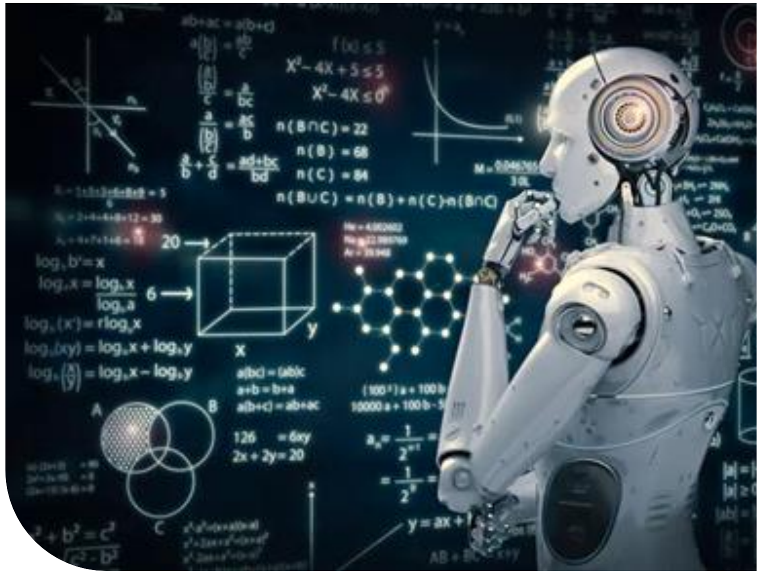
AI-Based Business Information Systems

Explainable AI



Prof. Dr. Ulrich Gnewuch

Course Organization

UNIVERSITY OF PASSAU

Lecture

AI-Enabled Business Capabilities

- AI-Enabled Innovation
- AI-Enabled Insights & Decisions
- AI-Enabled Engagement
- AI-Enabled Automation

AI Technologies & Trends

- AI Ethics & Responsible AI
- Generative AI
- Explainable AI
- Conversational AI

Foundations

- Introduction to AI in Business & Information Systems
- Design & Management of AI-Based Information Systems

Exercise

- **Exercise 4:** Generative AI & Innovation
- **Exercise 3:** Explainable AI Techniques
- **Exercise 2:** Human-Centered Chatbot Design
- **Exercise 1:** Robotic Process Automation Case Study

Industry Talk ZF Group

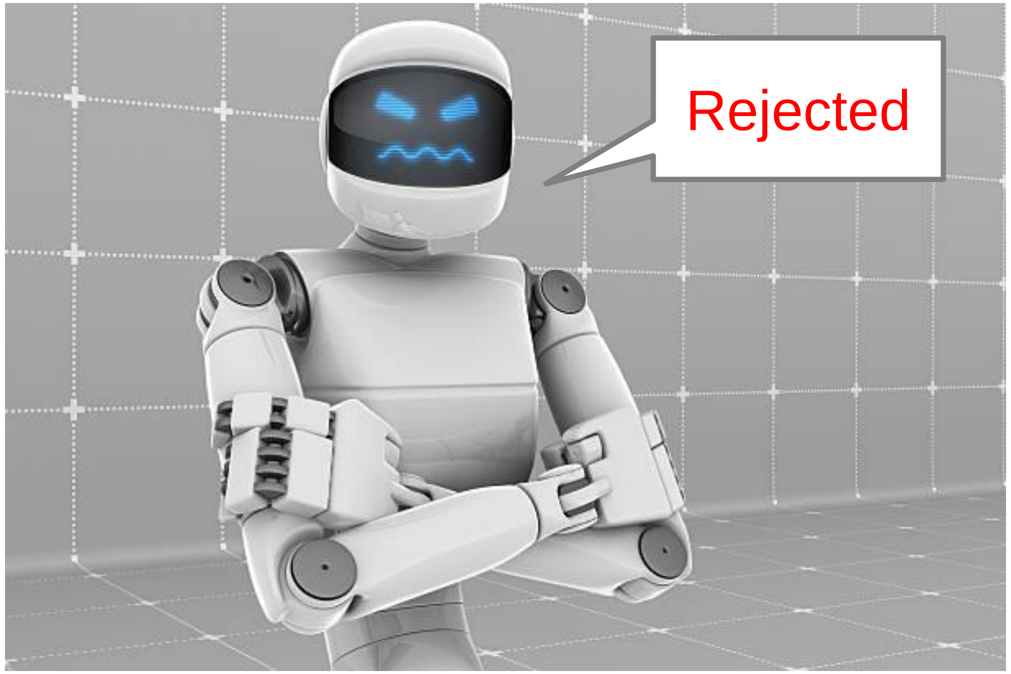
RECAP FROM LAST LECTURE:

- Please organize the following concepts based on the order in which they appear in the information value chain.
- What are key differences between the top-down knowledge-driven paradigm and the bottom-up data-driven paradigm?
- What are typical reasons why decision-makers ignore AI-enabled insights and recommendations?

Learning Goals

- Explain the concept of explainable AI (XAI) and its historical roots
- Describe the relationship between XAI stakeholders' explainability needs and the design of explanations
- Distinguish between different XAI approaches and name popular techniques
- Discuss the challenges and limitations of current XAI approaches

Why is Explainability Important?



AI is increasingly used to make consequential decisions about us

Companies need to comply with European Union regulation

Goodman & Flaxman 2017

Why is Explainability Important?

Decision-makers need to know how much they can rely on AI output

Developers want to debug and improve their AI models

Liao & Varshney 2021; Hind 2019

Explainability & Explainable AI (XAI)

Explainability is the ability for humans to understand the algorithm’s behavior. (based on Rosenfeld & Richardson 2019)

Explainable AI (XAI) is the ability of AI-based systems to explain their behaviors in understandable terms to humans. (based on Du et al. 2020)

- Explainability is often used interchangeably with other terms, such as interpretability or transparency, but there are differences
- Explainability is not just a (technical) property of a machine learning (ML) model but also considers the human side of explanations

Miller 2019; Gunning & Aha 2019; Berente et al. 2021

Brief History of Explainability

Expert Systems

Machine Learning

DARPA's Explainable AI Program

Time

1970 1980 1990 2000 2010 2020

Explainable AI is not a new topic.
The problem of explainability is as old as AI itself.

Gregor & Benbasat 1999; Mueller et al. 2019

The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice

By: L. Richard Ye
Department of Accounting and MIS, BA&E
California State University, Northridge
Northridge, CA 91330-8245
U.S.A.
rye@vax.csun.edu

Paul E. Johnson
Department of Information and Decision Sciences
University of Minnesota
271 19th Ave. So.
Minneapolis, MN 55455
U.S.A.
pjohnson@csom.umn.edu

Abstract

Providing explanations for recommended actions is deemed one of the most important capabilities of expert systems (ES). There is little empirical evidence, however, that explanation facilities indeed influence user confidence in, and acceptance of, ES-based decisions and recommendations. This paper investigates the impact of ES explanations on changes in user beliefs toward ES-generated conclusions. Grounded on a theoretical model of argument, three alternative types of ES explanations—trace, justification, and strategy—were provided in a simulated diagnostic expert system performing auditing tasks. Twenty practicing auditors evaluated the outputs of the system in a laboratory setting. The results indicate that explanation facilities can make ES-generated advice more acceptable to users and that justification is the most effective type of explanation to bring about changes in user attitudes toward the system. These findings are expected to be generalizable to application domains that exhibit similar characteristics to those of auditing: domains in which decision making tends to be judgmental and yet highly consequential, and the correctness or validity of such decisions cannot be readily verified.

Keywords: Auditing, expert systems, explanation facilities, justification, user acceptance

ISRL Categories: AI0105, EI0201, EI0208, GB02, HA04

Introduction

Expert systems (ES) are computer programs capable of performing specialized tasks based on an understanding of how human experts perform the same tasks. Few ESs, however, are targeted at replacing their human counterparts; usually they are intended to function as assistants or advisers to professional people with different technical background and problem-solving experience (Berry and Hart, 1990; Feigenbaum, et al., 1988; Leonard-Barton and Sviokla, 1988). To be useful and acceptable, it has been argued, an ES must not only perform at a level comparable to a human expert's, but also must be able to *explain*, in a form understandable to users, the reasoning processes it employs to solve problems and make recommendations (Duda and Shortliffe, 1983; Moore and Swartout, 1988; Teach and Shortliffe, 1981).

Central to the issue of explanation are two unique characteristics of ES applications. First, ESs are often developed to help make relatively unstructured decisions, and a time lag may exist between when such decisions must be made and when their quality can be assessed. As a result, the acceptance of ES-generated advice is more likely to be determined by its reasonableness than by its correctness. Second, real-world decisions have practical—financial, legal, political, and social—consequences. If users are to remain responsible for the decisions made, they are unlikely to accept a system's recommendation if they do not understand its underlying reasoning processes (Hollnagel,

The Use and Effects of Knowledge-based System Explanations: Theoretical Foundations and a Framework for Empirical Evaluation

Jasbir S. Dhaliwal • Izak Benbasat
Department of Decision Sciences, Faculty of Business Administration, National University of Singapore, Kent Ridge, Singapore 0511
fbajs@leonis.nus.sg
Management Information Systems Division, Faculty of Commerce and Business Administration, University of British Columbia, 2053, Main Mall, Vancouver, British Columbia, Canada V6T 1Z2
izak@unixg.ubc.ca

Ever since MYCIN introduced the idea of computer-based explanations to the artificial intelligence community, it has come to be taken for granted that all knowledge-based systems (KBS) need to provide explanations. While this widely-held belief has led to much research on the generation and implementation of various kinds of explanations, there has been no theoretical basis to justify the use of explanations by KBS users. This paper discusses the role of KBS explanations to provide an understanding of both the specific factors that influence explanation use and the consequences of such use.

The first part of the paper proposes a model based on cognitive learning theories to identify the reasons for the provision of KBS explanations from the perspective of facilitating user learning. Using the feedforward and feedback operators of cognitive learning the paper develops strategies for providing KBS explanations and classifies the various types of explanations found in current KBS applications.

This second part of the paper presents a two-part framework to investigate empirically the use of KBS explanations. The first part of the framework focuses on the potential factors that influence the explanation seeking behavior of KBS users, including user expertise, the types of explanations provided and the level of user agreement with the KBS. The second part of the framework explores the potential effects of the use of KBS explanations and specifically considers four distinct categories of potential effects: explanation use behavior, learning, perceptions, and judgmental decision making.

(*Knowledge-based System Explanations; Expert Systems; Cognitive Learning; Feedforward and Feedback Information*)

1. Introduction

Many types of knowledge-based systems (KBS) that capture, represent, and apply expert knowledge are currently being used successfully in various industrial and administrative applications (Hayes-Roth and Jacobstein 1994). They serve as independent decision makers, e.g., in situations where there are no human experts available or when they act as embedded controllers in smart

1047-7047/96/0703/0342$01.25

Ye & Johnson 1995; Dhaliwal & Benbasat 1996

Explanations in Expert Systems: MYCIN Example

UNIVERSITY OF PASSAU

- MYCIN was one of the best-known expert systems (developed in the 1970s)
- It was designed to provide advice for physicians regarding diagnosis and therapy for infectious diseases
- MYCIN offered two types of explanations:
  - Users could ask “HOW” in response to a recommendation and receive a trace of the rules fired
  - Users could ask “WHY” in response to being asked a question by the system, in which case MYCIN would provide a trace of the currently active goal

MYCIN: There is evidence that the type of infection is bacterial.

User: HOW? [How was it established that the type of infection is bacterial?]

MYCIN: The following rules concluded about the likelihood that the type of infection is bacterial:

Rule 148 (0.19)
Rule 500 (0.51)
Rule 501 (0.95)
Rule 502 (0.97)
Rule 526 (0.98)

Thus, it has been established that the type of infection is bacterial.

In light of the site from which the culture was obtained, and the method of collection, do you feel that a significant number of ORGANISM-1 were obtained?

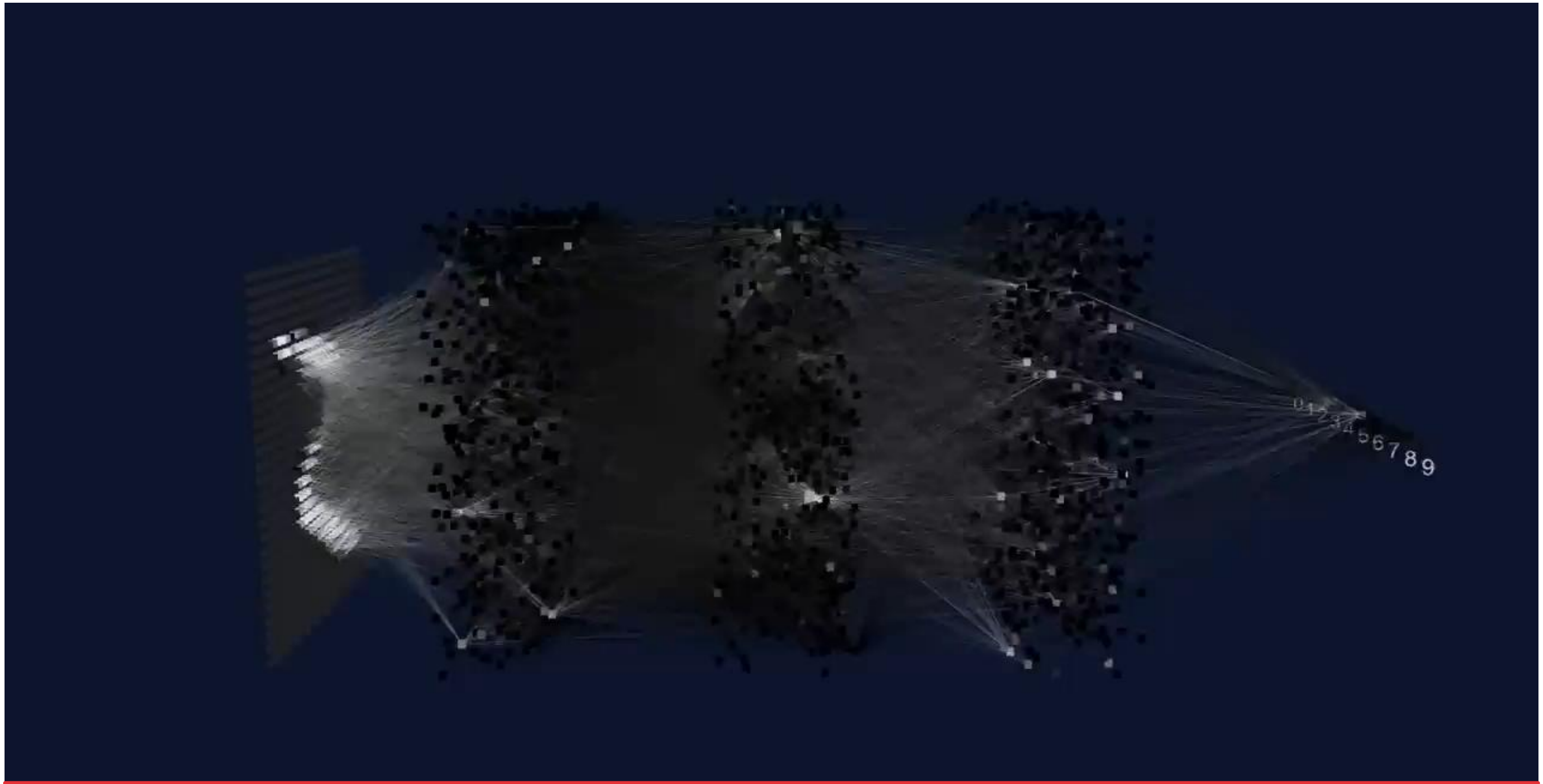
** *WHY 2*

We are trying to find out whether the organism has been observed in significant numbers in order to determine an organism or class of organism for which therapy should cover

Buchanan & Shortliffe 1984

Today, Explainability is More Difficult (and More Important)

UNIVERSITY OF PASSAU



How can we explain machine learning models?

Neural network for recognizing handwritten digits (MNIST dataset)

https://www.youtube.com/watch?v=Tsvxx-GGlTg

Explainability Needs & Explanation Design

XAI Stakeholders and Their Explainability Needs

AI Developers
(debug and improve
AI models)

Impacted Groups
(seek recourse or
contest the AI)

- Customers
- Patients
- Applicants
- …

Decision-Makers
(make informed decisions
based on an AI application)

Regulatory Bodies
(ensure that the AI is
safe, and society is not
negatively impacted)

**Business Owners and
Senior Management**
(assess an AI application's capability,
regularity compliance, …)

…

Liao & Varshney 2021; Hind 2019

XAI Stakeholders and Their Explainability Needs

AI Developers
(debug and improve AI models)

Impacted Groups
(seek recourse or contest the AI)

- Customers
- Patients
- Applicants
- …

Decision-Makers
(make informed decisions based on an AI application)

Regulatory Bodies
(ensure that the AI is safe, and society is not negatively impacted)

→ There can be no “one-fits-all” solution to XAI!

Liao & Varshney 2021; Hind 2019

Stakeholders & Explainability Needs in AI-based Loan Application Decision-Making

Please imagine the following scenario:

A bank employs an AI system (“RoboLoan”) to assist bank consultants in evaluating loan applications submitted by private consumers. The system analyzes applicants’ data and provides recommendations on whether a loan should be approved or rejected.

1. Who are the relevant XAI stakeholders in this scenario?
2. What are their specific explainability needs?

→ Discuss these questions with a partner for **~5 minutes** and be ready to share your answers

Explainability Needs Expressed as Questions

Task objectives	Main stakeholders who engage in this task	Example questions they may ask the AI
Debug and improve AI models	Model Developers	• *Is the AI's **performance** good enough?* • ***How** does the AI make predictions? How might it go wrong?* • ***Why** does the AI make such a mistake?*
To evaluate AI's capability and form appropriate trust	All stakeholders may engage in this task at some point	• *Is the AI's **performance** good enough? What are the risks and limitations?* • *What kinds of **output** can the AI give?* • *How does the AI work? Is it reasonable?*
Make informed decisions or take better actions	Decision-Makers, Impacted Groups	• ***Why** is this instance predicted to be X?* • ***Why** is this instance **not** predicted to be Y?* • ***How to change** this instance **to be** predicted Y?*
To adapt usage or control	Decision-Makers, Business Owners / Senior Management	• ***How** does the AI make predictions? What can I supply or change for it to work well?* • ***What if** I make this change?*
Ensure ethical or legal compliance	Regulatory Bodies	• ***How** does the AI make predictions? Are there any legal/ethical concerns, such as discrimination, privacy, or security concerns?* • ***Why** are the two instances/groups **not** treated the same by the AI?*
…	…	…

Liao et al. 2020; Liao & Varshney 2021

Question-Driven XAI

- **How**: asking about the general logic or process the AI follows (to have a global view)
- **Why**: asking about the reason behind a specific prediction
- **Why Not**: asking why the prediction is different from an expected or desired outcome

- **How to change to be that**: asking about ways to change the instance to get a different prediction
- **How to remain to be this**: asking what change is allowed for the instance to still get the same prediction
- **What if**: asking how the prediction changes if the input changes

- **Performance**: asking about the performance of the AI
- **Data**: asking about the training data
- **Output**: asking what can be expected or done with the AI's output

Liao et al. 2020

Mapping Between Questions and Explanations

Questions	Possible Ways to Explain
How (global model-wide)	• Describe the general model logic (e.g., as feature impact, rules, or decision-trees) • If a user is only interested in a high-level view, describe what are the top features or rules considered
Why	• Describe what key features of the instance determine the model's prediction of it • Describe rules that the instance fits to guarantee the prediction • Show similar examples with the same predicted outcome to justify the model's prediction
Why not	• Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction • Show prototypical examples that had the alternative outcome
How to be that (a different prediction)	• Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction • Show examples with minimum differences but had a different outcome than the prediction
How to still be this (the current prediction)	• Describe features/feature ranges or rules that could guarantee the same prediction • Show examples that are different from the particular instance but still had the same outcome
What if	• Show how the prediction changes corresponding to the inquired change
Performance	• Provide performance metrics of the model • Show uncertainty information for each prediction
Data	• Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.
Output	• Describe the scope of output or system functions

Liao & Varshney 2021

Explanations

- Content
- Presentation Format
- Provision Mechanism
- Timing
- Interactivity
- …

Explanation Presentation Format

Presentation Format

Text-based

Multimedia

Explanations are provided in **natural language**:

- Rules of an expert system
- Plain natural language
- …

Explanations are provided in **non-text** format:

- Graphs
- Histograms
- Images
- …

Gregor & Benbasat 1999

Explanation Presentation Format: Examples

“If the applicant’s income had been $10,000 higher, the loan would have been approved”

Counterfactual Explanations

Saliency Maps

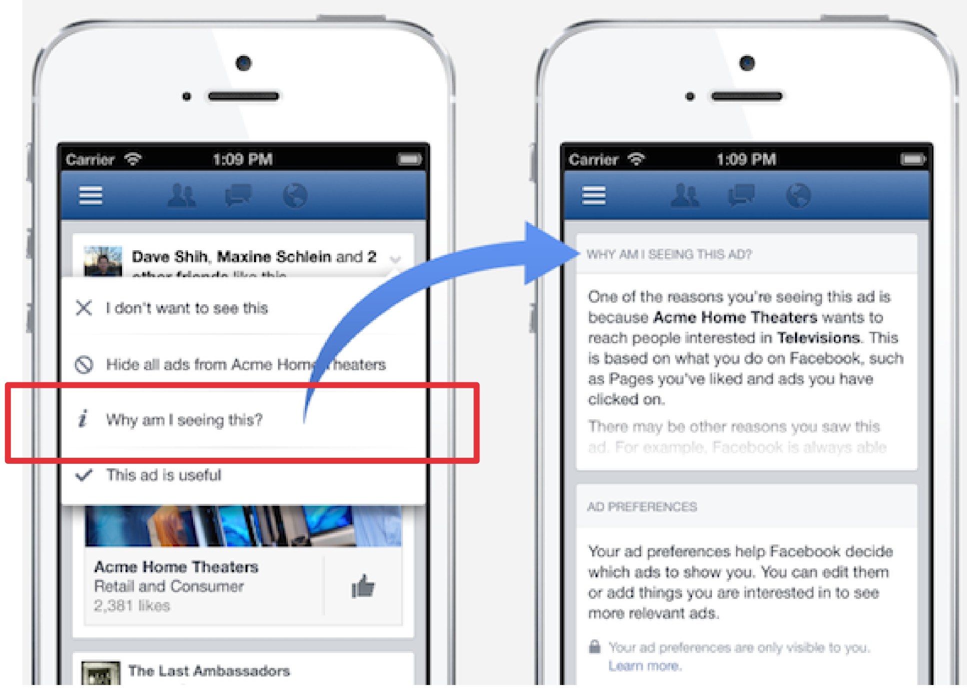
Explanation Provision Mechanism

Provision Mechanism

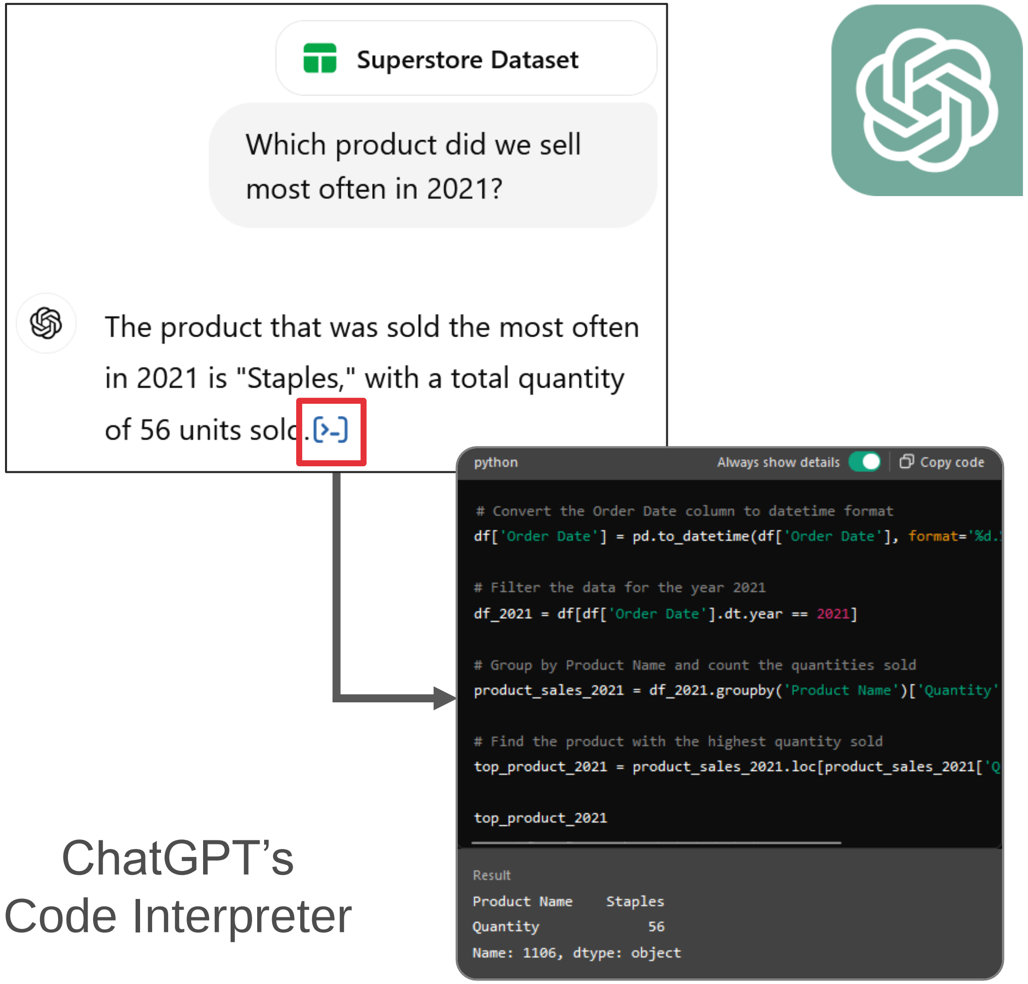
User-Invoked	Automatic	Intelligent
Explanations are provided **at the requests of users**.	Explanations are provided **automatically** without users' requests.	Explanations are provided to users **by the AI when** it considers such explanations are **needed**.

Gregor & Benbasat 1999

User-Invoked Explanation Provision: Example



Facebook's "Why am I seeing this ad?"



ChatGPT's
Code Interpreter

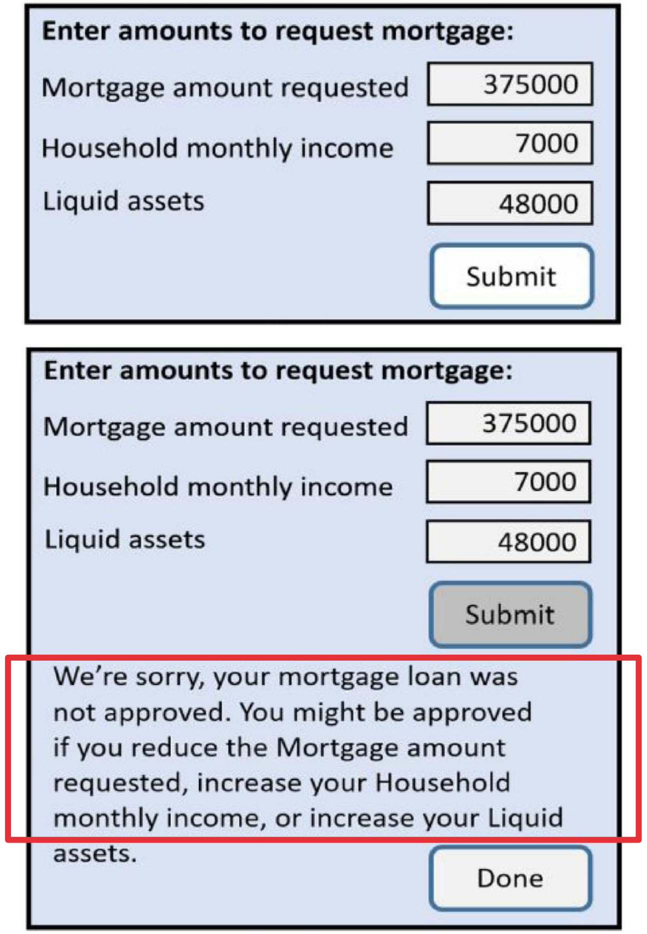
Explanation Timing

Timing

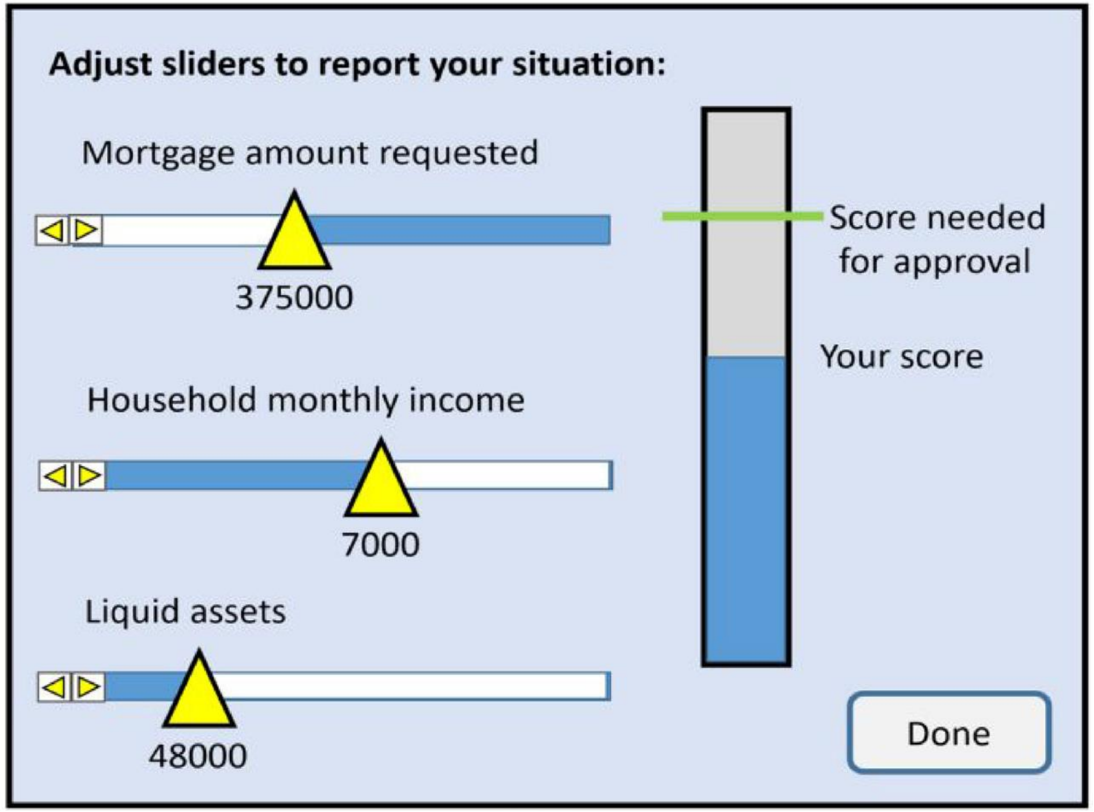
- Feedforward Explanation
  Explanations are provided **before** a certain behavior is performed by the AI.
- Feedback Explanation
  Explanations are provided **after** a certain behavior is completed by the AI.

Dhaliwal & Benbasat 1996

Interactivity



Static Explanation

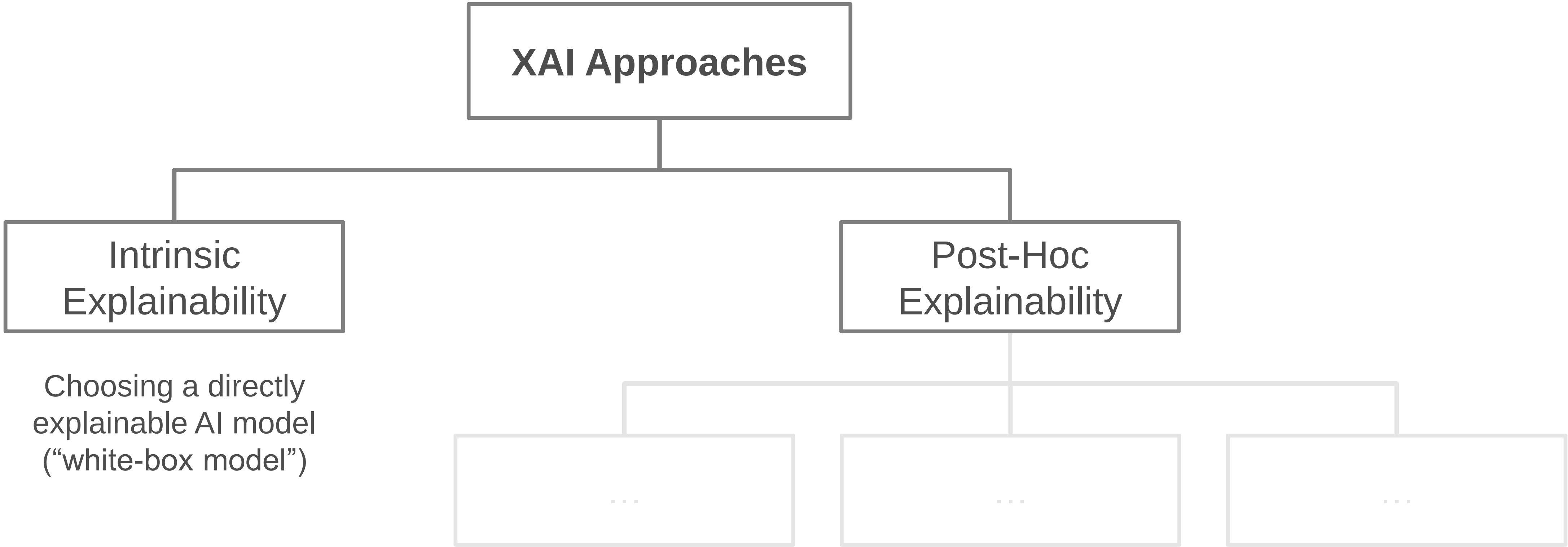


Interactive Explanation Interface

Shneiderman 2020

Explainable AI Approaches

Overview of Explainable AI Approaches



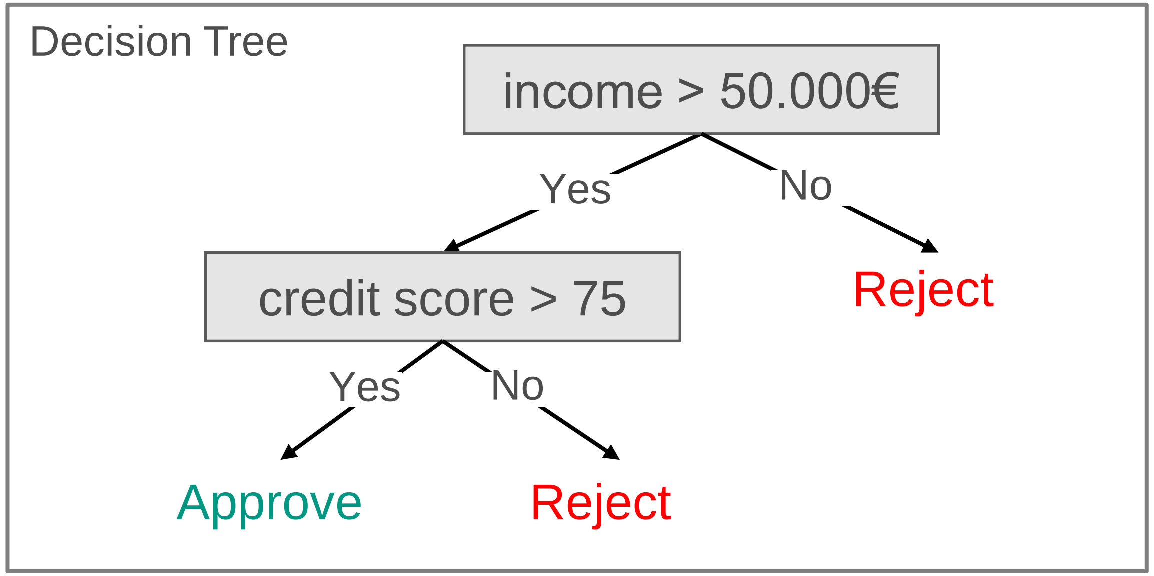
Choosing a directly explainable AI model (“white-box model”)

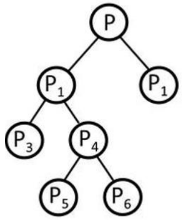
Choosing a not self-explanatory AI model (“black-box model”) and then using a post-hoc technique to generate explanations

Guidotti et al. 2019

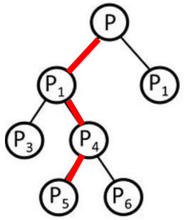
Intrinsic Explainability: White-Box Models

- White-box models incorporate explainability directly into their structures
- Examples: decision tree, linear/logistic regression, rule-based models
- Sometimes not possible and can also get quite complex





Splitting Criteria
(for the overall model)



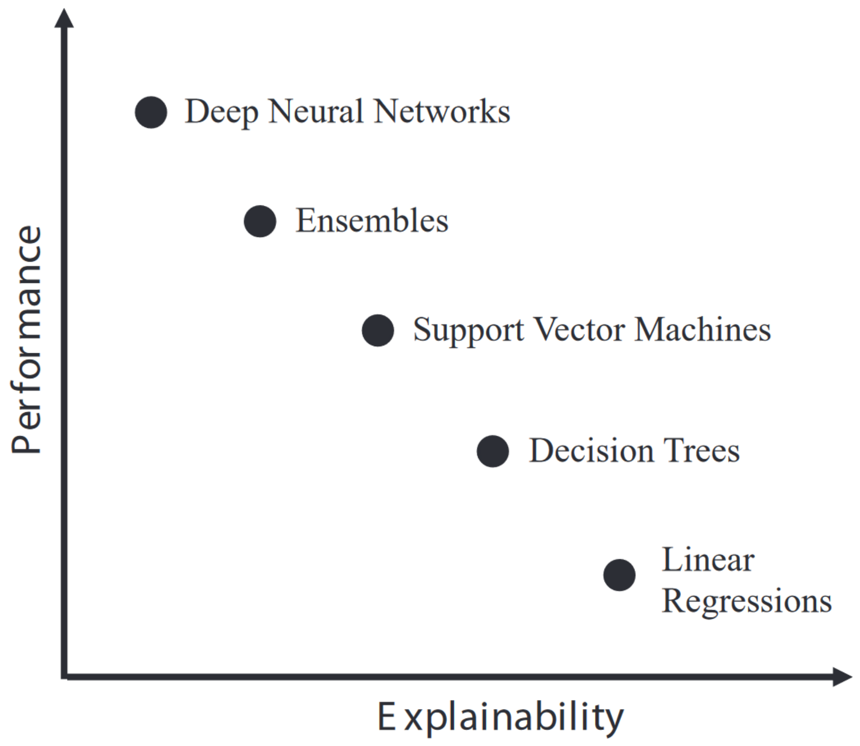
Decision path
(for a single decision)

Du et al. 2019

Performance – Explainability Tradeoff

Why not always choose a white-box model?

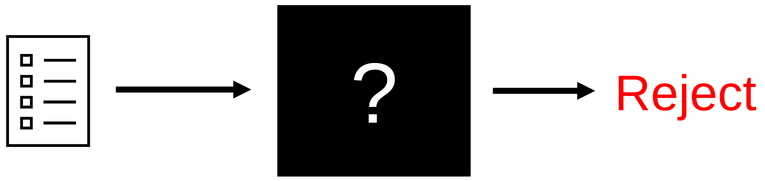
- Black-box models often outperform white-box models due to their ability to capture high non-linearity and interactions between features
- The choice between the two is discussed under the term “performance–explainability tradeoff”
- This **tradeoff is not always true**: In many contexts, especially with well-structured datasets and meaningful features, white-box models can reach comparable performance



Herm et al. 2023; Liao & Varshney 2021

Black Box Models

- Black-box models are complex and not self-explanatory
- Examples: neural networks, ensemble models
- Post-hoc explainability techniques can be used to generate explanations for their output (e.g., predictions)
  - Can be applied to any model (model-agnostic)
  - But **usually an approximation** and not always faithful!



Rudin 2019

XAI Approaches: Post-Hoc Explainability

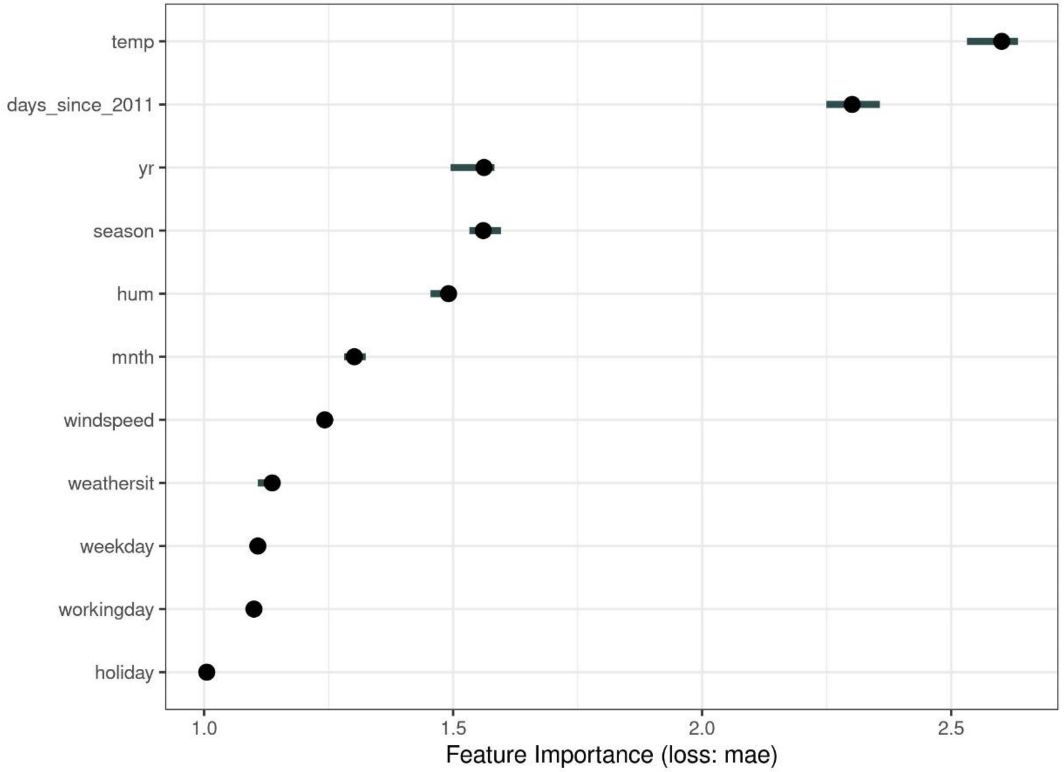
XAI Approaches

- Intrinsic Explainability
- Post-Hoc Explainability
  - Explaining the Model (**Global**)
    Explaining the overall logic of the model (e.g., feature importance)
  - Explaining a Single Output (**Local**)
    Explaining a particular model output (e.g., local feature contribution)
  - Inspecting Counterfactual
    Explaining how the model would behave with an alternative input (e.g., contrastive feature)

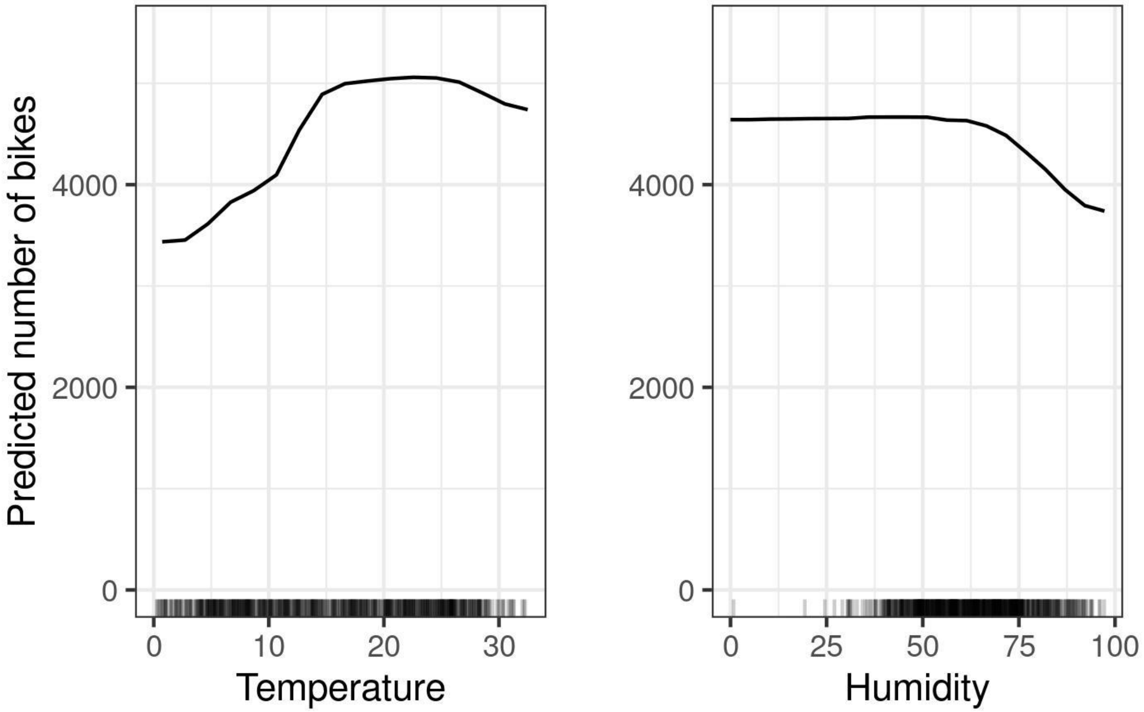
Guidotti et al. 2019; Du et al. 2019

Global Explanations: Examples

Permutation Feature Importance



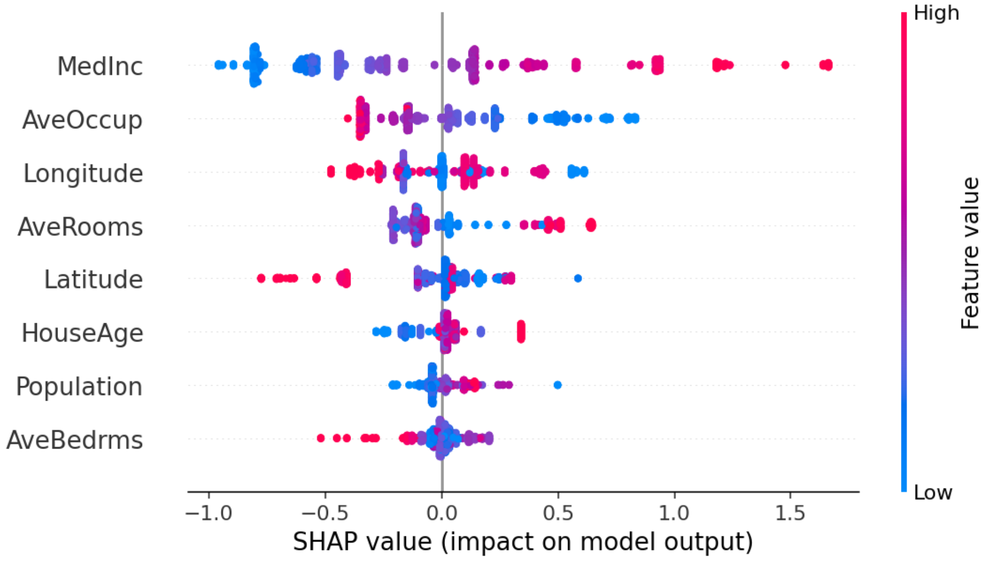
Partial Dependence Plot



https://christophm.github.io/interpretable-ml-book/pdp.html

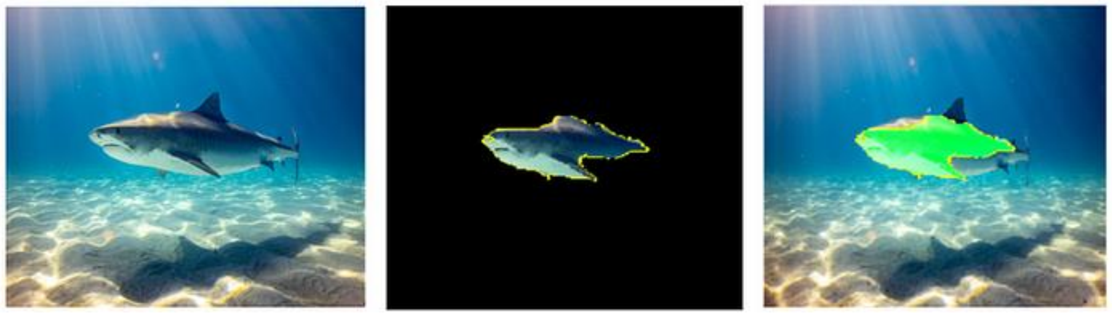
Local Explanations: Examples

SHAP



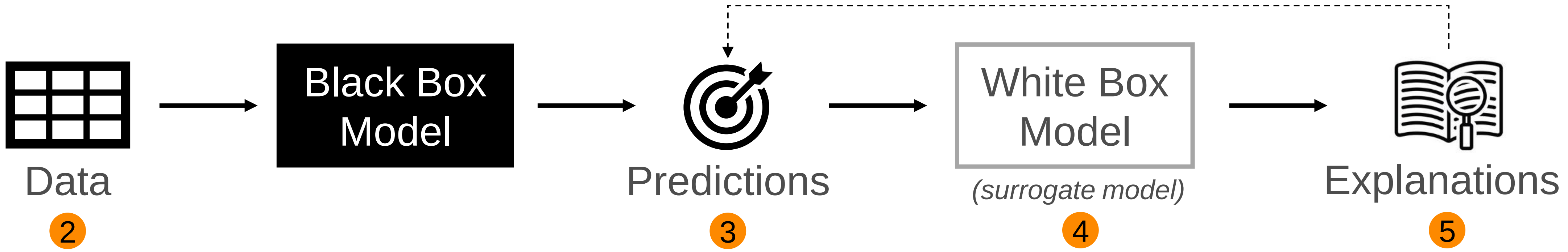
(SHAP values can be used for both
local and global explanations)

LIME



https://christophm.github.io/interpretable-ml-book/local-methods.html

LIME (Local Interpretable Model-Agnostic Explanations)



1. Select an instance for which you need an explanation of its black-box model prediction
2. Create a dataset of similar instances (“perturbing”)
3. Get the black-box model predictions for all these instances
4. Train a white-box model (“surrogate model”) on the new dataset consisting of instances and corresponding black-box model predictions
5. Use the white-box model to generate explanations for the black-box model’s prediction

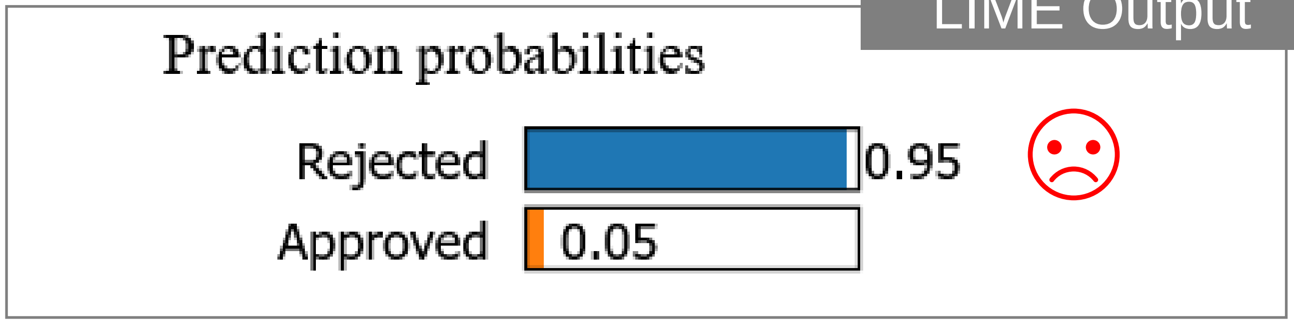
Ribeiro et al. 2016

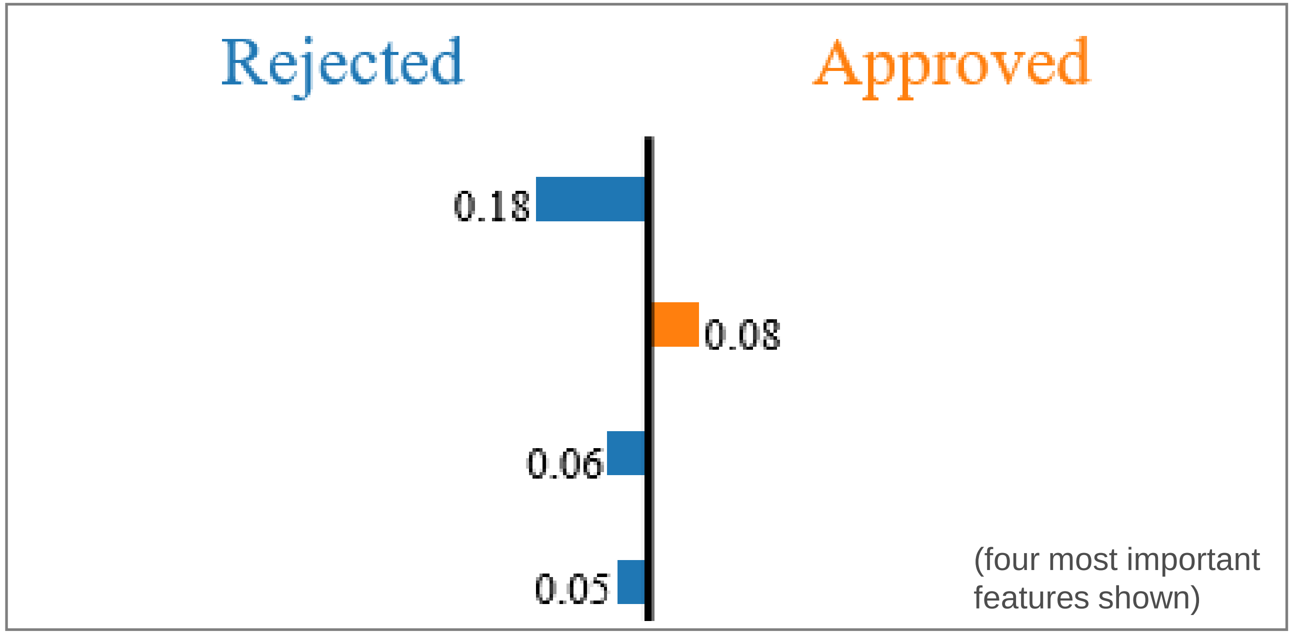
Example: LIME

	Applicant Information
Amount (EUR]	2.835€
Duration (months)	24
Purpose	furniture/equipment
Checking Account	no checking account at this bank
Loan History	previous loans paid back duly
Employment	Longer than 7 years

Prediction: Rejected

LIME Output





LIME Python package: https://github.com/marcotcr/lime

Dataset: https://doi.org/10.24432/C5NC77

LIME Results

Examine the explanation generated by LIME for the loan application data on the previous slide.

1. What do you think is the most important feature in this prediction (0.18)?
2. Which features do you think support the approval of the loan application (orange) and which ones contribute to the rejection (blue)?

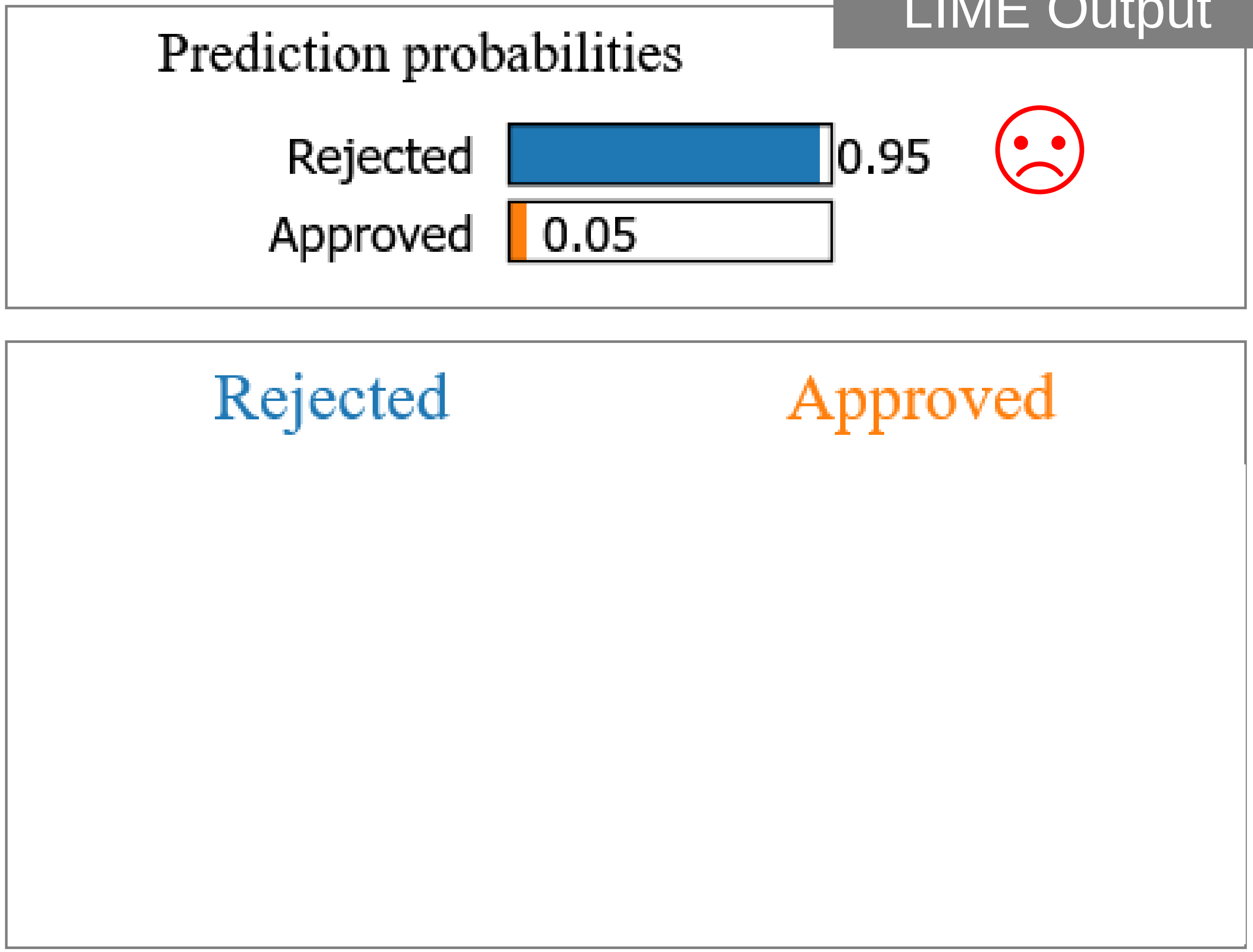
→ Discuss these questions with a partner for **2-3 minutes** and be ready to share your answers

Example: LIME

	Applicant Information
Amount (EUR]	2.835€
Duration (months)	24
Purpose	furniture/equipment
Checking Account	no checking account at this bank
Loan History	previous loans paid back duly
Employment	Longer than 7 years

Prediction: Rejected

LIME Output



LIME Python package: https://github.com/marcotcr/lime

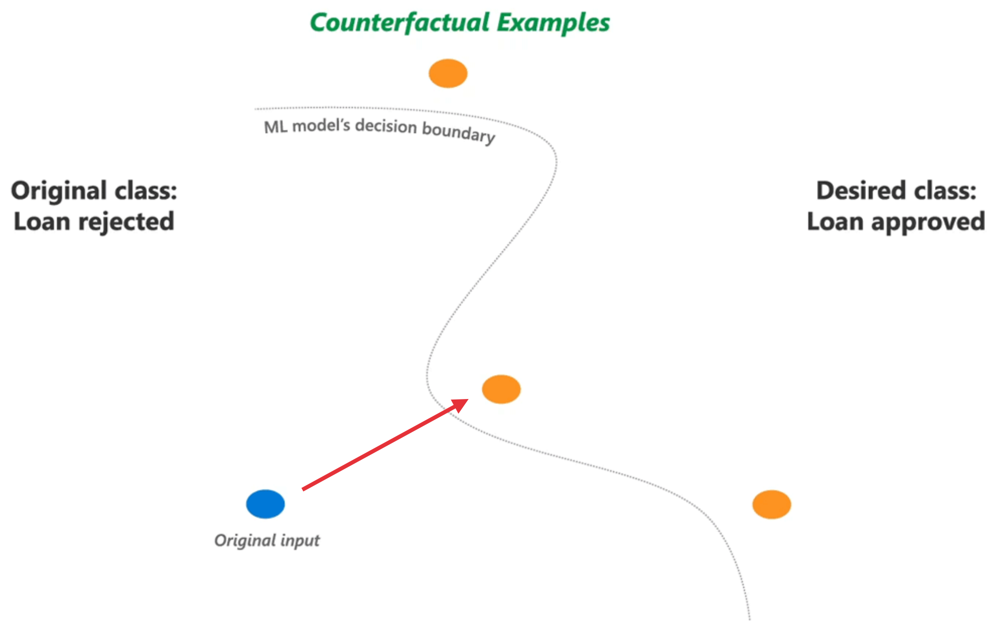
Dataset: https://doi.org/10.24432/C5NC77

Counterfactual Explanations: Example

- Counterfactual explanations provide an understanding of model outputs by posing hypothetical "what if" scenarios
- Counterfactuals show how the prediction would change if certain features were different

Example:

"If the applicant's income had been $10,000 higher, the loan application would have been approved."



Fernández-Loría et al. 2022

Mapping Between Questions, Explanations, and XAI Techniques

Questions	Possible Ways to Explain	Example XAI Technique
How (global model-wide)	• Describe the general model logic (e.g., as feature impact, rules, or decision-trees) • If a user is only interested in a high-level view, describe what are the top features or rules considered	**PFI, PDP, SHAP, …**
Why	• Describe what key features of the instance determine the model's prediction of it • Describe rules that the instance fits to guarantee the prediction • Show similar examples with the same predicted outcome to justify the model's prediction	**SHAP, LIME, …**
Why not	• Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction • Show prototypical examples that had the alternative outcome	**Counterfactuals, CEM, …**
How to be that (a different prediction)	• Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction • Show examples with minimum differences but had a different outcome than the prediction	**Counterfactuals, CEM, …**
How to still be this (the current prediction)	• Describe features/feature ranges or rules that could guarantee the same prediction • Show examples that are different from the particular instance but still had the same outcome	**CEM, …**
What if	• Show how the prediction changes corresponding to the inquired change	**PDP, …**

Note: Permutation Feature Importance (PFI), Partial Dependence Plot (PDP), SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), Contrastive Explanations Method (CEM)

Liao & Varshney 2021

More XAI Techniques …

- The development of new XAI techniques is a rapidly evolving and highly active research area

- A comprehensive overview of existing XAI techniques can be found here:

https://kdd-lab.github.io/XAISurvey/

Bodria et al. 2023

Challenges and Limitations of XAI

Miscalibrated Trust & Inappropriate Reliance

Explanations can lead to a false sense of confidence and unwarranted trust

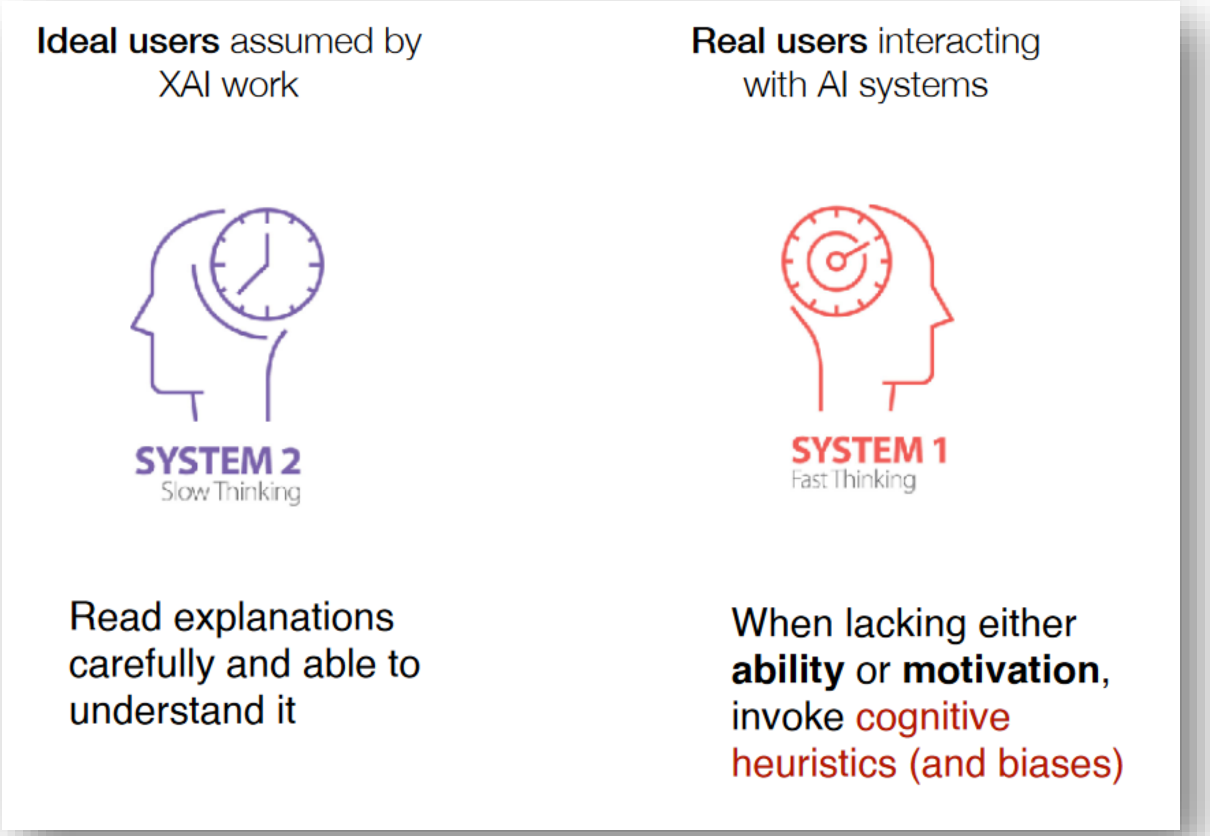
vs.

Explanations can make people lose trust in AI and under-rely on it

Bauer et al. 2023; Zhang et al. 2020; Ostinelli et al. 2024

Information Overload & Lack of Expertise

- Overly complex and detailed explanations can cause information overload
- This creates frustration and confusion
- People might misinterpret or not fully understand the explanations

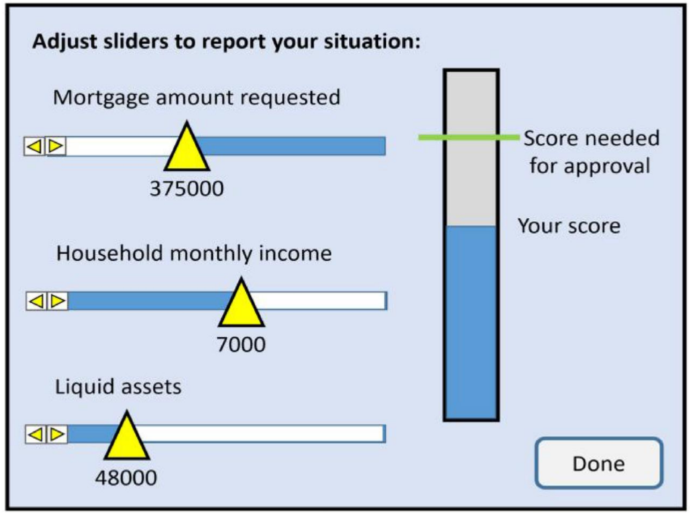


de Bruijn 2022; Poursabzi-Sangdeh et al. 2021

“Gaming the System”

- Explanations may enable people to “game” the AI system

- If people have access to information about how a decision or recommendation has been made, they might alter their behavior to gain a more favorable outcome



“If the applicant’s income had been $10,000 higher, the loan application would have been approved.”

Khosravi et al. 2020

Data Privacy and Intellectual Property Concerns

- Providing detailed explanations might risk exposing **proprietary information** about the model's architecture or training data
  - This could create tensions between transparency and **competitive advantage**

- Explanations can also introduce privacy risks by inadvertently revealing **sensitive information** embedded in the model's training data
  - This could create tensions between transparency and **data protection**

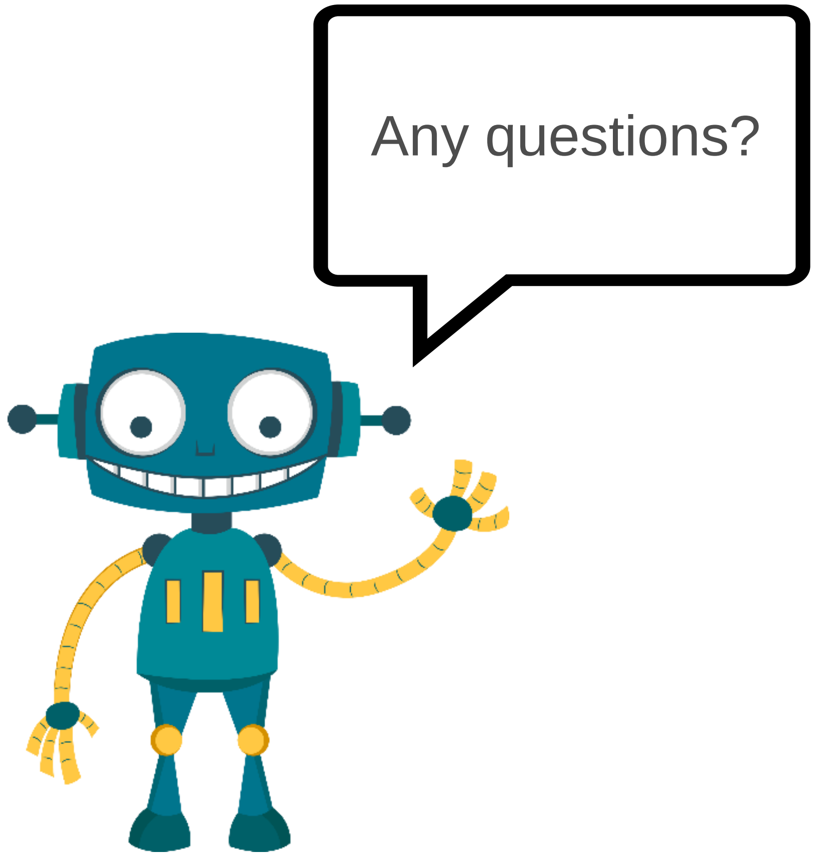
Goethals et al. 2023

Key Takeaways From This Lecture

UNIVERSITY OF PASSAU

- XAI is a long-standing, sociotechnical challenge involving a technical side and a human side
- Different people have different explainability needs (e.g., developer vs. decision-maker vs. customer)
- Explanations can differ in a variety of ways (e.g., content, presentation format, timing)
- Two main approaches to XAI exist: *intrinsic* explainability and *post-hoc* explainability
  - There are many different post-hoc XAI techniques (e.g., LIME)
- Despite its benefits, XAI has several downsides (e.g., information overload, data privacy concerns) and risks (e.g., miscalibrated trust, overreliance)

Thank you for your attention!



References (1)

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. Information systems research, 34(4), 1582-1602.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. MIS quarterly, 45(3).

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box models. Data Mining and Knowledge Discovery, 37(5), 1719-1778.

Buchanan, B., & Shortliffe, E. (1984). Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison Wesley Publishing Company, Reading, MA.

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, 39(2), 101666.

Dhaliwal, J. S., & Benbasat, I. (1996). The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation. Information systems research, 7(3), 342-362.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Communications of the ACM, 63(1), 68–77.

Fernández-Loría, C., Provost, F., & Han, X. (2022). Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach. MIS Quarterly, 46(3), 1635-1660.

Goethals, S., Sörensen, K., & Martens, D. (2023). The privacy issue of counterfactual explanations: explanation linkage attacks. ACM Transactions on Intelligent Systems and Technology, 14(5), 1-24.

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine, 38(3), 50-57.

Gregor, S., & Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. MIS Quarterly, 23(4), 497.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AI magazine, 40(2), 44-58. https://www.darpa.mil/program/explainable-artificial-intelligence

Herm, L. V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. International Journal of Information Management, 69, 102538.

References (2)

Hind, M. (2019). Explaining explainable AI. XRDS: Crossroads, The ACM Magazine for Students, 25(3), 16-19.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074.

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. arXiv preprint arXiv:2110.10790.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-15).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. https://apps.dtic.mil/sti/citations/trecms/AD1073994

Ostinelli, M., Bonezzi, A., & Lisjak, M. (2024). Unintended effects of algorithmic transparency: The mere prospect of an explanation can foster the illusion of understanding how an algorithm works. Journal of Consumer Psychology.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-52).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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 (pp. 1135-1144).

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. Autonomous Agents and Multi-Agent Systems, 33, 673-705.

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.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-31.

Ye, L. R., & Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. MIS Quarterly, 157-172.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 295-305).