

MAI4CAREU

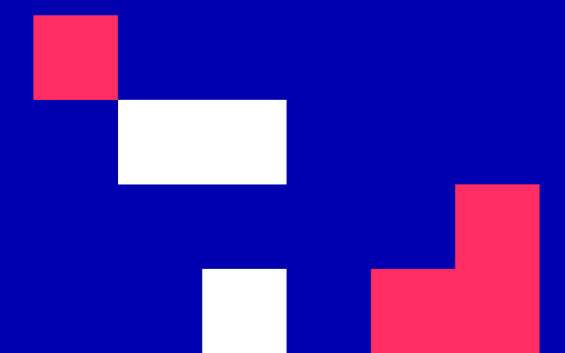
Master programmes in Artificial
Intelligence 4 Careers in Europe

University of Cyprus

HUMAN-CENTERED INTELLIGENT USER INTERFACES - MAI648

Marios Belk

2022



CONTENT 11

Explainable AI

CONTENTS

- Definitions
 - Blackbox Problem
 - Importance of Explainable AI
 - Explainable AI and GDPR
 - Design Guidelines for Explainability
- Types of Explainability
 - Challenges for Designing Explainable AI

CONTENT 11

Learning Outcomes

- Know terms and definitions of explainable AI
- Understand the benefits of how explainable can increase user trust and acceptance in interactive systems
- Evaluate the challenges for designing explainable AI-based user interfaces

CONTENT 11

Definitions

- *Explainable AI (XAI) is AI in which the results of the solution can be understood by humans. It contrasts with the concept of the "black box" in machine learning where even its designers cannot explain why an AI arrived at a specific decision - Wikipedia*
- Also know as **Interpretable AI, Explainable Machine Learning**
- https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

The “Clever Hans” Problem



By Unknown author - www.kryptozoologie.net, Public Domain,
<https://commons.wikimedia.org/w/index.php?curid=4479308>

CONTENT 11

More on definitions

- Explainable Artificial Intelligence (XAI) is an emerging research topic of machine learning aimed at **unboxing how AI systems' black-box choices** are made

Guang Yang, Qinghao Yede, Jun Xiaf (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion, 77, 29-52

CONTENT 11

Main directions

- How to make a model interpretable
- How to explain the information to the user

CONTENT 11

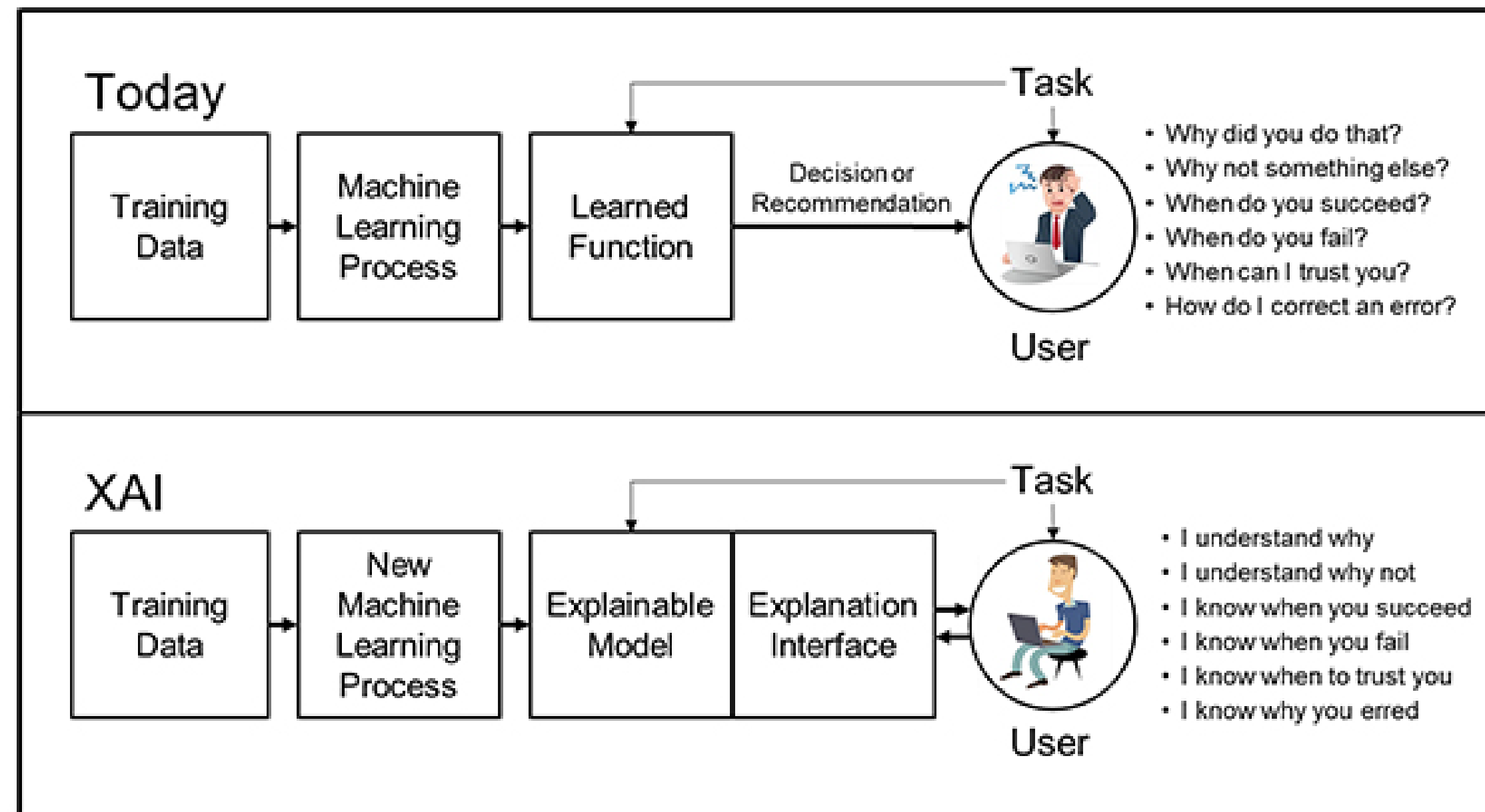
Black Box Problem

- In a nutshell
- For an AI mechanism, people don't know how the mechanism works to produce the output

CONTENT 11

- Explainable AI aims to make the AI-based decision understandable

CONTENT 11



Source: Darpa

<https://www.birlasoft.com/articles/demystifying-explainable-artificial-intelligence>

CONTENT 11

Supervised Machine Learning

Training Data

Explaining Data

Explaining model facts: performance, limitations

Labels

Cats



Dogs



Learning Model - Machine Learning Algorithm

Prediction Label: Dog

XAI aims to explain the model's decision

New Instance



CONTENT 11

Supervised Machine Learning

Training Data

Labels
Cats



Dogs



Learning Model - Machine Learning Algorithm

Prediction Label:
Cat

USER

How did you decide on that?

Why did you fail?

When do I get a successful result?

Can I trust the machine?

New Instance



CONTENT 11

Supervised Machine Learning

Training Data

Labels
Cats



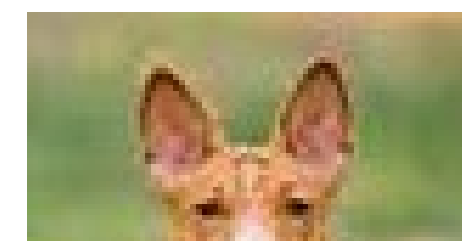
Dogs



Learning Model -
Machine Learning Algorithm

Prediction Label:
Cat

Explanation Interface
This is a cat because it has pointy ears



New Instance



CONTENT 11

Why is XAI important?

CONTENT 11

GDPR - “rights to explanation”

- Article 13 (2) f and Article 15 (1) h
 - “the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.”
- Article 22
 - “The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.”

CONTENT 11

Problems of Black Box in a variety of domains

- Bias in court
- Lack of transparency in financing
- Discrimination in recruiting
- Bias in sexual orientation
- Lack of transparency about algorithm limitations like in translators

Source: Sarah Theres Völkel. Explainable AI. Introduction to Intelligent User Interfaces

CONTENT 11

Explainability

- IBM Design for AI: Explainability -
<https://www.ibm.com/design/ai/ethics/explainability/>
- Google PAIR: Explainability+Trust
- SAP Design Guidelines for Explainability
- UXAI for Designers

CONTENT 11

IBM's Recommended actions

<https://www.ibm.com/design/ai/ethics/explainability/>

- 01. Allow for questions. A user should be able to ask why an AI is doing what it's doing on an ongoing basis. This should be clear and up front in the user interface at all times.
- 02. Decision making processes must be reviewable, especially if the AI is working with highly sensitive personal information data like personally identifiable information, protected health information, and/or biometric data.

CONTENT 11

IBM's Recommended actions

<https://www.ibm.com/design/ai/ethics/explainability/>

- 03. When an AI is assisting users with making any highly sensitive decisions, the AI must be able to provide them with a sufficient explanation of recommendations, the data used, and the reasoning behind the recommendations.
- 04. Teams should have and maintain access to a record of an AI's decision processes and be amenable to verification of those decision processes.

CONTENT 11

SAP Design Guidelines for Explainability

- SAP suggests three explanation levels: minimum, simple, and expert

<https://experience.sap.com/fiori-design-web/explainable-ai/>

Level 1 WHAT



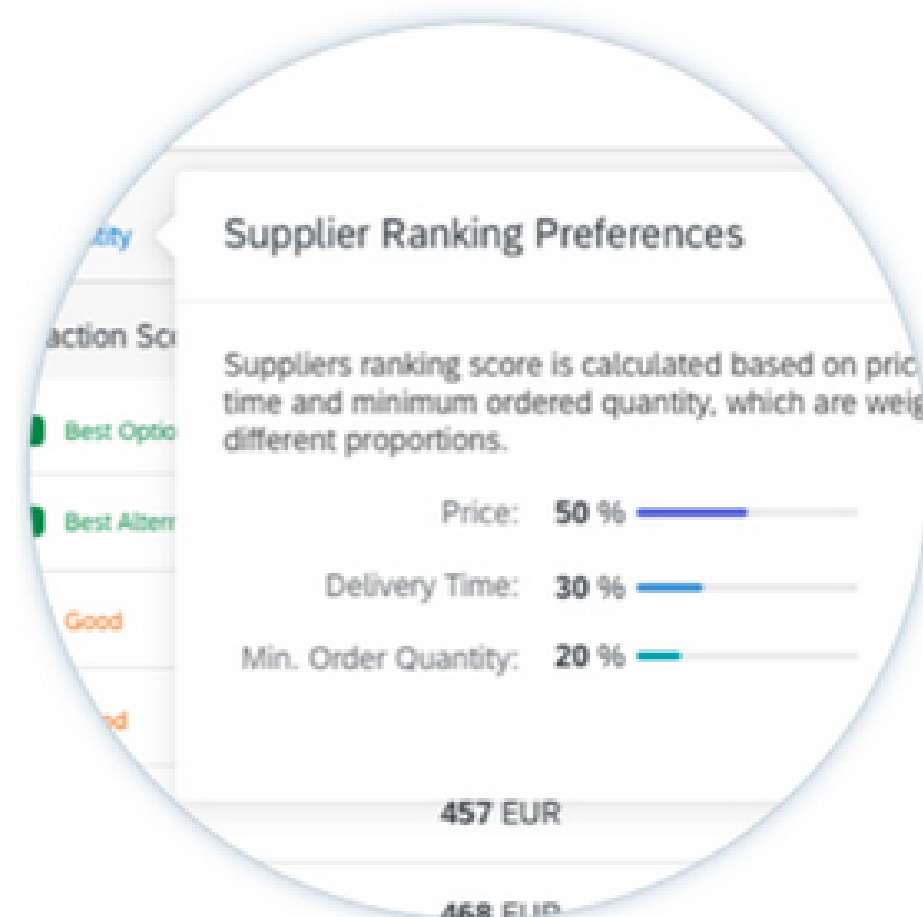
Minimum

Explanation level 1

Level 1: Indicator (What?)

The minimum explanation level. An indicator is required whenever AI (machine learning) output is provided. The indicator is also the access point for the next explanation level (if required).

Level 2 WHY



Simple

Explanation level 2

Level 2: Abstract (Why?)

A condensed view of the relevant properties, amounts, and contextual information. An abstract helps users to better understand the AI proposals. It can contain links to the last and most detailed explanation level.

Level 3 HOW



Expert

Explanation level 3

Level 3: Detail (How?)

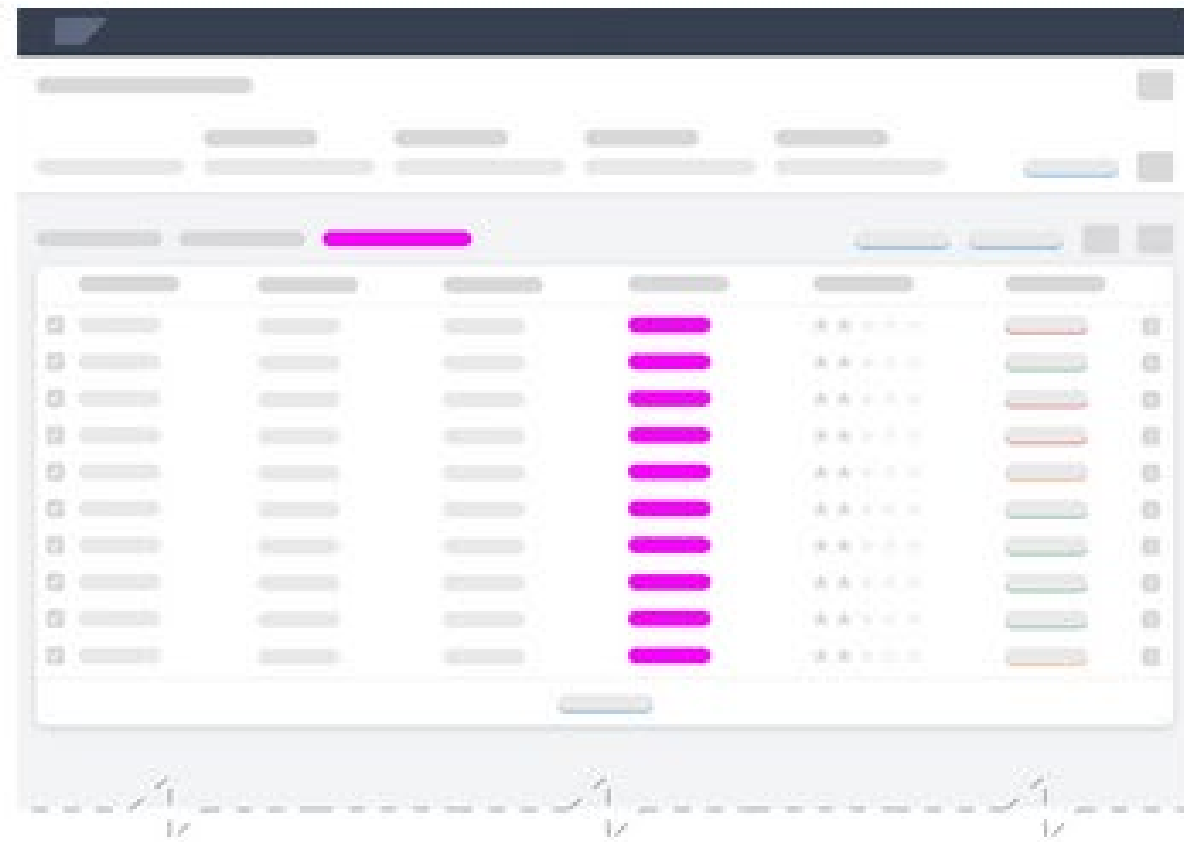
An extended report specifically for advanced users. It covers all aspects processed by the intelligent system, the AI performance, and any further context and conditions that help users to monitor AI operations.

<https://experience.sap.com/fiori-design-web/explainable-ai/>

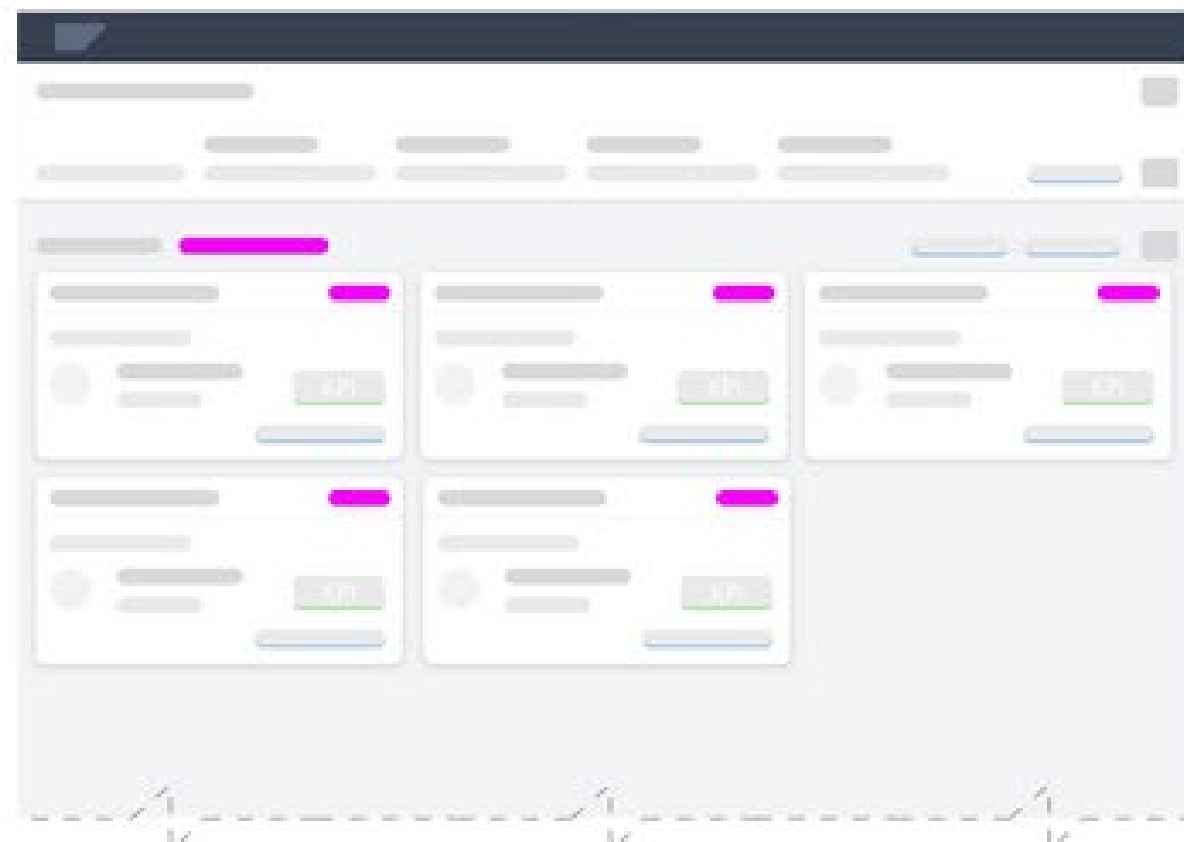


Level 1

Explanation indicator



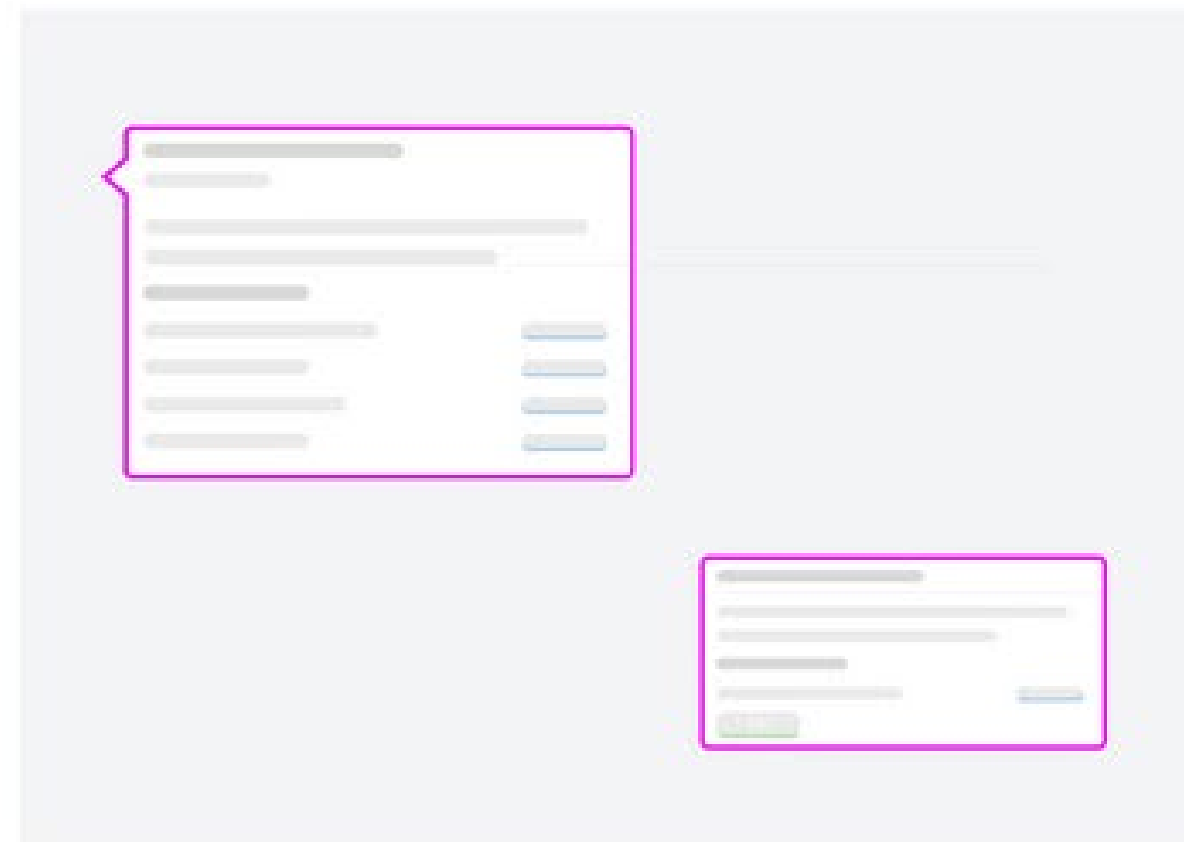
Global / local indicator in lists and tables



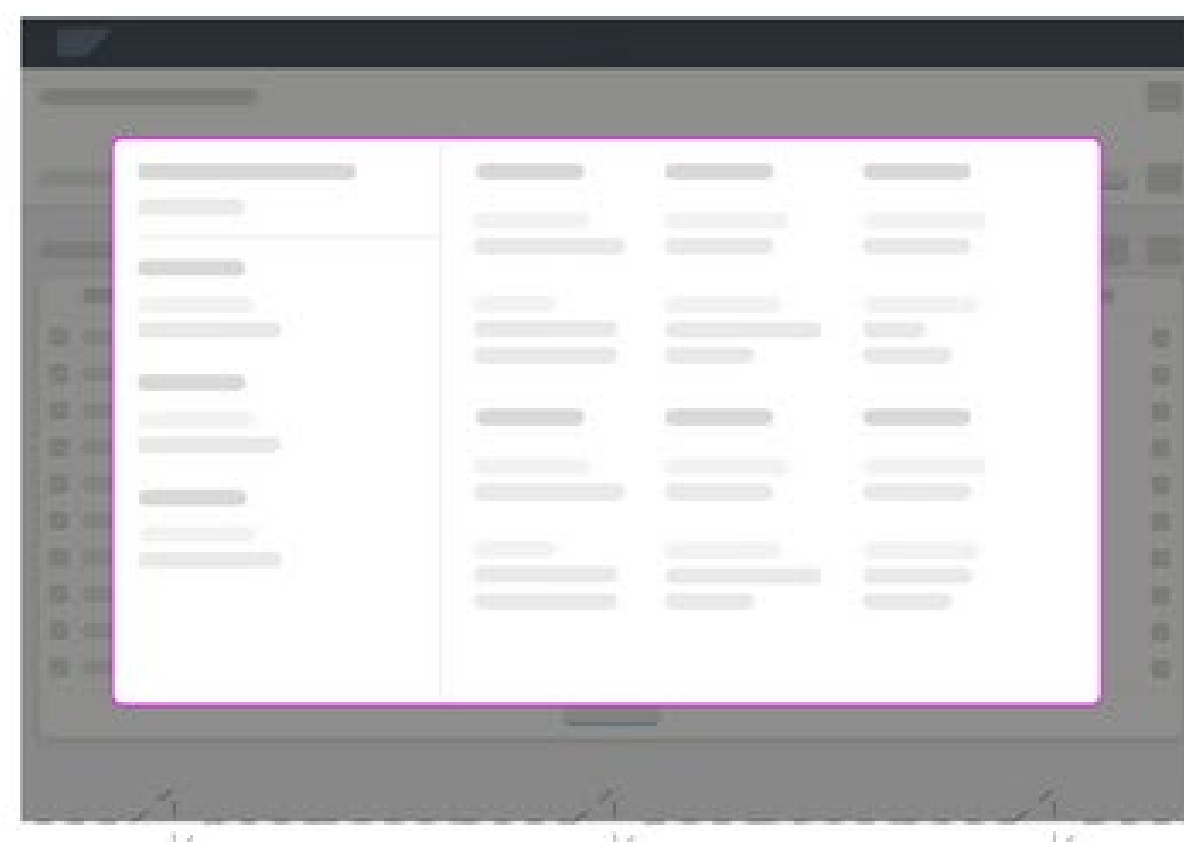
Global / local indicator for cards

Level 2

Simple explanation



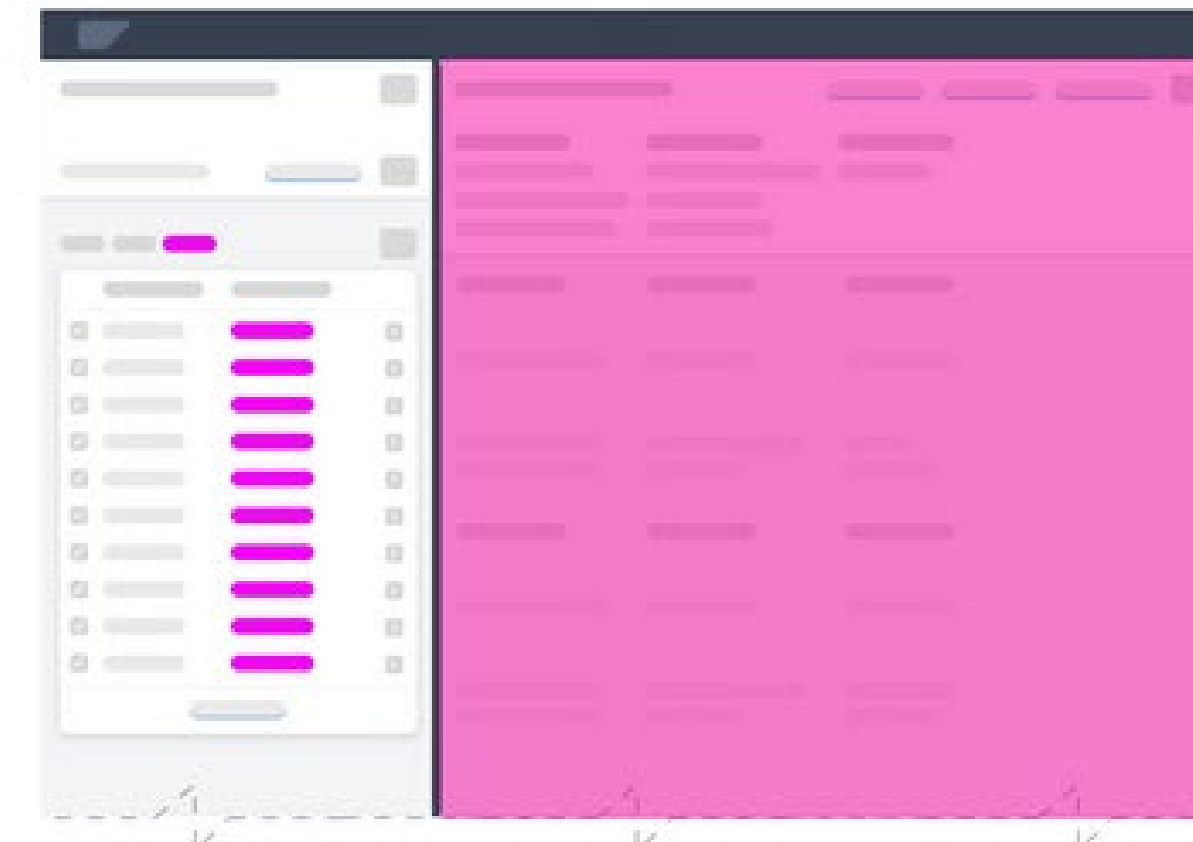
Explanation popover, conversation items



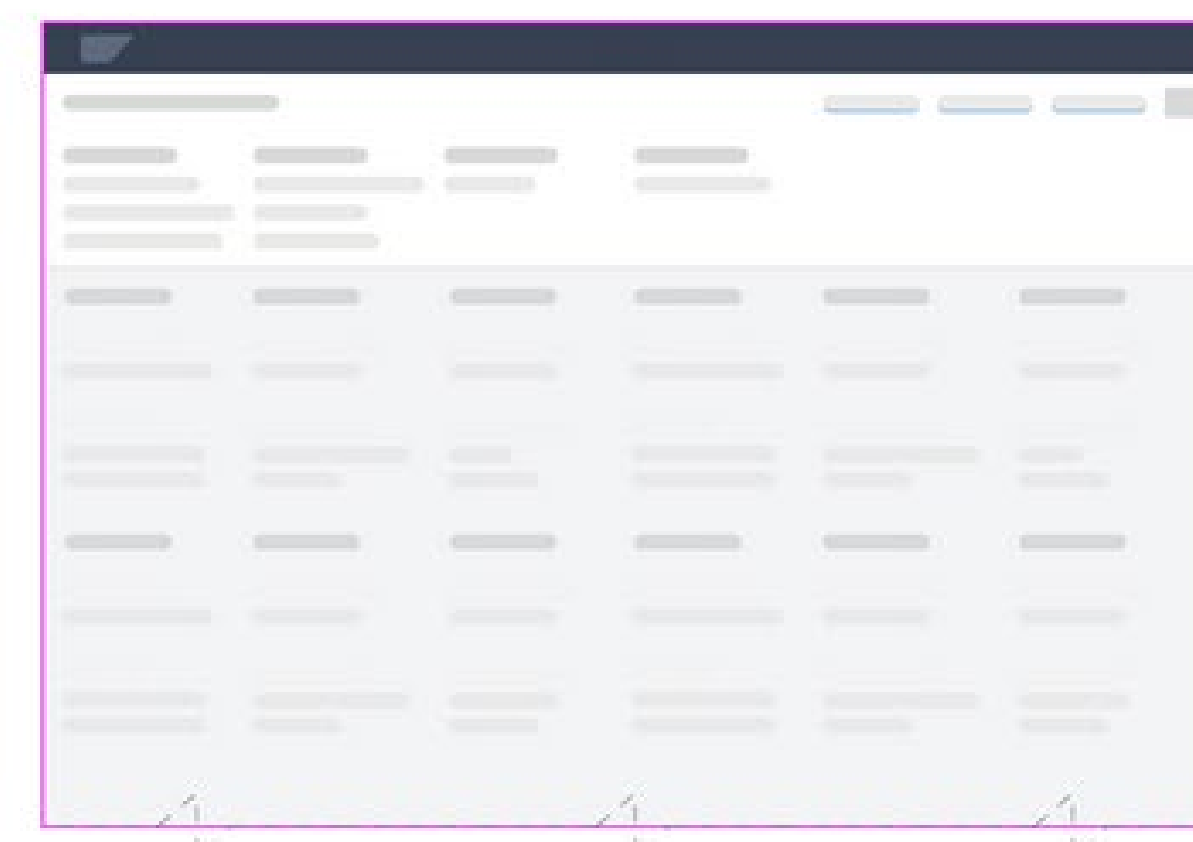
Overlay

Level 3

Extended explanation



Explanation page



Dedicated explanation application

<https://experience.sap.com/fiori-design-web/explainable-ai/>

CONTENT 11

Types of Explainability

- **Validating Models:** Approaches that aim to eliminate bias in the training data
- **Debugging Models:** Provide insights on wrong predictions
- **Knowledge Discovery:** Provide new insights through data analysis

Sarah Theres Völkel. Explainable AI. Introduction to Intelligent User Interfaces

CONTENT 11

Validating Models



Source: Brandon Messner | Unsplash

Classified as Dog



Source: Jose Carls Ichiro | Unsplash

Classified as Wolf

From Sarah Theres Völkel. Explainable AI. Introduction to Intelligent User Interfaces

CONTENT 11

Validating Models



Classified as Wolf



LIME-Explanation (idealised)

[Ribeiro et al. 2016]

From Sarah Theres Völkel. Explainable AI. Introduction to Intelligent User Interfaces

CONTENT 11

How can we design effective explanations?

- Provide applications that explain the “why” behind a recommendation?
- Are these explanations helpful?

CONTENT 11

What are the main challenges from an HCI perspective?

CONTENT 11

What are the main challenges from an HCI perspective?

- Help users to understand the information behind the model by helping them build correct mental models
- Help users to increase their trust towards the model
- Help users to make corrections to the model

Sarah Theres Völkel. Explainable AI. Introduction to Intelligent User Interfaces

CONTENT 11

Explanations



Why is this message in spam? It is similar to messages that were identified as spam in the past.

Report as not spam

CONTENT 11

TV Shows Genres

MAID DOWN TO EARTH WITH ZAC EFRON

Breaking Bad Dead to Me

ON ABBEY IS IT CAKE?

GOOD GIRLS Gilmore Girls

SUITS

Peaky Blinders

95% Match 2022 16+ 6 Seasons HD AD

Cast: Cillian Murphy, Sam Neill, Helen McCrory, more

Genres: Crime TV Shows, British, Period Pieces

This show is: Violent

Season 1

Episodes

1 Episode 1 57m

Ambitious gang leader Thomas Shelby recognizes an opportunity to move up in the world thanks to a missing crate of guns.

CONTENT 11

The screenshot displays a streaming service interface. At the top left, there's a 'TV Shows' header with a 'Genres' dropdown. Below it is a grid of show thumbnails including 'GOOD GIRLS', 'Gilmore', 'SUITS', 'Say I DO', and 'SKIN DEEP BEFORE & AFTER'. The central focus is the 'PEAKY BLINDERS' series page, which features a large background image of a man in a trench coat. A 'Love this!' tooltip is visible over the 'Like' button. The page includes a 'Play' button, a '95% Match' indicator, and details such as '2022', '16+', '6 Seasons', 'HD', and 'AD'. A description reads: 'A notorious gang in 1919 Birmingham, England, is led by the fierce Tommy Shelby, a crime boss set on moving up in the world no matter the cost.' The 'Cast' list includes Cillian Murphy, Sam Neill, and Helen McCrory. Genres listed are 'Crime TV Shows, British, Period Pieces'. A warning says 'This show is: Violent'. The 'Episodes' section shows 'Season 1' selected, with 'Episode 1' (57m) listed as 'Ambitious gang leader Thomas Shelby recognizes an opportunity to move up in the world thanks to a missing crate of guns.' On the right, a 'Suggestions For You' section shows thumbnails for 'MAID', 'DOWN TO EARTH WITH ZAC EFRON', 'Breaking Bad', 'Dead to Me', 'ON ABBEY', and 'IS IT CAKE?'.

CONTENT 11


The screenshot displays a streaming service interface. On the left, a grid of TV show thumbnails is visible, including 'Good Girls', 'Suits', 'Dead to Me', and 'Is It Cake?'. The central focus is on the TV show 'Peaky Blinders'. A 'Remove from My List' button is highlighted with a red box. Below it are 'Play', 'Checkmark', and 'Share' icons. The show's details include a '95% Match' rating, '2022' release year, '16+' rating, '6 Seasons', and 'HD AD' quality. The description reads: 'A notorious gang in 1919 Birmingham, England, is led by the fierce Tommy Shelby, a crime boss set on moving up in the world no matter the cost.' The cast list includes Cillian Murphy, Sam Neill, and Helen McCrory. Genres are listed as 'Crime TV Shows, British, Period Pieces'. A warning 'This show is: Violent' is present. The 'Episodes' section shows 'Season 1' selected, with 'Episode 1' (57m) listed as 'Ambitious gang leader Thomas Shelby recognizes an opportunity to move up in the world thanks to a missing crate of guns.'

CONTENT 11


People who viewed this also viewed

Page 1 of 8

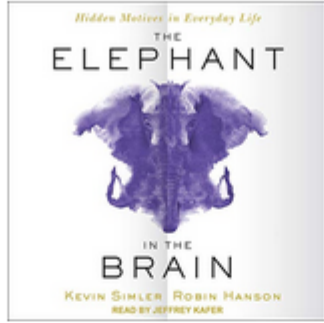
<




The Pragmatic Programmer: 20th Anniversary Edition, 2nd Edition
David Thomas
★★★★★ 1,903
#1 Best Seller in Software Testing
Audible Audiobook
\$0.00 Free with Audible trial




Elon Musk: Tesla, SpaceX, and the Quest for a Fantastic Future
Ashlee Vance
★★★★★ 16,466
#1 Best Seller in Engineering Patents & Inventions
Audible Audiobook
\$0.00 Free with Audible trial



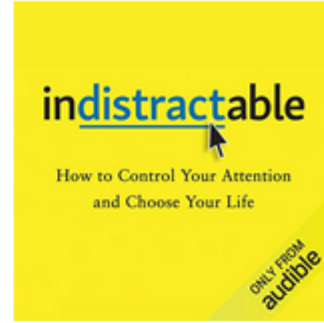
The Elephant in the Brain: Hidden Motives in Everyday Life
Kevin Simler
★★★★★ 656
Audible Audiobook
\$0.00 Free with Audible trial




Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems
Martin Kleppmann
★★★★★ 2,975
#1 Best Seller in Enterprise Data Computing
Audible Audiobook
\$0.00 Free with Audible trial



Clean Code: A Handbook of Agile Software Craftsmanship
Robert C. Martin
★★★★★ 4,337
Audible Audiobook
\$0.00 Free with Audible trial



Indistractable: How to Control Your Attention and Choose Your Life
Nir Eyal
★★★★★ 3,896
Audible Audiobook
\$0.00 Free with Audible trial



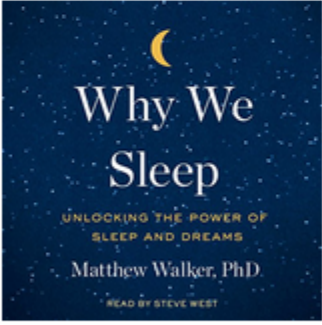
AI 2041: Ten Visions for Our Future
Kai-Fu Lee
★★★★★ 818
Audible Audiobook
\$0.00 Free with Audible trial

>


People who bought this also bought

Page 1 of 6

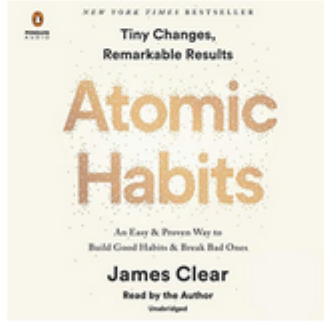
<



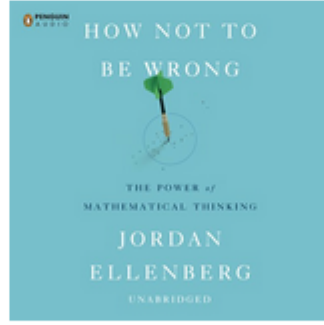
Why We Sleep: Unlocking the Power of Sleep and Dreams
Matthew Walker
★★★★★ 18,876
#1 Best Seller in Sleep Disorders
Audible Audiobook
\$0.00 Free with Audible trial




The Complete Software Developer's Career Guide: How to Learn Programming, Get a Job, and Advance Your Career
John Sonmez
★★★★★ 842
Audible Audiobook
\$0.00 Free with Audible trial



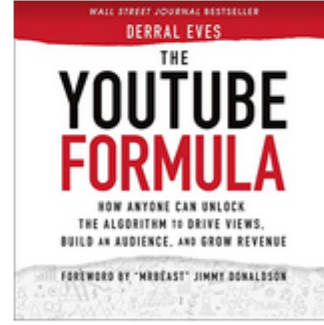
Atomic Habits: An Easy & Proven Way to Build Good Habits & Break Bad Ones
James Clear
★★★★★ 100,351
#1 Best Seller in Medical Social Psychology & Interactions
Audible Audiobook
\$0.00 Free with Audible trial



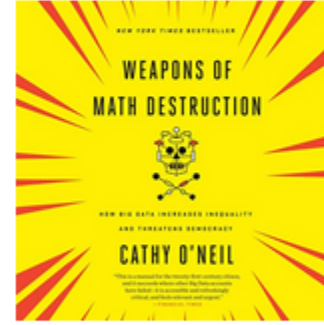
How Not to Be Wrong: The Power of Mathematical Thinking
Jordan Ellenberg
★★★★★ 2,668
Audible Audiobook
\$0.00 Free with Audible trial



Building Microservices: Designing Fine-Grained Systems
Sam Newman
★★★★★ 203
Audible Audiobook
\$0.00 Free with Audible trial



The YouTube Formula: How Anyone Can Unlock the Algorithm to Drive Views, Build an Audience, and Grow Revenue
Derral Eves
★★★★★ 565
Audible Audiobook
\$0.00 Free with Audible trial



Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy
Cathy O'Neil
★★★★★ 2,959
Audible Audiobook
\$0.00 Free with Audible trial

>



CONTENT 11

Input	<ul style="list-style-type: none"> • 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? 	Why	<ul style="list-style-type: none"> • 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/how is this instance NOT predicted...? • Why is this instance predicted P instead of Q? • Why are [instance A and B] given different predictions?
Output	<ul style="list-style-type: none"> • 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) ? 	Why not	<ul style="list-style-type: none"> • 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]?
Performance	<ul style="list-style-type: none"> • 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... 	What If	<ul style="list-style-type: none"> • 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 (global)	<ul style="list-style-type: none"> • How does the system make predictions? • What features does the system consider? <ul style="list-style-type: none"> • * Is [feature X] used or not used for the predictions? • What is the system's overall logic? <ul style="list-style-type: none"> • 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? <ul style="list-style-type: none"> • * How are the parameters set? 	How to be that	<ul style="list-style-type: none"> • 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?
		How to still be this	<ul style="list-style-type: none"> • * 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)
		Others	

Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–15. DOI:<https://doi.org/10.1145/3313831.3376590>

CONTENT 11

Explanations

- Explanations are **contrastive**: Why C instead of Y?
- Explanations are **selective**: Show the most important information that contributed to a decision (at the cost of completeness)
- Explanations are **credible**: Be consistent with users' prior knowledge
- Explanations are **conversational**: Who reads an explanation? Allow users to raise queries

From Sarah Theres Völkel. Explainable AI. Introduction to Intelligent User Interfaces

CONTENT 11

- Why not book Z?
- Why not book X first?
- What would happen if?

People who viewed this also viewed

Page 1 of 8

A screenshot of an Audible recommendation carousel. The main book is 'The Pragmatic Programmer: 20th Anniversary Edition, 2nd Edition' by David Thomas, a #1 Best Seller in Software Testing. The carousel shows seven other audiobooks:

- Elon Musk: Tesla, SpaceX, and the Quest for a Fantastic Future** by Ashlee Vance (#1 Best Seller in Engineering Patents & Inventions)
- The Elephant in the Brain: Hidden Motives in Everyday Life** by Kevin Simler
- Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems** by Martin Kleppmann (#1 Best Seller in Enterprise Data Computing)
- Clean Code: A Handbook of Agile Software Craftsmanship** by Robert C. Martin
- Indistractable: How to Control Your Attention and Choose Your Life** by Nir Eyal
- AI 2041: Ten Visions for Our Future** by Kai-Fu Lee

People who bought this also bought

Page 1 of 6

A screenshot of an Audible recommendation carousel. The main book is 'Why We Sleep: Unlocking the Power of Sleep and Dreams' by Matthew Walker, a #1 Best Seller in Sleep Disorders. The carousel shows six other audiobooks:

- The Complete Software Developer's Career Guide: How to Learn Practical Skills and Advance Your Career** by John Sonmez
- Atomic Habits: An Easy & Proven Way to Build Good Habits & Break Bad Ones** by James Clear (#1 Best Seller in Medical Social Psychology & Interactions)
- How Not to Be Wrong: The Power of Mathematical Thinking** by Jordan Ellenberg
- Building Microservices: Designing Fine-Grained Systems** by Sam Newman
- The YouTube Formula: How Anyone Can Unlock the Algorithm to Drive Views, Build an Audience, and Grow Revenue** by Derral Eves
- Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy** by Cathy O'Neil



CONTENT 11

Discussion

- Choose an AI system you interact with
- How would you improve explanations about the system's recommendations?

CONTENT 11

Sources

- Sarah Theres Völkel. Explainable AI. Introduction to Intelligent User Interfaces
- Introduction to eXplainable AI (XAI) Q. Vera Liao, Moninder Singh, Yunfeng Zhang, Rachel Bellamy. ACM CHI 2021 Course on Intro to Explainable AI
<https://hcixaitutorial.github.io/>

MAI4CAREU

Master programmes in Artificial
Intelligence 4 Careers in Europe

Thank you.