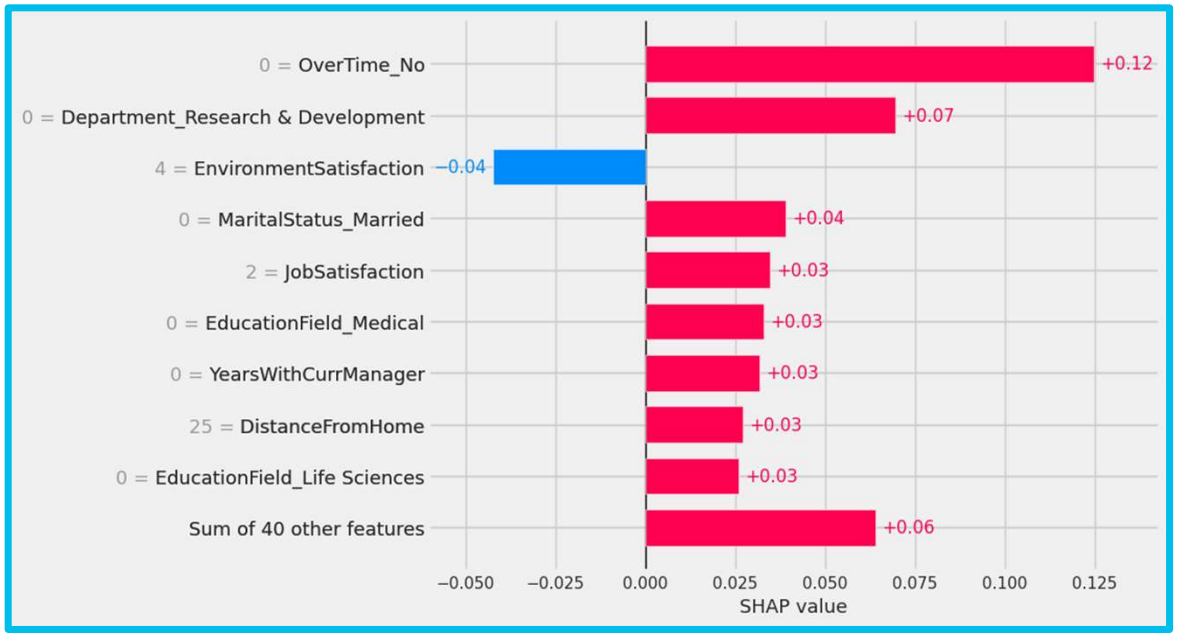
Forming (Global) Rationales from Local Explanations

Interactive Explanations



Exercise

Get back into pairs. Discuss the pros and cons of the feature based and CF explanations in helping the manager understand the model's prediction.



Counterfactuals

CF	Feature	Current	Counterfactual
1	Overtime	Yes	No
2	Job Satisfaction	2	4

dice-ml 0.11

```
pip install dice-ml
```

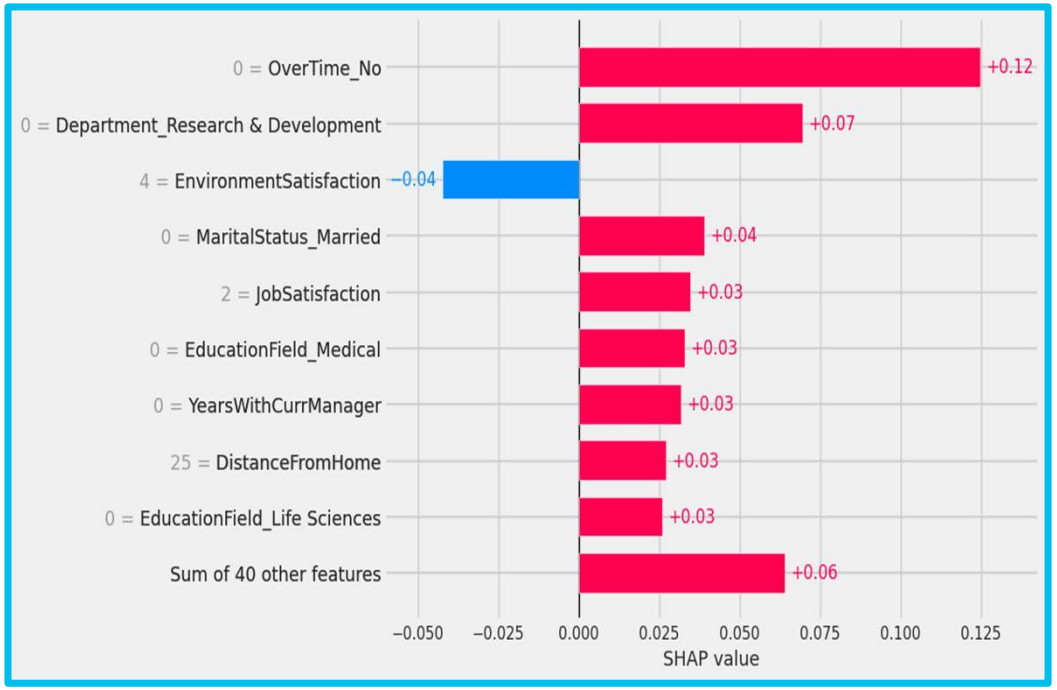
SHAP (SHapley Additive exPlanations)

Breaks down the prediction into contributions from each feature

Shows how much each feature **increased or decreased** the prediction

This SHAP bar plot explains **why the model made a specific prediction** by visualising:

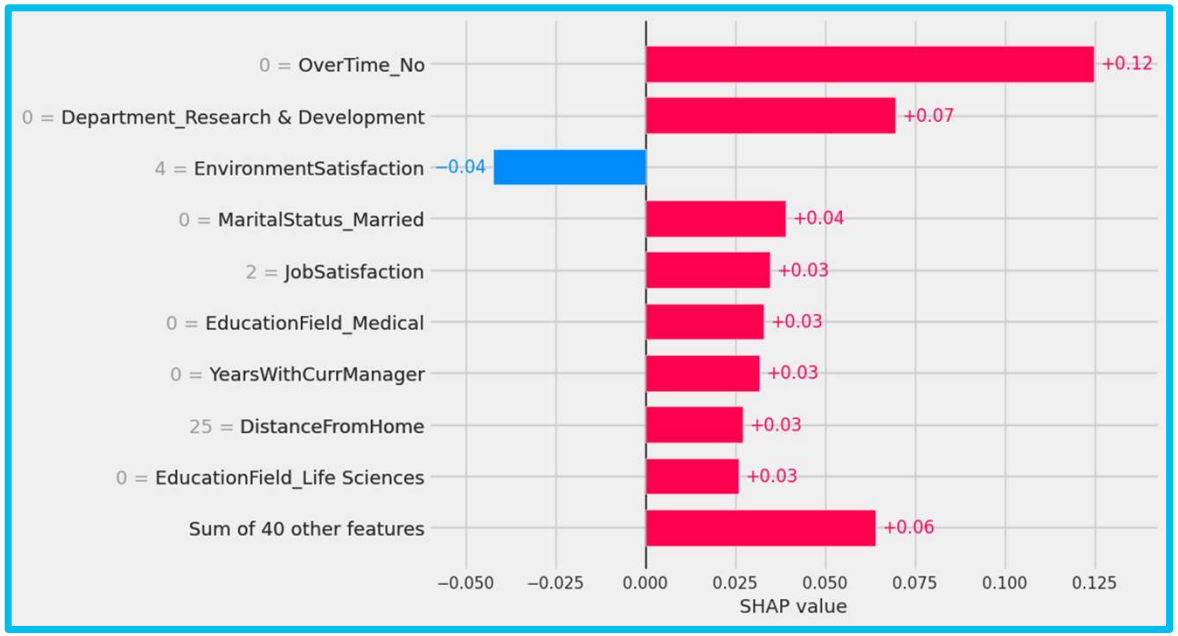
- **Red Bars:** Features that increased the prediction (positive contributions).
- **Blue Bars:** Features that decreased the prediction (negative contributions).
- **Bar Length:** The size of the contribution (e.g., "OverTime_No" has the largest positive impact).



Scott, M., and Lee Su-In. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017): 4765-4774.

Exercise

Get back into pairs. Discuss the pros and cons of the feature based and CF explanations in helping the manager understand the model's prediction.



Counterfactuals

CF	Feature	Current	Counterfactual
1	Overtime	Yes	No
2	Job Satisfaction	2	4

dice-ml 0.11

```
pip install dice-ml
```

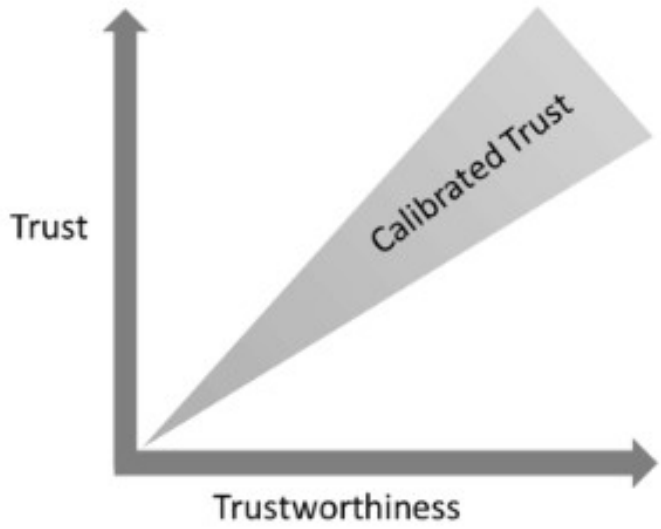
Promote Trust Calibration

integrity and reliability.

The confidence one element has in another, that the second element will behave as expected. (NIST)

Calibrated trust prevents over-trust or under-trust, enabling users to rely on AI appropriately.

XAI?

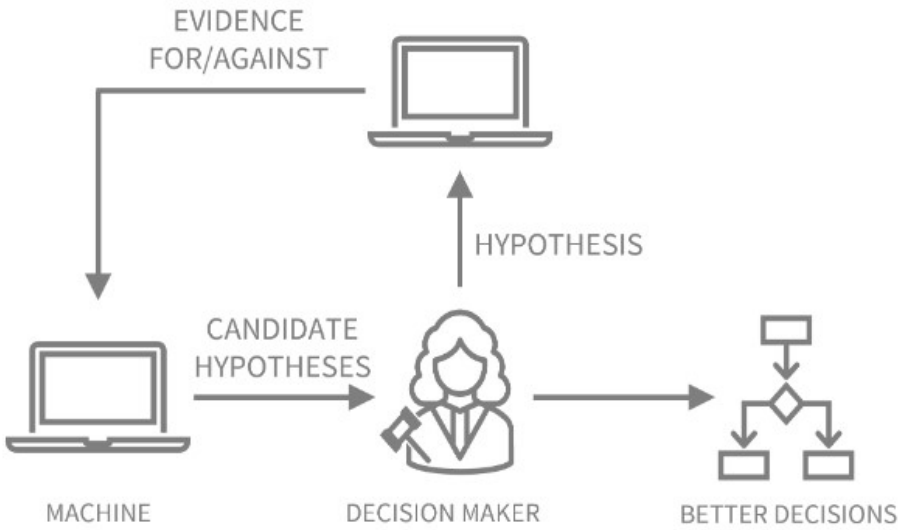


Source: McDermott et al. (2019)

1. Kaur, Davinder, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. "Trustworthy Artificial Intelligence: A Review." ACM Comput. Surv., 39, 55 (2): 1–38.
2. Jacovi, Alon, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 624–35. FAccT '21. New York, NY, USA: Association for Computing Machinery.
3. Miller, Tim. 2022. "Are We Measuring Trust Correctly in Explainability, Interpretability, and Transparency Research?" arXiv [Cs.HC]. arXiv. http://arxiv.org/abs/2209.00651.
4. McDermott, Patricia L., and Ronna N. Ten Brink. 2019. "Practical Guidance for Evaluating Calibrated Trust." Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting 63 (1): 362–66.

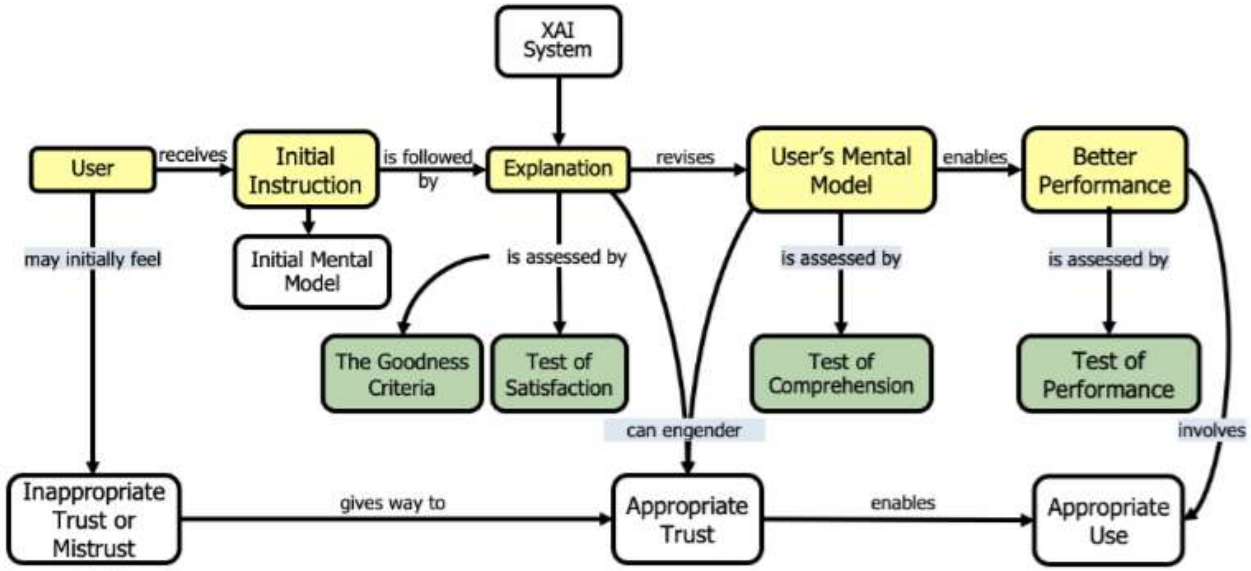
Preservation of Human Agency and Control

Ensure that users maintain control over decision-making processes, with AI acting as a support tool rather than an authoritative entity.

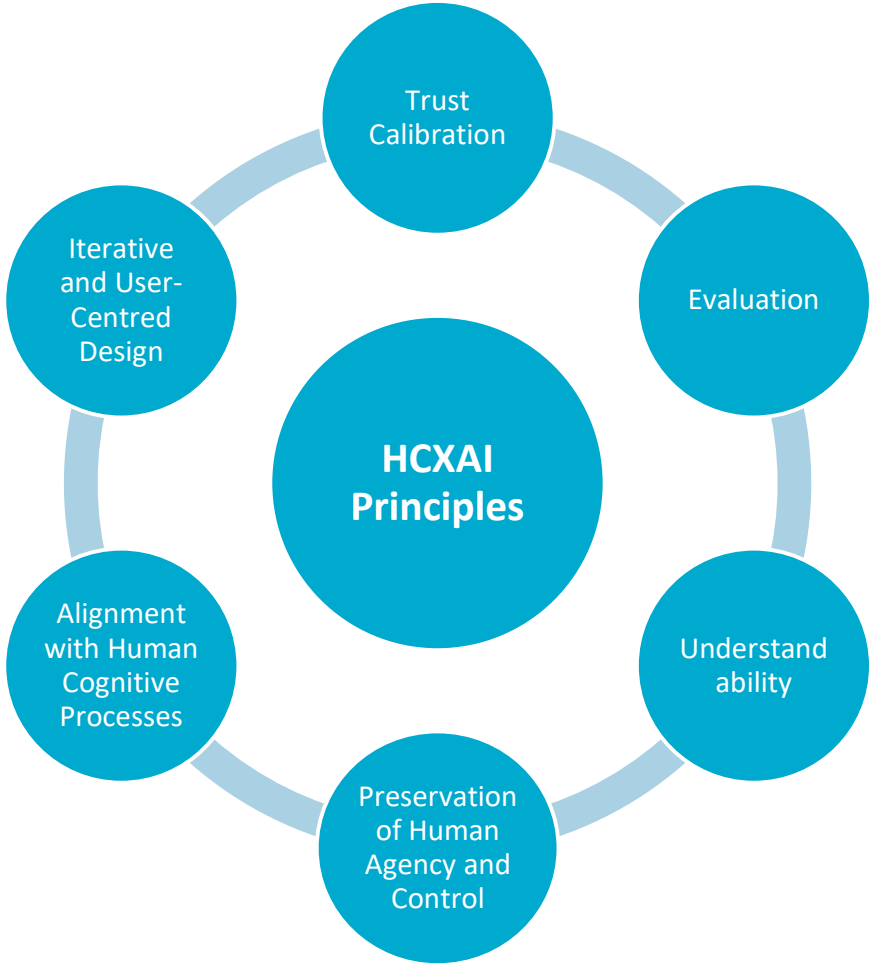


Miller, Tim. 2023. “Explainable AI Is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI.” In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 333–42. FAccT ’23. New York, NY, USA: Association for Computing Machinery.

Evaluation of XAI Systems



1. Gunning, David, and D. Aha. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program." AI Magazine 40 (2): 44–58.
2. Gunning, David, E. Vorm, Jennifer Yunyan Wang, and Matt Turek. 2021. "DARPA's Explainable AI (XAI) Program: A Retrospective." Applied AI Letters, November.
3. Hoffman, Robert R., Shane Mueller, Gary Klein, and Jordan Litman. 2023. "Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance." Frontiers in Computer Science 5.



1. Schauer, Andreas. 2024. "In the Eyes of the User: A Systematic Literature Review of User Perceptions in Human-XAI Interaction." In ICIS 2024 Proceedings. aisel.aisnet.org. https://aisel.aisnet.org/icis2024/lit_review/lit_review/12/.

2. Miller, Tim. 2023. "Explainable AI Is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI." In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 333–42. FAccT '23. New York, NY, USA: Association for Computing Machinery.

3. Gunning, David, and D. Aha. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program." AI Magazine 40 (2): 44–58.

4. Gunning, David, E. Vorm, Jennifer Yunyan Wang, and Matt Turek. 2021. "DARPA's Explainable AI (XAI) Program: A Retrospective." Applied AI Letters, November. https://doi.org/10.22541/au.163699841.19031727/v1.

5. Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. 2021. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems." ACM Trans. Interact. Intell. Syst., 24, 11 (3–4): 1–45.

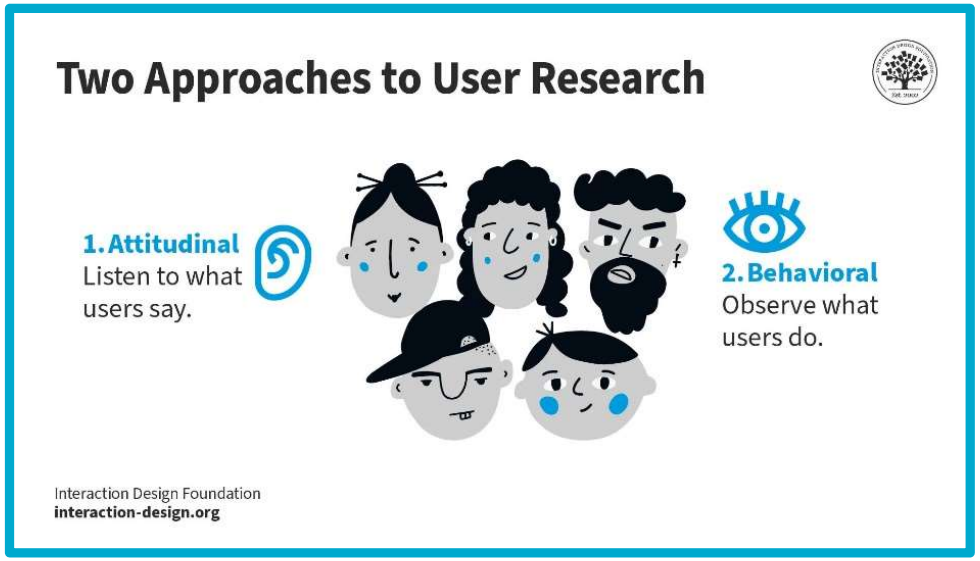
6. Kim, Sunnie S. Y., Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "'Help Me Help the AI': Understanding How Explainability Can Support Human-AI Interaction." In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17. CHI '23 250. New York, NY, USA: Association for Computing Machinery.

BREAK (15-20 mins)

Part 4: Designing for User Engagement and Control

Start by understanding user needs

1. Interviews
2. Surveys and Questionnaires
3. Focus Groups
4. Ethnographic Studies
5. Usability Testing
6. Cognitive Task Analysis



Source: https://www.interaction-design.org/literature/topics/user-needs

Start by understanding user needs

Interviews: Each participant engaged in a **one-hour Zoom interview**

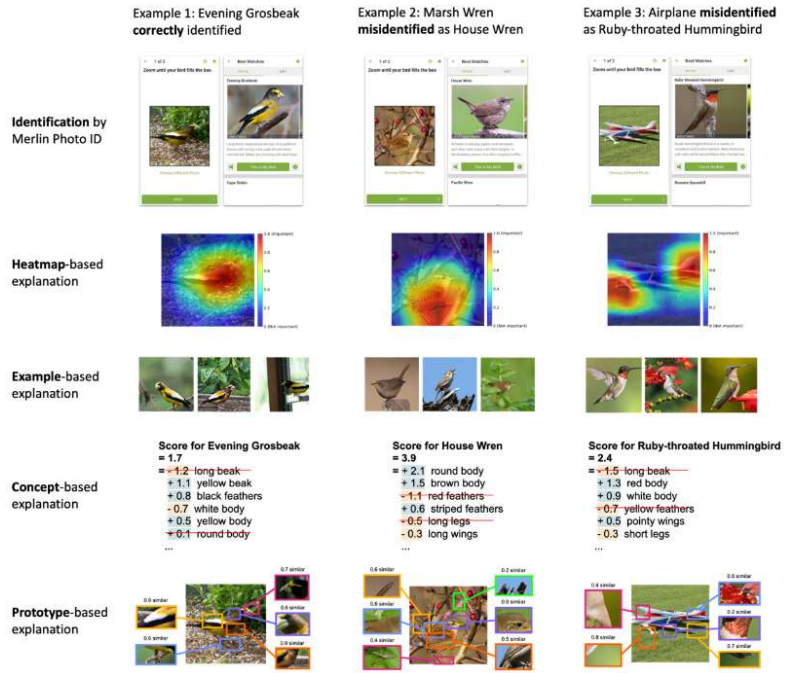
Survey: Participants rated their **curiosity and understanding**; e.g., "Why did the AI make this decision?").

Data Collection and Analysis:

- Interview transcripts and survey responses.
- **Thematic analysis** to identify **conceptual themes**.

Mock-Up Testing: Four XAI mock-ups/prototypes used to **gather feedback**.

Iterative Refinement: **Group discussions** among the researchers



Kim, Sunnie S. Y., Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "'Help Me Help the AI': Understanding How Explainability Can Support Human-AI Interaction." In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17. CHI '23 250. New York, NY, USA: Association for Computing Machinery.

Start by understanding user needs

A question bank could be used to design **initial list of questions** that users may ask

Liao, Q. Vera, Daniel Gruen, and Sarah Miller. "Questioning the AI: informing design practices for explainable AI user experiences." Proceedings of the 2020 CHI conference on human factors in computing systems. 2020.

Input

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

Give control to the users

AI Recommendations
Assumption: Decision makers will carefully consider recommendations.
Reality: Empirical evidence shows this is not the case.
Reason: Lack of trust?

Explainable AI
Assumption: Distrust can be mitigated by with reasons or explanations.
Reality: Does not help decision makers determine whether machine or themselves are right when they disagree.
Reason: Lack of information to resolve uncertainty.

Miller, Tim. "Explainable AI is dead, long live explainable ai! hypothesis-driven decision support using evaluative AI." Proceedings of the 2023 ACM conference on fairness, accountability, and transparency. 2023.

Exercise: Meet Bluster and Prudence

RECOMMENDATION-DRIVEN
EXPLAINABLE AI

HYPOTHESIS-DRIVEN
EVALUATIVE AI

Give control to the users

Cognitive forcing functions → interventions to encourage analytical thinking

CF could potentially reduce overreliance

Need to focus on system usability – users may not prefer the best system!

Adaptively apply CF based on task requirements and user characteristics

- CF benefits more engaged users!

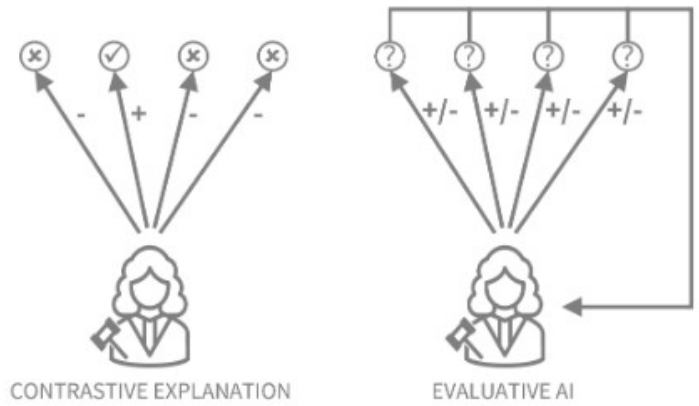
Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making." Proceedings of the ACM on Human-computer Interaction 5.CSCW1 (2021): 1-21.

Presenting evidence for user hypotheses

Explainable AI is Dead, Long Live Explainable AI!

Hypothesis-driven Decision Support using Evaluative AI

Tim Miller
tmiller@unimelb.edu.au
The University of Melbourne
Melbourne, VIC, Australia



From human-in-the-loop to machine-in-the-loop

AI is a 'sounding board', not a decision maker

Implement features that facilitate hypothesis testing

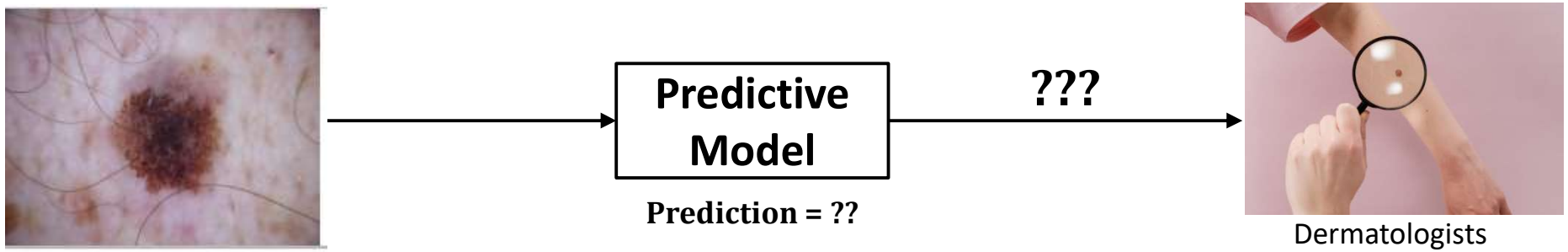
Filter out unlikely hypotheses or rank options

Miller, Tim. "Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI." Proceedings of the 2023 ACM conference on fairness, accountability, and transparency. 2023.

Case Study on Skin Cancer

Suppose you are working on developing hypothesis-driven XAI for a tool that helps dermatologists detecting skin cancer. For an image that is uploaded by a dermatologists, you have to determine what explainability information to show to them.

What could hypothesis driven approach look like?



Presenting evidence for user hypotheses

Towards the New XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence

Thao Le[a,*], Tim Miller[b], Liz Sonenberg[a] and Ronal Singh[c]

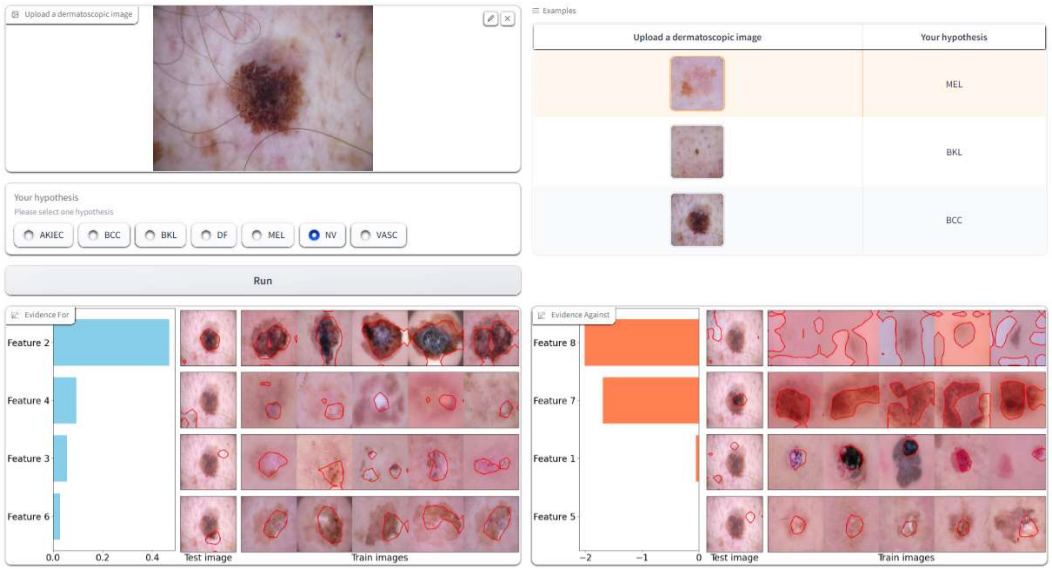
[a]School of Computing and Information Systems, The University of Melbourne
[b]School of Electrical Engineering and Computer Science, The University of Queensland
[c]CSIRO's Data61

Visual Evaluative AI: A Hypothesis-Driven Tool with Concept-Based Explanations and Weight of Evidence

Thao Le[a,*], Tim Miller[b], Ruihan Zhang[a], Liz Sonenberg[a] and Ronal Singh[c]

[a]School of Computing and Information Systems, The University of Melbourne, Australia
[b]School of Electrical Engineering and Computer Science, The University of Queensland, Australia
[c]CSIRO/Data61, Australia

https://github.com/thaole25/EvaluativeAI



Le, T., Miller, T., Singh, R., & Sonenberg, L. (2024). Towards the new XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence. ArXiv, abs/2402.01292. ***ACCEPTED AT ECAI 2024***

Le, T., Miller, T., Zhang, R., Sonenberg, L., & Singh, R. (2024). Visual Evaluative AI: A Hypothesis-Driven Tool with Concept-Based Explanations and Weight of Evidence. *arXiv preprint arXiv:2407.04710*.

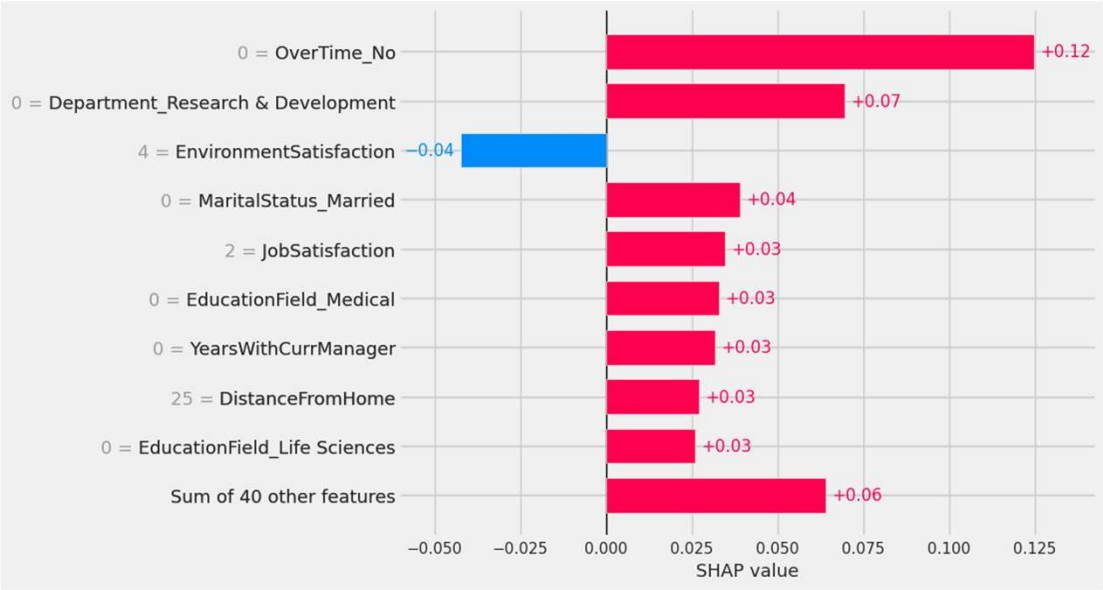
What would you do differently on your next explainability project?

Part 5: Evaluating XAI and User Studies

Exercise: Employee Attrition

Suppose you have determined that a one the of XAI methods that may be relevant is Feature-based explanation.

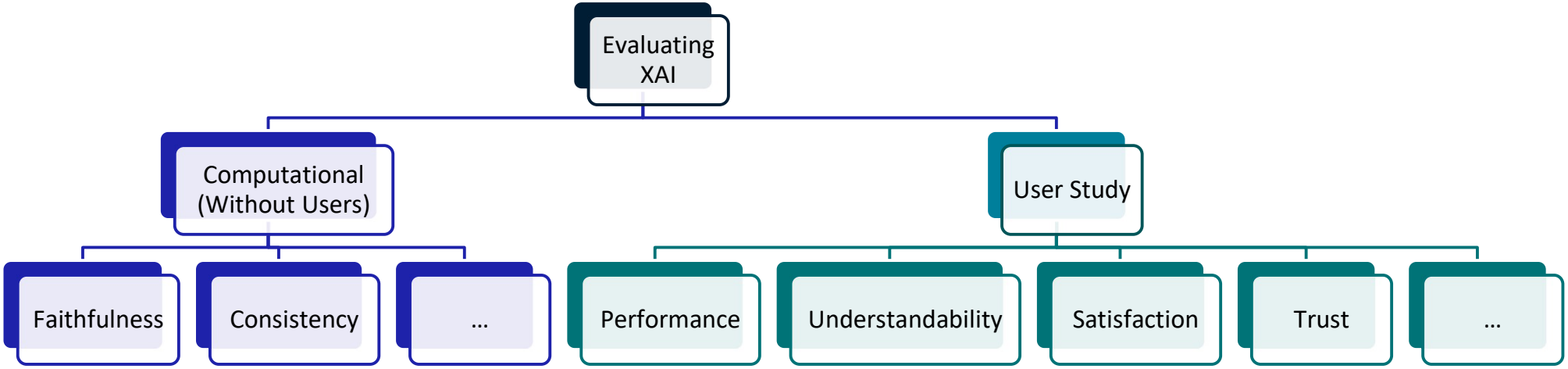
How would you go about evaluating these?



Assessing explainable AI

Pawlicki, Marek, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'Antonio, Rafał Kozik, and Michał Choraś. 2024. "Evaluating the Necessity of the Multiple Metrics for Assessing Explainable AI: A Critical Examination." Neurocomputing 602 (128282): 128282.

Computational and Human Evaluations



1. Hoffman, Robert R., Shane Mueller, Gary Klein, and Jordan Litman. 2023. "Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance." Frontiers in Computer Science 5. https://doi.org/10.3389/fcomp.2023.1096257.
2. Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI." ACM Computing Surveys 55 (13s): 1–42.
3. Pawlicki, Marek, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'Antonio, Rafał Kozik, and Michał Choraś. 2024. "Evaluating the Necessity of the Multiple Metrics for Assessing Explainable AI: A Critical Examination." Neurocomputing 602 (128282): 128282.

	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is.
		Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is.
		Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets.
		Key idea: Answers "why not?" or "what if?" questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are.
		Key idea: Human-understandable concepts in the explanation
Presentation	**Compactness**	Describes the size of the explanation.
		Key idea: Less is more
	Composition	Describes the presentation format and organization of the explanation.
		Key idea: *How* something is explained
	Confidence	Describes the presence and accuracy of probability information in the explanation.
		Key idea: Confidence measure of the explanation or model output
User	**Context**	Describes how relevant the explanation is to the user and their needs.
		Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs.
		Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user.
		Key idea: Can the user influence the explanation?

Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI." ACM Computing Surveys 55 (13s): 1–42.

Packages to assist you!

BEExAI is an open-source tool to quantitatively evaluate, compare and benchmark post-hoc explainable AI - features attribution - methods on multiples Machine Learning models with an end-to-end pipeline.

- https://github.com/SquareResearchCenter-AI/BEExAI

Quantus is an eXplainable AI toolkit for responsible evaluation of neural network explanations.

- https://quantus.readthedocs.io/en/latest/

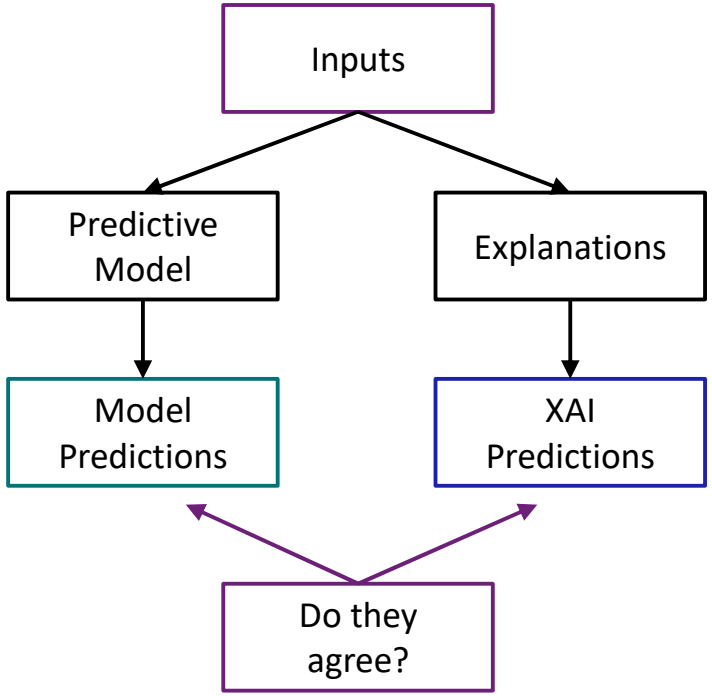
OpenXAI is a general-purpose lightweight library that provides a comprehensive list of functions to systematically evaluate the reliability of post hoc explanation methods.

- https://github.com/AI4LIFE-GROUP/OpenXAI

XAI-Bench is a library for benchmarking feature attribution techniques using synthetic data

- https://github.com/abacusai/xai-bench

Faithfulness



Agreement between the output of the predictive model and the explanation when applied to the input, how well the explanations mimic the output of the predictive model if used to make predictions

Faithfulness – How?

CORRECTNESS (Section 6.1)	
Model Parameter Randomization Check—*Feature importance, Heatmap, Localization* Randomly perturb the internals of the predictive model and check that the explanation changes.	[3, 145, 209, 232, 284]
Explanation Randomization Check—*Feature importance, Heatmap* Randomly perturb the explanation (which is built into the predictive model) and check that the output of the predictive model changes.	[165, 232]
White Box Check—*Feature importance, Decision Rules, White-box model, Localization* Apply the explanation method to an interpretable white box model and check the correspondence of the explanation with the white box reasoning.	[54, 114, 117, 136, 202, 204, 307]
Controlled Synthetic Data Check *Feature importance, Heatmap, Prototypes, Localization, White-box model, Graph* Controlled experiment: Create a synthetic dataset such that the predictive model should follow a particular reasoning, known a priori (important: checking this assumption by, e.g., reporting almost-perfect accuracy). Evaluate whether the explanation shows the same reasoning as the data generation process.	[3, 40, 73, 104, 110, 125, 151, 156, 158, 193, 196, 197, 211, 237, 242, 250, 260, 287]
Single Deletion—*Feature importance, Heatmap* Delete, mask or perturb a single feature in the input and evaluate the change in output of the predictive model. Measure correlation with explanation's importance score.	[8, 17, 42, 49, 73, 179, 218, 219, 221, 299, 299]
Incremental Deletion (or Incremental Addition)—*Feature importance, Heatmap* One by one delete (or perturb) or add features to the input, based on explanation's order, and measure for each new input the change in output of the predictive model. Report average change in log-odds score, AUC, steepness of curve or number of features needed for a different decision. Compare with random ranking or other baselines.	[25, 34, 38, 41, 70, 75, 79, 88, 90, 95, 102, 106, 110, 122, 129, 150, 155, 158, 165, 173, 181, 200, 202, 216, 221, 231, 261, 271, 278, 283, 284]

Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI." ACM Computing Surveys 55 (13s): 1–42.

Faithfulness

An explanation method should also verify the property of Faithfulness. It means that attributions should be aligned with the important features for the model prediction. Most faithfulness metrics quantifies this aspect by perturbating a sample relatively to feature importances and compare the way the prediction of the model changes.

Faithfulnesss correlation

Faithfulness Correlation (original paper) computes the correlation between the sum of the attributions of x_S and the difference in model prediction, x_S being an input with a subset of indices S features replaced by a baseline value.

With a predictor f, an explanation function g, an instance x and a subset of features S, faithfulness correlation is defined by:

$$\mu_F(f, g; x) = \underset{S \in \binom{|d|}{|S|}}{corr} \left(\sum_{i \in S} g(f, x)_i, f(x) - f(x_S) \right)$$

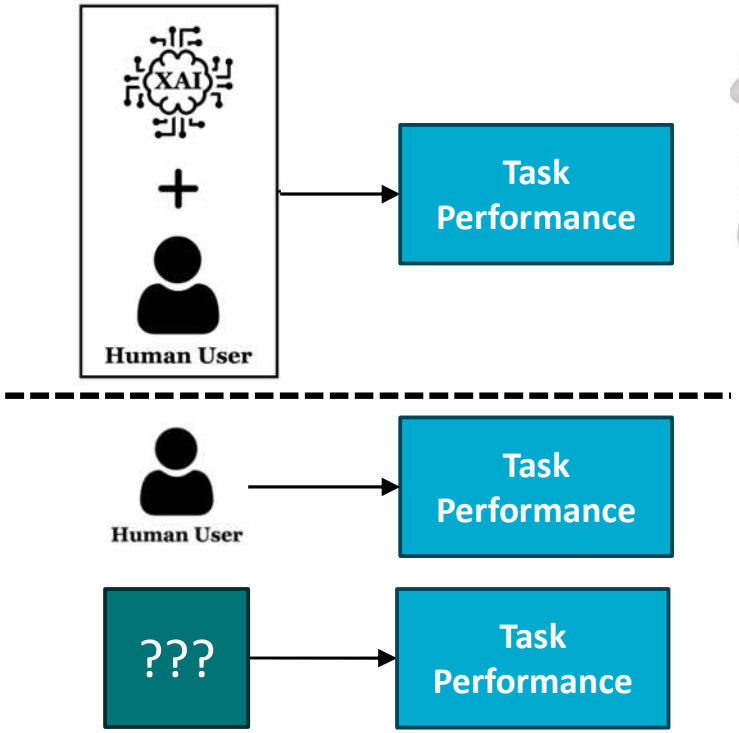
All subsets of $\binom{|d|}{|S|}$ are not explored entirely and the size of subsets |S| is fixed. By definition, the values of the correlation lie in [-1,1] with the ideal case being $\mu_F = 1$.

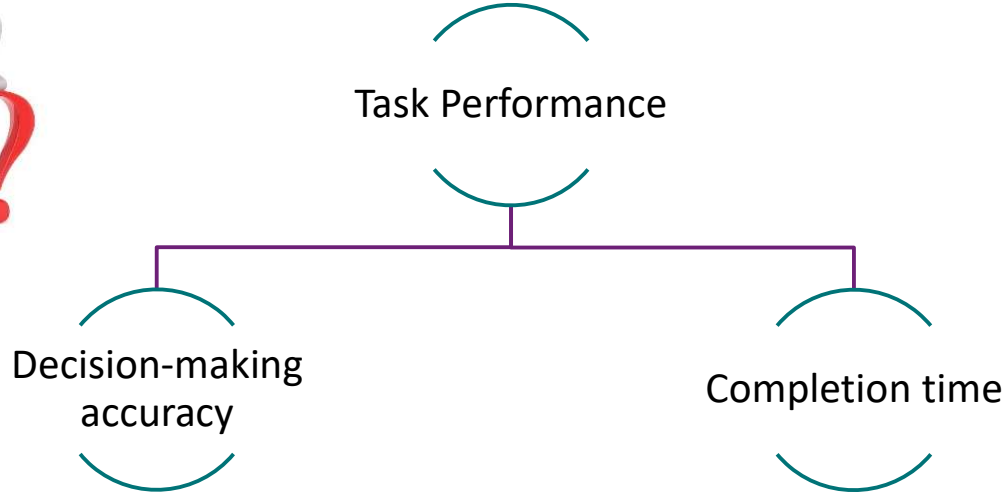
In our experiments, we use a percentage of the total number of features for the subset size |S| with a default value of 20% which we found to yield sufficiently discriminative results between different explainability methods. We choose to repeat this process for 20 iterations which we consider a good trade-off between computation time and quality of the approximation.

https://beexai.readthedocs.io/en/latest/metrics.html#robustness

User Related

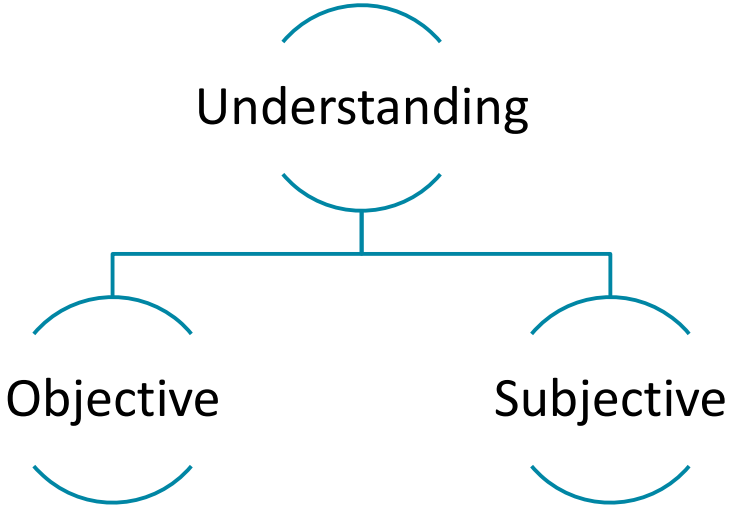
Task Performance



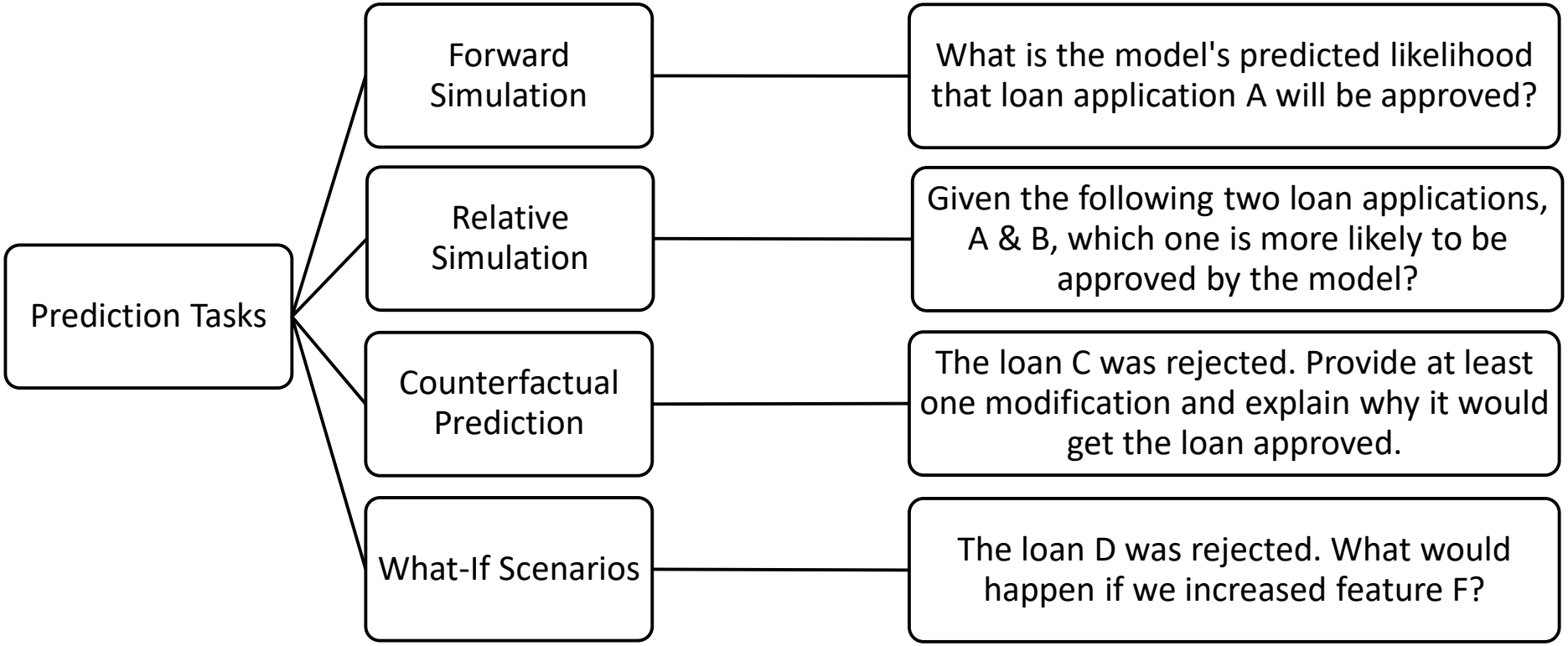


Understanding

Understanding in XAI is a user's ability to form a **mental model of a predictive model**, enabling comprehension, prediction, and insight into its decision-making process.



Understanding: Objective



Understanding: Objective

Feature Understanding Tasks	Ranking Features	Rank the following features in order of importance for determining loan approval.
	Feature Impact	How does an increase in the following feature have a positive, negative or null effect on the loan approval?
	Feature Variability	Does the model assign the same level of importance to the feature X for these two loan applications? Justify

Understanding: Subjective

Perceived Understandability

Self-Reported Understanding

How well do you understand the explanation of how the properties contribute to the prediction?

User Reflection

Do you feel the explanation provided was sufficient for understanding the prediction?

Hoffman, Robert R., et al. "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance."

Trust

Trust Measures

Self-Reported Trust

Jian et al. (2000) and Hoffman et al. (2023)

Observed Trust

User agreement with model decisions
high agreement + incorrect decisions = **over trust**
low agreement + correct decisions = **under trust**
Weight of Advice (WOA) metric

Focus Groups

Offer much richer insights into the mental model of participants

Usability

SATISFACTION

HELPFULNESS

WORKLOAD (COGNITIVE LOAD)

FAIRNESS PERCEPTION

1. Hoffman, Robert R., Shane Mueller, Gary Klein, and Jordan Litman. 2023. "Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance." Frontiers in Computer Science 5. https://doi.org/10.3389/fcomp.2023.1096257.
2. Rong, Yao, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2024. "Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (4): 2104–22.

Part 6: Wrap up and Q&A

List 3 things that you have learnt today regarding HCXAI.

1

Centre User Needs in XAI Design

2

Balance Explanation Quality with Human Control and Autonomy

3

Combine Objective Measure with User-Centric Evaluations

Key takeaways

Readings

1. Miller, Tim. 2023. “Explainable AI Is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI.” In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–42. FAccT ’23. ACM.
2. Buçinca, Zana, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2024. “Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills.” *arXiv [Cs.HC]*. arXiv. http://arxiv.org/abs/2410.04253.
3. Ma, Shuai. 2024. “Towards Human-Centered Design of Explainable Artificial Intelligence (XAI): A Survey of Empirical Studies.” *arXiv [Cs.HC]*. arXiv. http://arxiv.org/abs/2410.21183.
4. Rong, Yao, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2024. “Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (4): 2104–22.
5. Schauer, Andreas. 2024. “In the Eyes of the User: A Systematic Literature Review of User Perceptions in Human-XAI Interaction.” In *ICIS 2024 Proceedings*. aisel.aisnet.org. https://aisel.aisnet.org/icis2024/lit_review/lit_review/12/.
6. Schmitt, Vera, Balázs Patrik Csomor, Joachim Meyer, Luis-Felipe Villa-Areas, Charlott Jakob, Tim Polzehl, and Sebastian Möller. 2024. “Evaluating Human-Centered AI Explanations: Introduction of an XAI Evaluation Framework for Fact-Checking.” In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, 91–100. MAD ’24. ACM.
7. Vera Liao, Q., and Kush R. Varshney. 2021. “Human-Centered Explainable AI (XAI): From Algorithms to User Experiences.” *arXiv [Cs.AI]*. arXiv. http://arxiv.org/abs/2110.10790.
8. Zhang, Ruihan, et al. "Invertible concept-based explanations for cnn models with non-negative concept activation vectors." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 13. 2021.

Readings

9. Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. “Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods.” In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–86. AIES ’20. New York, NY, USA: Association for Computing Machinery.
10. Le, T., Miller, T., Singh, R., & Sonenberg, L. (2024). Towards the new XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence. ArXiv, abs/2402.01292. ECAI 2024.
11. Le, T., Miller, T., Zhang, R., Sonenberg, L., & Singh, R. (2024). Visual Evaluative AI: A Hypothesis-Driven Tool with Concept-Based Explanations and Weight of Evidence. arXiv preprint arXiv:2407.04710.
12. Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.” Harvard Journal of Law & Technology 31 (2): 841–87.
13. Verma, Sahil, John Dickerson, and Keegan Hines. 2020. “Counterfactual Explanations for Machine Learning: A Review.” arXiv Preprint arXiv:2010.10596. arXiv. http://arxiv.org/abs/2010.10596.
14. Mothilal, Ramaravind K., 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, 607–17. FAT* ’20. New York, NY, USA: Association for Computing Machinery.
15. Verma, Sahil, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. “Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review.” ACM Computing Surveys, July. https://doi.org/10.1145/3677119.
16. Singh, Ronal, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. 2023. “Directive Explanations for Actionable Explainability in Machine Learning Applications.” ACM Transactions on Interactive Intelligent Systems, January. https://doi.org/10.1145/3579363.
17. Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.
18. Madumal, Prashan, et al. "A grounded interaction protocol for explainable artificial intelligence." arXiv preprint arXiv:1903.02409 (2019).

Readings

19. Dominguez, Vicente, et al. "The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images." Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019.
20. Millecamp, Martijn, et al. "To explain or not to explain: the effects of personal characteristics when explaining music recommendations." *Proceedings of the 24th international conference on intelligent user interfaces*. 2019.
21. Schoeffer, Jakob, Niklas Kuehl, and Yvette Machowski. "“There is not enough information”: On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.
22. Lai, Vivian, and Chenhao Tan. "On human predictions with explanations and predictions of machine learning models: A case study on deception detection." Proceedings of the conference on fairness, accountability, and transparency. 2019.
23. Buçinca, Zana, et al. "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems." Proceedings of the 25th international conference on intelligent user interfaces. 2020.
24. Kaur, Davinder, et al. "Trustworthy artificial intelligence: a review." ACM computing surveys (CSUR) 55.2 (2022): 1-38.
25. Vered, Mor, et al. "The effects of explanations on automation bias." Artificial Intelligence 322 (2023): 103952.
26. Cai, Carrie J., Jonas Jongejan, and Jess Holbrook. "The effects of example-based explanations in a machine learning interface." Proceedings of the 24th international conference on intelligent user interfaces. 2019.
27. Kaur, Harmanpreet, et al. "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning." Proceedings of the 2020 CHI conference on human factors in computing systems. 2020.
28. Wang, Xinru, and Ming Yin. "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making." Proceedings of the 26th International Conference on Intelligent User Interfaces. 2021.
29. Al-Ansari, Noor, Dena Al-Thani, and Reem S. Al-Mansoori. "User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review." Human Behavior and Emerging Technologies 2024.1 (2024): 4628855.
30. Ehsan, U., Riedl, M.O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In: Stephanidis, C., Kurosu, M., Degen, H., Reinerman-Jones, L. (eds) HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence. HCII 2020. Lecture Notes in Computer Science(), vol 12424. Springer, Cham.

Tutorial on Human-Centred Explainable AI (HCXAI)
Principles, Guidelines, and Practical Examples

THE 22ND AUSTRALASIAN DATA SCIENCE AND MACHINE LEARNING CONFERENCE (AUSDM'24) | Melbourne, Australia | 25-27 November 2024

Dr Ronal Singh
Research Scientist
Human-Centric Security & CINTEL FSP,
CSIRO's Data61

Prof Tim Miller
Professor of AI
School of Electrical Engineering and Computer Science
The University of Queensland