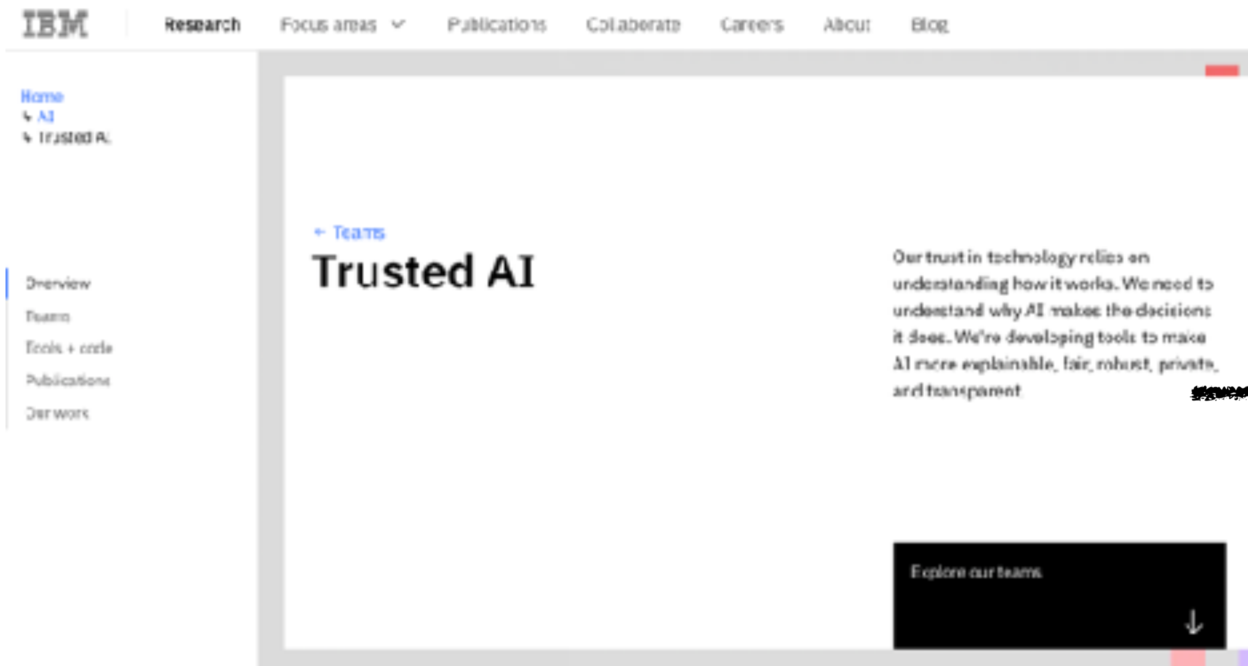
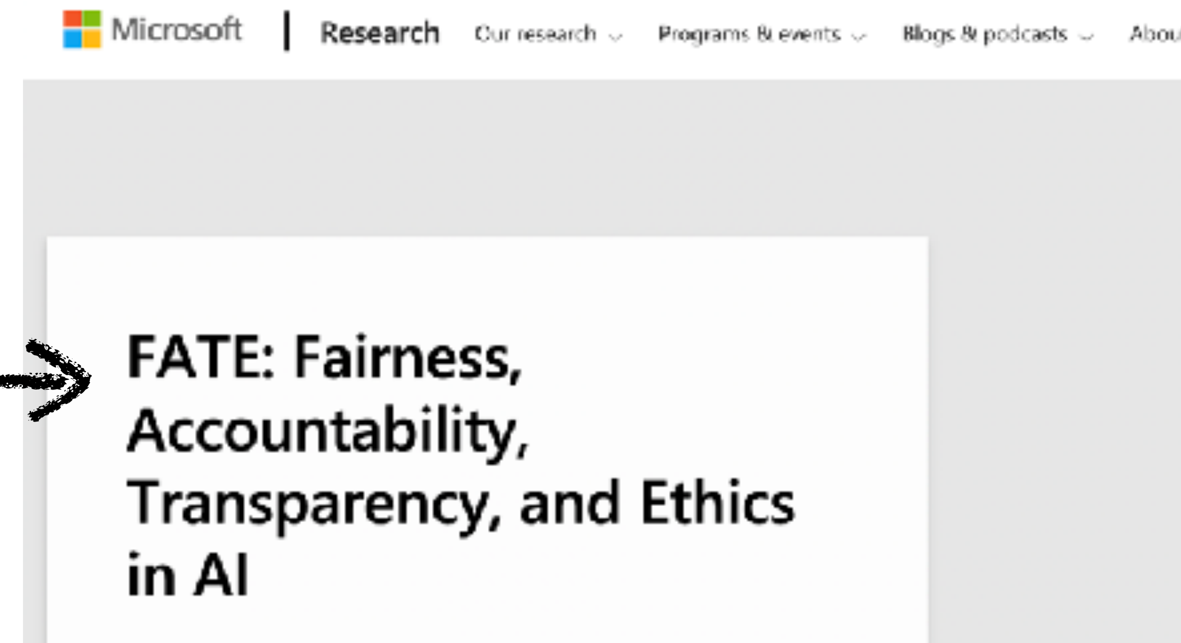
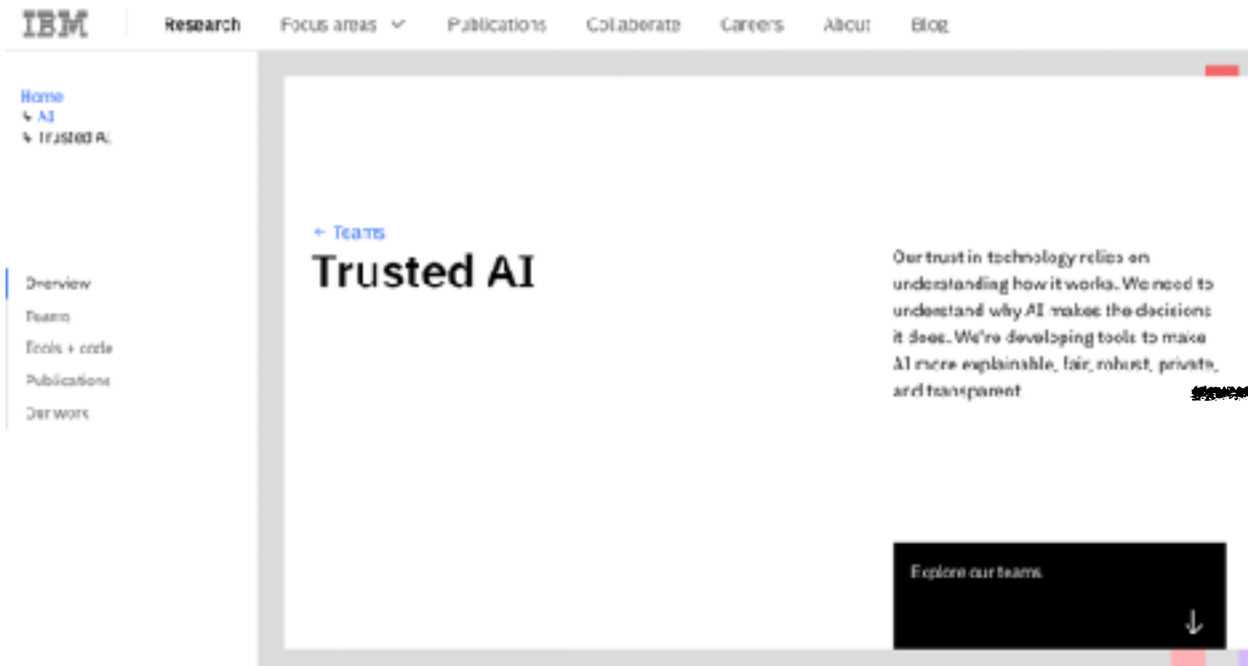
Human-Centered Explainable AI (XAI): From Algorithms to User Experiences

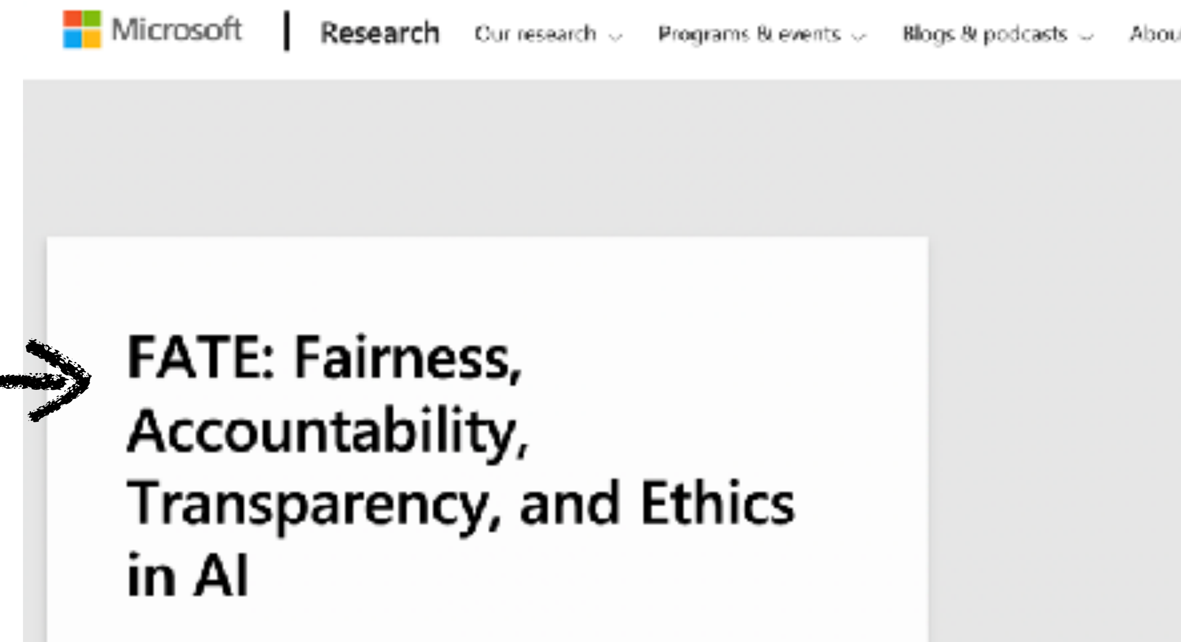
Q. Vera Liao

Microsoft Research


IBM
Research
Focus areas
Publications
Collaborate
Careers
About
Blog
Home
AI
Trusted AI
Overview
Teams
Tools + code
Publications
Our work
← Teams
Trusted AI
Our trust in technology relies on understanding how it works. We need to understand why AI makes the decisions it does. We're developing tools to make AI more explainable, fair, robust, private, and transparent.
Explore our teams


Microsoft
Research
Our research
Programs & events
Blogs & podcasts
About
FATE: Fairness, Accountability, Transparency, and Ethics in AI





Human-Centered Explainable AI (XAI): From Algorithms to User Experiences

Q. VERA LIAO*, Microsoft Research, Canada
KUSH R. VARSHNEY, IBM Research, United States

of artificial intelligence (AI), explainable AI (XAI) has produced a vast collection practitioners to build XAI applications. With the rich application opportunities, entists or researchers to comprehend the models they are developing, to become ot AI deployed in numerous domains. However, explainability is an inherently brace human-centered approaches. Human-computer interaction (HCI) research ning increasingly important. In this chapter, we begin with a high-level overview ectively survey our own and other recent HCI works that take human-centered id methodological tools for XAI. We ask the question "*what are human-centered* it they play in shaping XAI technologies by helping navigate, assess and expand xplainability needs, to uncover pitfalls of existing XAI methods and inform new iuman-compatible XAI.

An overview of recent HCI works on XAI

https://arxiv.org/abs/2110.10790

What are human-centered approaches doing for XAI?

Human-Centered Explainable AI (XAI): From Algorithms to User Experiences

Q. VERA LIAO*, Microsoft Research, Canada
KUSH R. VARSHNEY, IBM Research, United States

(Book Chapter Draft 10/2021) As a technical sub-field of artificial intelligence (AI), explainable AI (XAI) has produced a vast collection of algorithms, providing a toolbox for researchers and practitioners to build XAI applications. With the rich application opportunities, explainability has moved beyond a demand by data scientists or researchers to comprehend the models they are developing, to become an essential requirement for people to trust and adopt AI deployed in numerous domains. However, explainability is an inherently human-centric property and the field is starting to embrace human-centered approaches. Human-computer interaction (HCI) research and user experience (UX) design in this area are becoming increasingly important. In this chapter, we begin with a high-level overview of the technical landscape of XAI algorithms, then selectively survey our own and other recent HCI works that take human-centered approaches to design, evaluate, provide conceptual and methodological tools for XAI. We ask the question "*what are human-centered approaches doing for XAI*" and highlight three roles that they play in shaping XAI technologies by helping navigate, assess and expand the XAI toolbox: to drive technical choices by users' explainability needs, to uncover pitfalls of existing XAI methods and inform new methods, and to provide conceptual frameworks for human-compatible XAI.

The quest for explainable AI (XAI)

Companies Grapple With AI's Opaque Decision-Making Process

We Need AI That Is Explainable, Auditable, and Transparent

Why "Explainability" Is A Big Deal In AI

From black box to white box: Reclaiming human power in AI

How Explainable AI Is Helping Algorithms Avoid Bias

A large collection of **XAI algorithms:** aiming to make models understandable

Explaining Explanations: An Overview of Interpretability of Machine Learning

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{lgilpin, davidbau, bzy, abajwa, specter, lkagal}@ mit.edu

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received August 5, 2018, accepted September 4, 2018, date of publication September 17, 2018, date of current version October 12, 2018.
Digital Object Identifier 10.1109/ACCESS.2018.2870052

Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

AMINA ADADI AND MOHAMMED BERRADA
Computer and Interdisciplinary Physics Laboratory, Sidi Mohammed Ben Abdellah University, Fez 30050, Morocco
Corresponding author: Amina Adadi (amina.adadi@gmail.com)

electronics MDPI

Review

Machine Learning Interpretability: A Survey on Methods and Metrics

Diogo V. Carvalho [1,2,*], Eduardo M. Pereira [1] and Jaime S. Cardoso [2,3]

Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI

A Survey of Methods for Explaining Black Box Models

RICCARDO GUIDOTTI, ANNA MONREALE, SALVATORE RUGGIERI, and FRANCO TURINI, KDDLab, University of Pisa, Italy
FOSCA GIANNOTTI, KDDLab, ISTI-CNR, Italy
DINO PEDRESCHI, KDDLab, University of Pisa, Italy

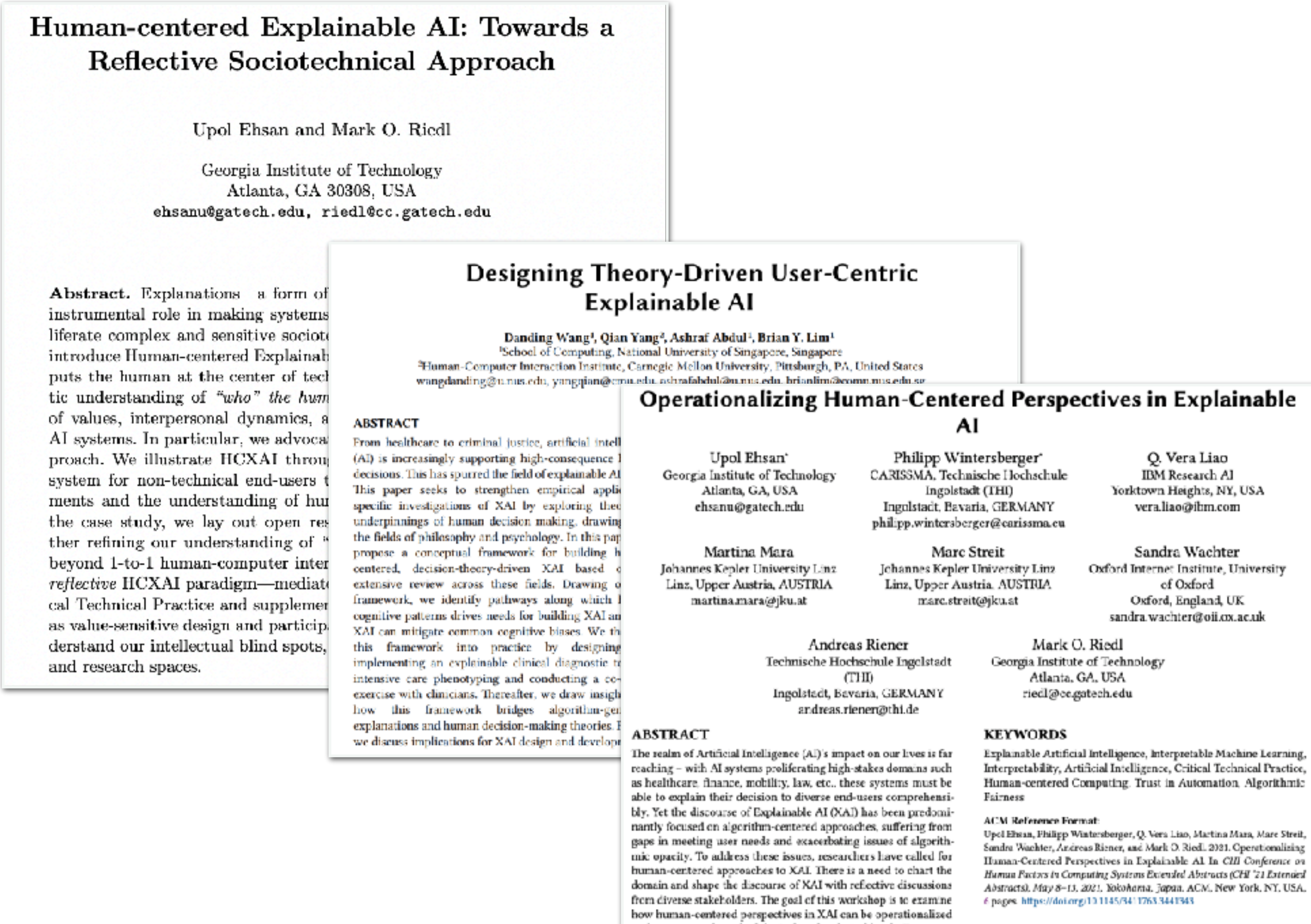
In recent years, many accurate decision support systems have been constructed as black boxe systems that hide their internal logic to the user. This lack of explanation constitutes both a prac ethical issue. The literature reports many approaches aimed at overcoming this crucial weakness at the cost of sacrificing accuracy for interpretability. The applications in which black box decisi can be used are various, and each approach is typically developed to provide a solution for a speci and, as a consequence, it explicitly or implicitly delineates its own definition of interpretability a tion. The aim of this article is to provide a classification of the main problems addressed in the lite respect to the notion of explanation and the type of black box system. Given a problem definit box type, and a desired explanation, this survey should help the researcher to find the proposals

Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*

Gabriëlle Ras, Marcel van Gerven, Pim Haselager
Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands
{g.ras, m.vangerven, w.haselager}@donders.ru.nl

Abstract

Issues regarding explainable AI involve four components: users, laws & regulations, explanations and algorithms. Together these components provide a context in which explanation methods can be evaluated regarding their adequacy. The goal of this chapter is to bridge the gap between expert users and lay users. Different kinds of users are identified and their concerns revealed, relevant statements from the General Data Protection Regulation are analyzed in the context of Deep Neural Networks (DNNs), a taxonomy for the classification of existing



Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach

Upol Ehsan and Mark O. Riedl

Georgia Institute of Technology
Atlanta, GA 30308, USA
ehsanu@gatech.edu, riedl@cc.gatech.edu

Abstract. Explanations a form of
instrumental role in making systems
liferate complex and sensitive sociot
introduce Human-centered Explainab
puts the human at the center of tech
tic understanding of *"who" the hum*
of values, interpersonal dynamics, a
AI systems. In particular, we advoca
proach. We illustrate HCXAI throu
system for non-technical end-users t
ments and the understanding of hu
the case study, we lay out open res
ther refining our understanding of "
beyond 1-to-1 human-computer inter
reflective HCXAI paradigm—mediat
cal Technical Practice and supplemen
as value-sensitive design and particip
derstand our intellectual blind spots,
and research spaces.

Designing Theory-Driven User-Centric Explainable AI

Danding Wang[1], Qian Yang[2], Ashraf Abdul[1], Brian Y. Lim[1]
[1]School of Computing, National University of Singapore, Singapore
[2]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, United States
wangdanding@u.nus.edu, yangqian@cmu.edu, ashrafabdul@u.nus.edu, brianlim@comp.nus.edu.sg

ABSTRACT
From healthcare to criminal justice, artificial intell
(AI) is increasingly supporting high-consequence
decisions. This has spurred the field of explainable AI
This paper seeks to strengthen empirical appli
specific investigations of XAI by exploring theo
underpinnings of human decision making, drawing
the fields of philosophy and psychology. In this pap
propose a conceptual framework for building h
centered, decision-theory-driven XAI based o
extensive review across these fields. Drawing o
framework, we identify pathways along which
cognitive patterns drives needs for building XAI an
XAI can mitigate common cognitive biases. We th
this framework into practice by designing
implementing an explainable clinical diagnostic to
intensive care phenotyping and conducting a co-
exercise with clinicians. Thereafter, we draw insigh
how this framework bridges algorithm-gen
explanations and human decision-making theories.
we discuss implications for XAI design and develop

Operationalizing Human-Centered Perspectives in Explainable AI

Upol Ehsan*
Georgia Institute of Technology
Atlanta, GA, USA
ehsanu@gatech.edu

Philipp Wintersberger*
CARISSMA, Technische Hochschule Ingolstadt (THI)
Ingolstadt, Bavaria, GERMANY
philipp.wintersberger@carissma.eu

Q. Vera Liao
IBM Research AI
Yorktown Heights, NY, USA
vera.liao@ibm.com

Martina Mara
Johannes Kepler University Linz
Linz, Upper Austria, AUSTRIA
martina.mara@jku.at

Marc Streit
Johannes Kepler University Linz
Linz, Upper Austria, AUSTRIA
marc.streit@jku.at

Sandra Wachter
Oxford Internet Institute, University of Oxford
Oxford, England, UK
sandra.wachter@oii.ox.ac.uk

Andreas Riener
Technische Hochschule Ingolstadt (THI)
Ingolstadt, Bavaria, GERMANY
andreas.riener@thi.de

Mark O. Riedl
Georgia Institute of Technology
Atlanta, GA, USA
riedl@cc.gatech.edu

ABSTRACT
The realm of Artificial Intelligence (AI)'s impact on our lives is far reaching – with AI systems proliferating high-stakes domains such as healthcare, finance, mobility, law, etc., these systems must be able to explain their decision to diverse end-users comprehensibly. Yet the discourse of Explainable AI (XAI) has been predominantly focused on algorithm-centered approaches, suffering from gaps in meeting user needs and exacerbating issues of algorithmic opacity. To address these issues, researchers have called for human-centered approaches to XAI. There is a need to chart the domain and shape the discourse of XAI with reflective discussions from diverse stakeholders. The goal of this workshop is to examine how human-centered perspectives in XAI can be operationalized

KEYWORDS
Explainable Artificial Intelligence, Interpretable Machine Learning, Interpretability, Artificial Intelligence, Critical Technical Practice, Human-centered Computing, Trust in Automation, Algorithmic Fairness

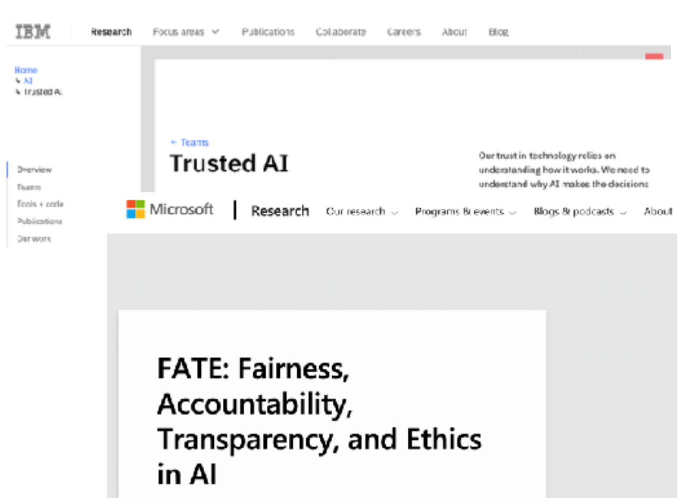
ACM Reference Format:
Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3411763.3441342

What are human-centered approaches doing for XAI?

- How to make XAI human-centered?
- What are the current trends and important problems?
- How should AI and HCI communities work together?

My lenses

(Cognitive) human-computer interaction



Intersecting with **AI researchers** and **practitioners**

IBM Research Trusted AI

Home | Demo | Resources | Events | Videos | Community

AI Explainability 360 Open Source Toolkit

This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. Containing eight state-of-the-art algorithms for interpretable machine learning as well as metrics for explainability, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs ↗ | Get Code ↗

Not sure what to do first? Start here!

Read More

Learn more about explainability concepts, terminology, and tools before you begin.

→

Try a Web Demo

Step through the process of explaining models to consumers with different personas in an interactive web demo that shows a sample of capabilities available in this toolkit.

→

Watch Videos

Watch videos to learn more about AI Explainability 360 toolkit.

→

Read a Paper

Read a paper describing how we designed AI Explainability 360 toolkit.

→

Use Tutorials

Step through a set of in-depth examples that introduce developers to code that explains data and models in different industry and application domains.

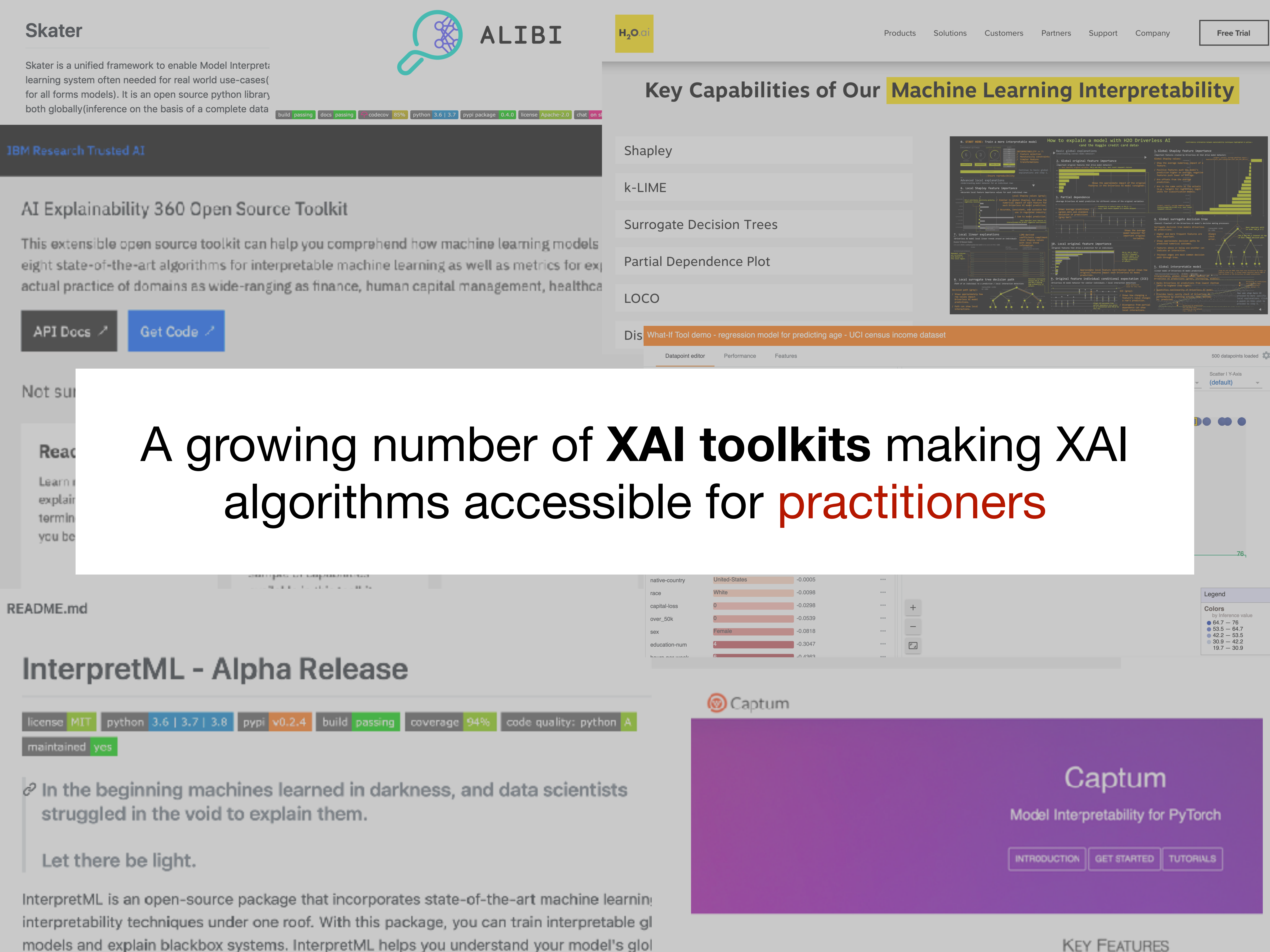
→

Ask

Join c
360 S
quest
and t
you u

→

https://aix360.mybluemix.net/



A growing number of **XAI toolkits** making XAI algorithms accessible for practitioners

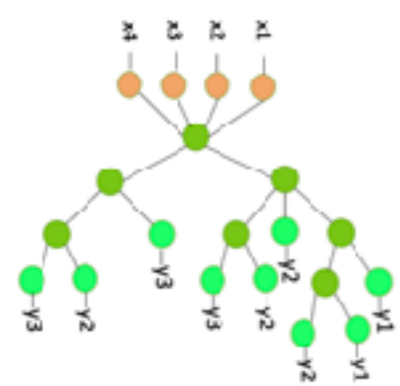
XAI

Directly interpretable model

Generalized Linear Rule Model

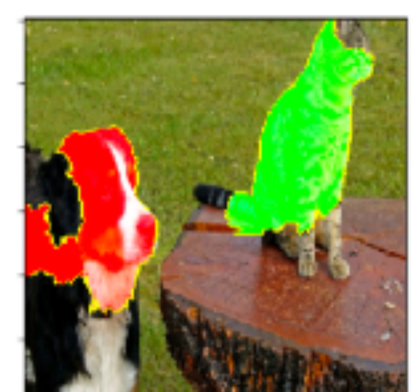
Post-hoc explainability

Explaining the model (global)



Model distillation

Explaining a decision (local)



Feature importance

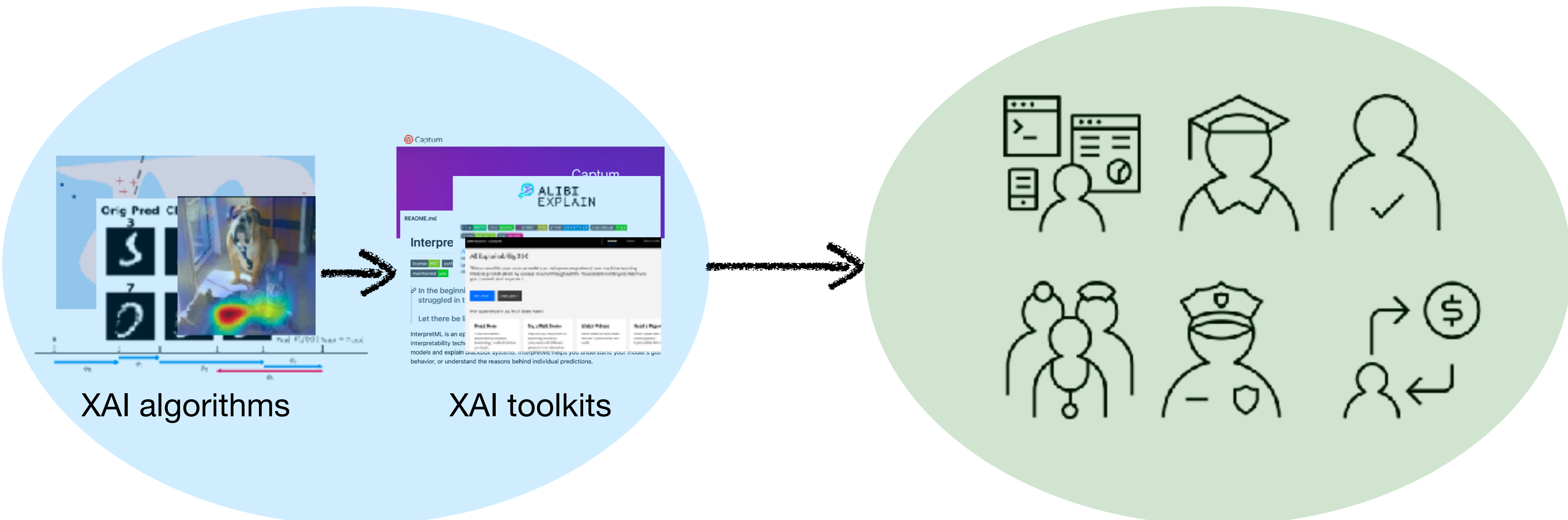
Inspecting counterfactual

- If **{debt percentage under 30%}**, you will no longer be predicted of high risk

Counterfactual explanation

Check out our **CHI2021 Course** materials, with links to AIX360 code libraries: https://hcixaitutorial.github.io/

HCXAI: bridging work from XAI algorithms to user experiences



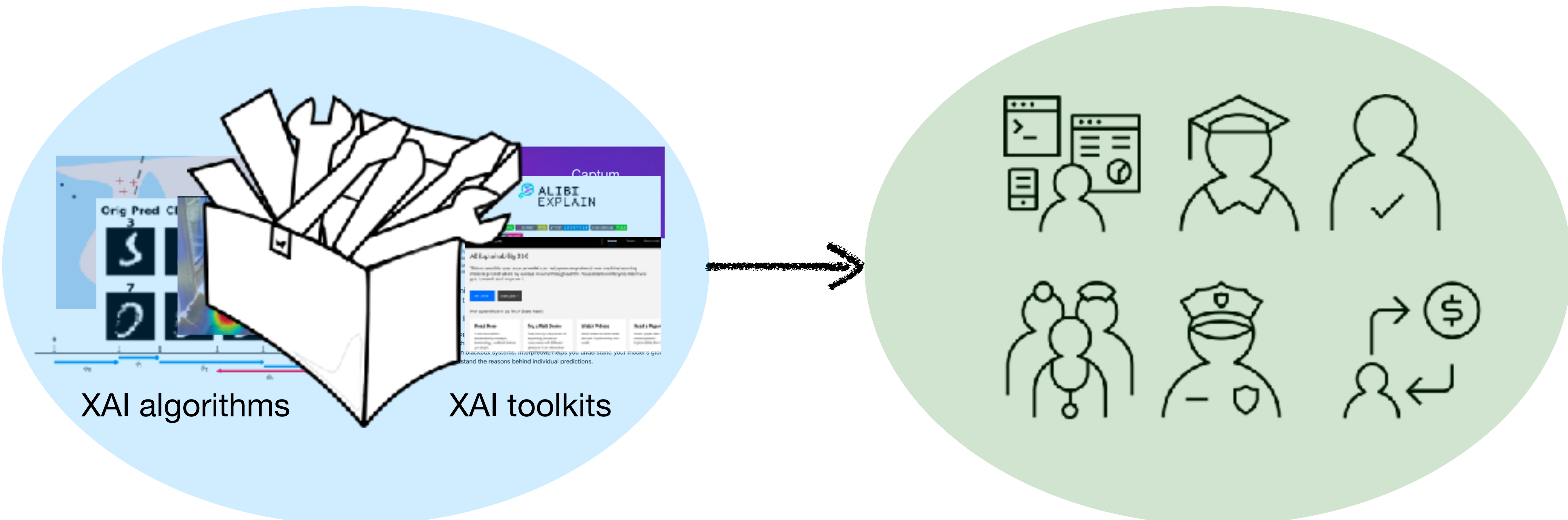
XAI techniques

Real-world XAI systems?

Built by practitioners

Serving many domains and user groups

HCXAI: bridging work from XAI algorithms to user experiences



A toolbox of XAI techniques

Real-world XAI systems?
Built by *practitioners*
Serving many domains and user groups

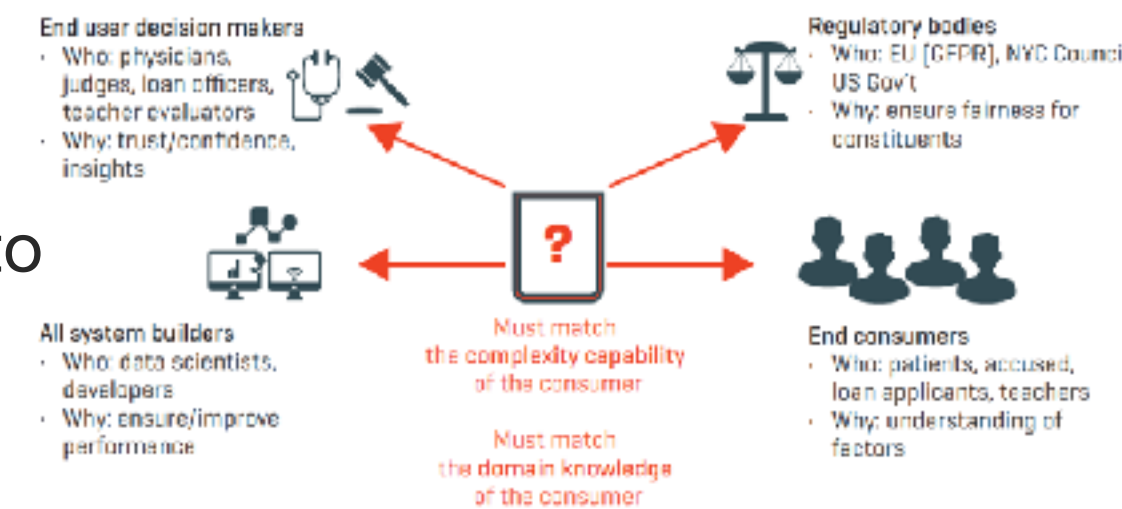
What are human-centered approaches doing for XAI?

- **Navigate** the toolbox: Drive technical choices by users' explainability needs
- **Assess** the toolbox: Uncover pitfalls of existing XAI methods through empirical studies
- **Expand** the toolbox: Inform new methods and conceptual frameworks for human-compatible XAI

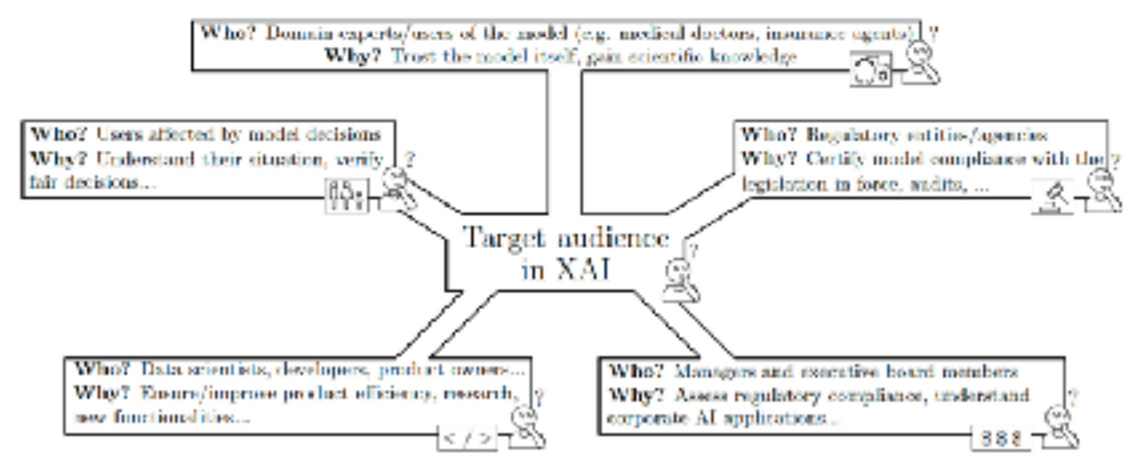
Navigate the toolbox: **Characterizing the space of users' explainability needs**

Who are the prototypical users of XAI?

- **Model developers**, to improve or debug the model.
- **Decision-makers**, who are direct users, to make informed decisions.
- **Impacted groups**, whose life could be impacted by the AI, to seek recourse or contest the AI.
- **Business owners or administrators**, to assess an AI application's capability, regulatory compliance, etc.
- **Regulatory bodies**, to audit for legal or ethical concerns such as fairness, safety, privacy, etc.

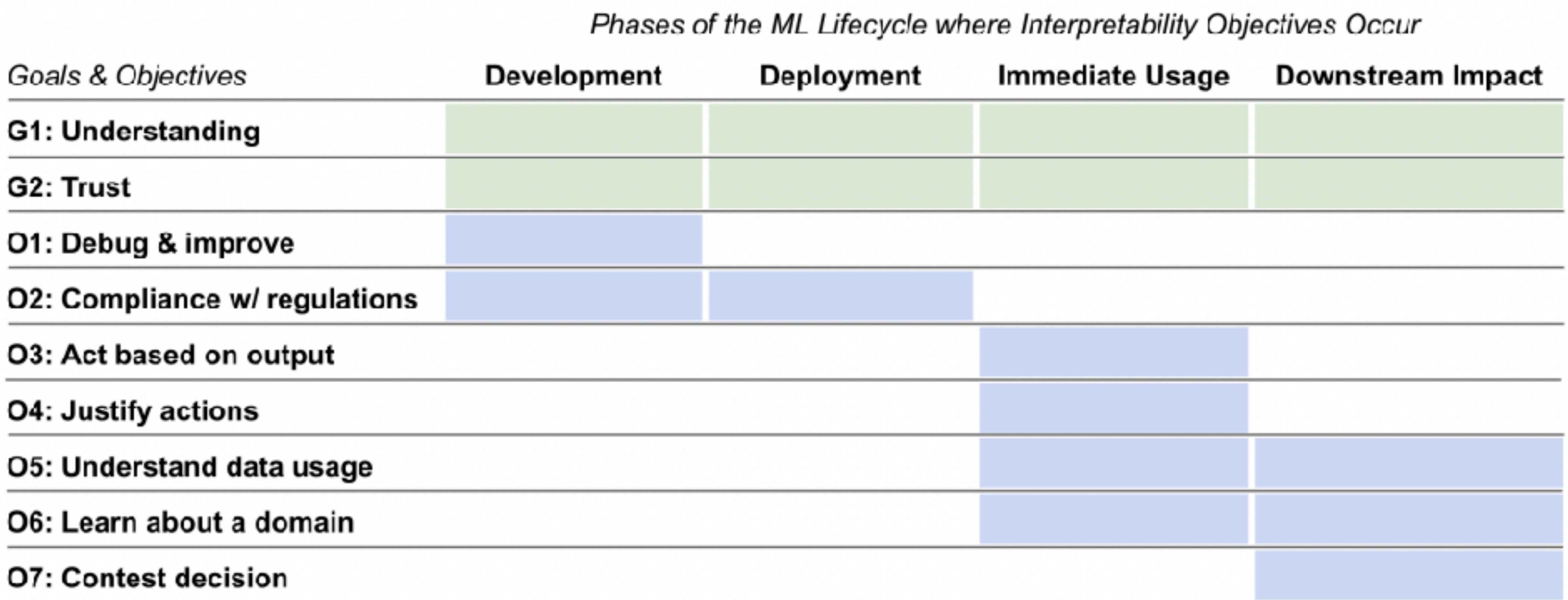


(Hind et al, 2019)



(Arrieta et al, 2019)

Persona is not enough: user objectives



Suresh et al. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. CHI 2021

Explainability needs expressed as questions

Task objectives	Users who may engage in this task	Example questions they may ask the AI
To improve or debug the model	Model Developers. Some applications would also allow other user groups to perform this task	- 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 user groups can 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?
To make informed decisions or take better actions	Decision-Makers, Impacted Groups, and more	- **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**? - **How to** make sure this instance **remains to be X**? What change is
To adapt usage or control	Decision-Makers, Business Owners, and more	- **How** does the AI make predictions? What can I supply or change for it to work well? - **What if** I make this change?
To learn new knowledge about a domain	Decision-Makers, Business Owners, Impacted Groups, and more	- **How** does the prediction task work? What are the key features to consider? - **What if** this feature changes? How does it impact the outcome? - **Why** is this instance **not** predicted to be Y as I would expect?
To ensure ethical or legal compliance	All user groups can engage in this task at some point	- **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?

Check out my blog post with IBM Data & AI

Lim and Dey. Toolkit to support intelligibility in context-aware applications. UbiComp 2010
Graesser et al. Question-driven explanatory reasoning. *Applied Cognitive Psychology (1996)*

Navigate the toolbox: User-centered **Question-Driven XAI Design**

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020
Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Working paper)

Where we started: Research into **XAI Design Practices**

Research questions:

- What is the design space of XAI UX?
- What are the design challenges?

Liao et al. [Questioning the AI: Informing Design Practices for Explainable AI User Experiences](). CHI 2020 🎖

Methodology

- Interviewed **20 designers** working on **16 AI products**

1. Walk through the AI system
2. Common questions users might ask
3. Discuss each question card
4. General challenges to create XAI products

Understanding input (training data): What kind of data does the system learn from?
- What is the *source* of the data?
- How are the *labels/ground-truth* produced?

Inspecting what if changing a case/counterfactual questions: what if, how to be that, how to still be this
- What would the system predict if the case changes to...?
- How should this case change to get a different prediction?
- What are the scope of changes permitted for this case to still get the same prediction?
- What kind of cases get a different/same prediction?

Understanding the model globally: How does the system make predictions (overall logic)?
- What algorithm is used?
- What *rules* does the system use to make predictions?
- *What features* does the model consider or not consider?
- How does the model *weigh/reason with these features*?

Understanding output: What kind of output/predictions does the system give?
- What does the system output *mean*?
- How can I use the output of the system?

Other category (add your own question)

Understanding prediction for a particular case: Why this? Why not that?
- Why is this case given this prediction? Why is it NOT predicted that?
- What *feature(s)* of this case lead to the model's prediction for it?
- *What kind of cases* are predicted this?
- Why are [cases A and B] given *the same prediction*?
- Why are [cases A and B] given *different predictions*?

Understanding model performance and certainty: How accurate/reliable are the system's predictions?
- *How often* does the system make mistakes?
- *When/under what situation* is the system likely to be correct/wrong?

XAI Algorithms

Opportunities for new methods

- Explain data limitations and generalizability
- Explain output of multiple models
- Explain system changes
- Multi-level global explanations
- Interactive counterfactual explanations
- Social explanations
- Personalized and adaptive explanations

XAI UX

Design guidelines to address user needs

Input: Provide comprehensive transparency of training data, especially the limitations

Output: Contextualize the system's output in downstream tasks and the users' overall workflow

Performance: Help users understand the limitations of the AI and make it actionable

Global model: Choose appropriate level of details to explain the model

Local: Distinguish between "why not" and "why"

Counterfactuals: Consider opportunities as utility features for analytics or exploration

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020

XAI Question Bank

Data

- **What kind of data was the system trained on?**
- What is the source of the training data?
- How were the labels/ground-truth produced?
- What is the sample size of the training data?
- What dataset(s) is the system NOT using?
- What are the potential limitations/biases of the data?
- What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

Output

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

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 model-wide explanation)

- **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 kind of rules does it follow?
  - How does [feature X] impact its predictions?
  - What are the top rules/features that determine its predictions?
- What kind of algorithm is used?
  - How were the parameters set?

Why

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

Why not

- **Why is this instance NOT predicted to be [a different outcome Q]?**
- Why is this instance predicted [P instead of a different outcome Q]?
- Why are [instance A and B] given different predictions?

How to be that
(a different prediction)

- **How should this instance change to get a different prediction Q?**
- What is the minimum change required for this instance to get a different prediction Q?
- How should a given feature change for this instance to get a differen prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

How to still be this
(the current prediction)

- **What is the scope of change permitted for this instance to still get the same prediction?**
- What is the range of value permitted for a given feature for this prediction to stay the same?
- What is the necessary feature(s)/feature-value(s) present or absent to guarantee this prediction?
- What kind of instance gets the same prediction?

What If

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

Others

- How/why will the system change/adapt/improve/drift over time? (change)
- Can I, and if so, how do I, improve the system? (improvement)
- Why is the system using or not using a given algorithm/feature/rule/dataset? (follow-up)
- What does [a machine learning terminology] mean? (terminological)
- What are the results of other people using the system? (social)

Liao et al. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI 2020

Question	Explanations	Example XAI techniques
Global how (global model-wide)	• Describe the general model logic as feature impact*, rules⁺ or decision-trees• (sometimes need to explain with a surrogate simple model) • If the user is only interested in a high-level view, describe what are the top features or rules considered	ProfWeight*⁺•, Global Feature Importance*, PDP*, DT Surrogate•
Why	• Describe how features of the instance, or what key features, determine the model's prediction of it* • Or describe rules⁺ that the instance fits to guarantee the prediction⁺ • Or show similar examples• with the same predicted outcome to justify the model's prediction	LIME*, SHAP*, LOCO*, Anchors⁺, ProtoDash•
Why not (a different prediction)	• Describe what features of the instance determine the current prediction and/or with what changes the instance would get the alternative prediction* • Or show prototypical examples⁺ that had the alternative outcome	CEM* , Counterfactuals⁺ , ProtoDash⁺ (on alternative prediction)
How to be that (a different prediction)	• Highlight feature(s) that if changed (increased, decreased, absent, or present) could alter the prediction to the alternative outcome, often with minimum effort required* • Or show examples with minimum differences but had the alternative outcome⁺	CEM*, Counterfactuals⁺, DiCE⁺
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*, Anchors⁺
What if	• Show how the prediction changes corresponding to the inquired change of input	PDP, ALE
Performance	• Provide performance metrics of the model • Show uncertainty information for each prediction • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC Uncertainty Qauntification 360 FactSheets, Model Cards
Data	• Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.	FactSheets, DataSheets
Output	• Describe the scope of output or system functions. • Suggest how the output should be used for downstream tasks or user workflow	FactSheets, Model Cards

Questions as *re-framing* the technical space of XAI

Questions as "*boundary objects*" supporting designer-engineer collaboration

Liao et al. Question-Driven Design Process for Explainable AI User Experiences. (Working paper)

Challenges for practitioners: “in the dark” design process

- Challenge **navigating the technical capabilities**

“ *finding the right pairing to put the ideas of what’s right for the user together with what’s doable given the tools or the algorithms*

- **Communication barriers** and **implementation cost** impeding buy-in from data scientists and the team

“ *It remains in this weird limbo where people know it's important. People see it happen. They don't know how to make it happen. And everybody's feeling their way in the dark with no lights.*

Liao et al. [Questioning the AI: Informing Design Practices for Explainable AI User Experiences](). CHI 2020

Question-Driven XAI Design

Step 1

Identify user questions

Elicit user needs for XAI as questions

Also gather user intentions and expectations for asking the questions

Designers, users

Step 2

Analyze questions

Cluster questions into categories and prioritize categories for the XAI UX to focus on

Summarize user intentions and expectations to identify key user requirements

Designers, product team

Step 3

Map questions to modeling solutions

Map prioritized question categories to candidate XAI techniques as a set of functional elements that the design should cover

A mapping guide for supervised ML is provided for reference

Designers, data scientists

Step 4

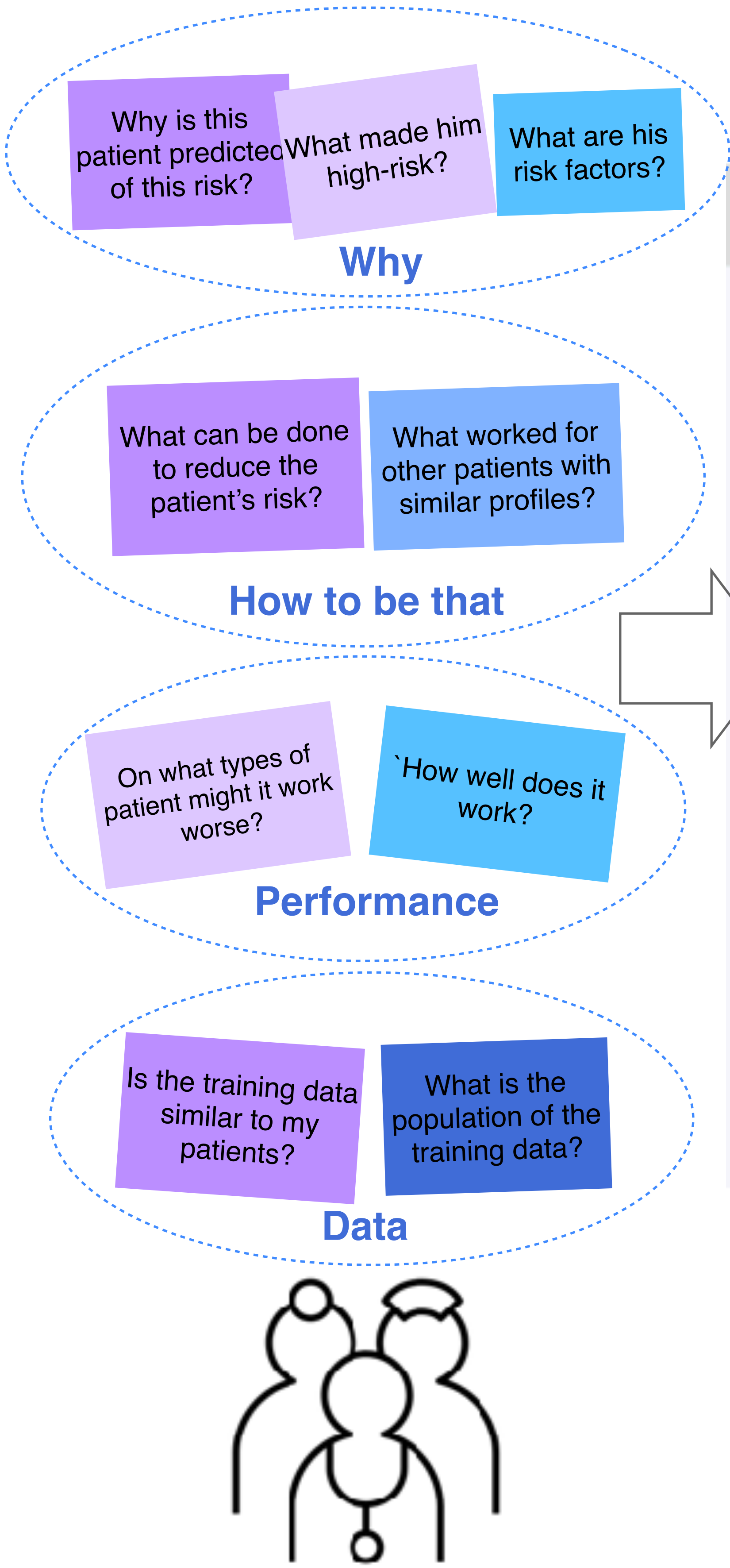
Iteratively design and evaluate

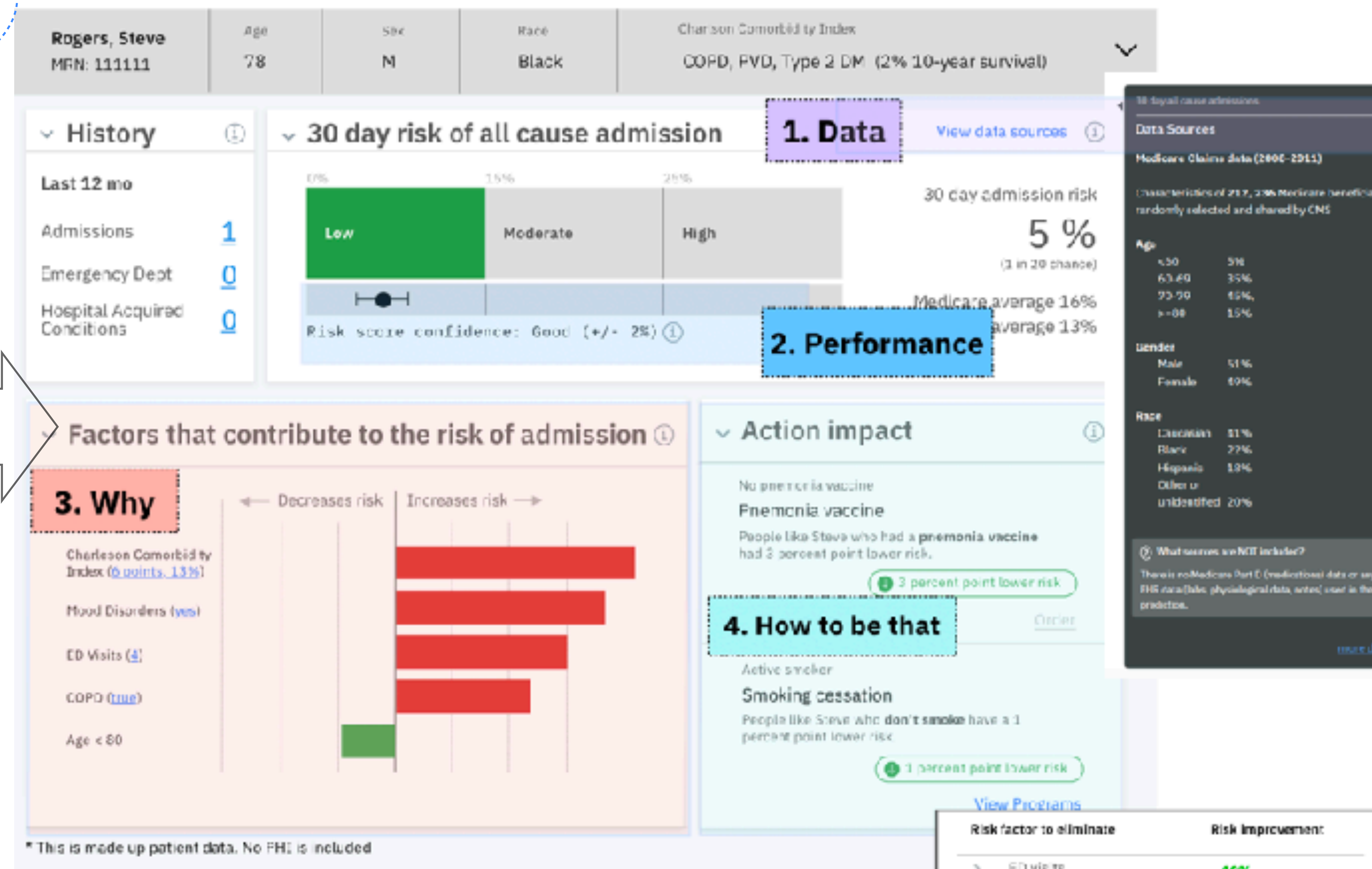
Create a design including the candidate elements identified in step 3

Iteratively valuate the design with the user requirements identified in step 2 and fill the gaps

Designers, data scientists, users

Liao et al. [Question-Driven Design Process for Explainable AI User Experiences](). (Working paper)





AI for Explainable Healthcare Adverse Event Risk Prediction

Liao et al. [Question-Driven Design Process for Explainable AI User Experiences](). (Working paper)

Assess the toolbox: **Uncovering pitfalls of existing XAI methods**

Pitfalls of XAI algorithms

- Disconnect with **user objectives and contexts** in deployment
  - Explainability defined in a vacuum v.s. *actionable understanding*
  - Current proxy evaluation tasks used by AI researchers have limited evaluative power (Buçinca, 2020; Zhang, 2020)

“Proxy evaluation tasks” disconnect with usage contexts and objectives

The AI must decide: Is 30% or more of the nutrients on this plate fat?

Fact: 30% or more of the nutrients on this plate is not fat.

Here are ingredients that the AI knows the fat content of and recognized as main nutrients:

avocado
bacon

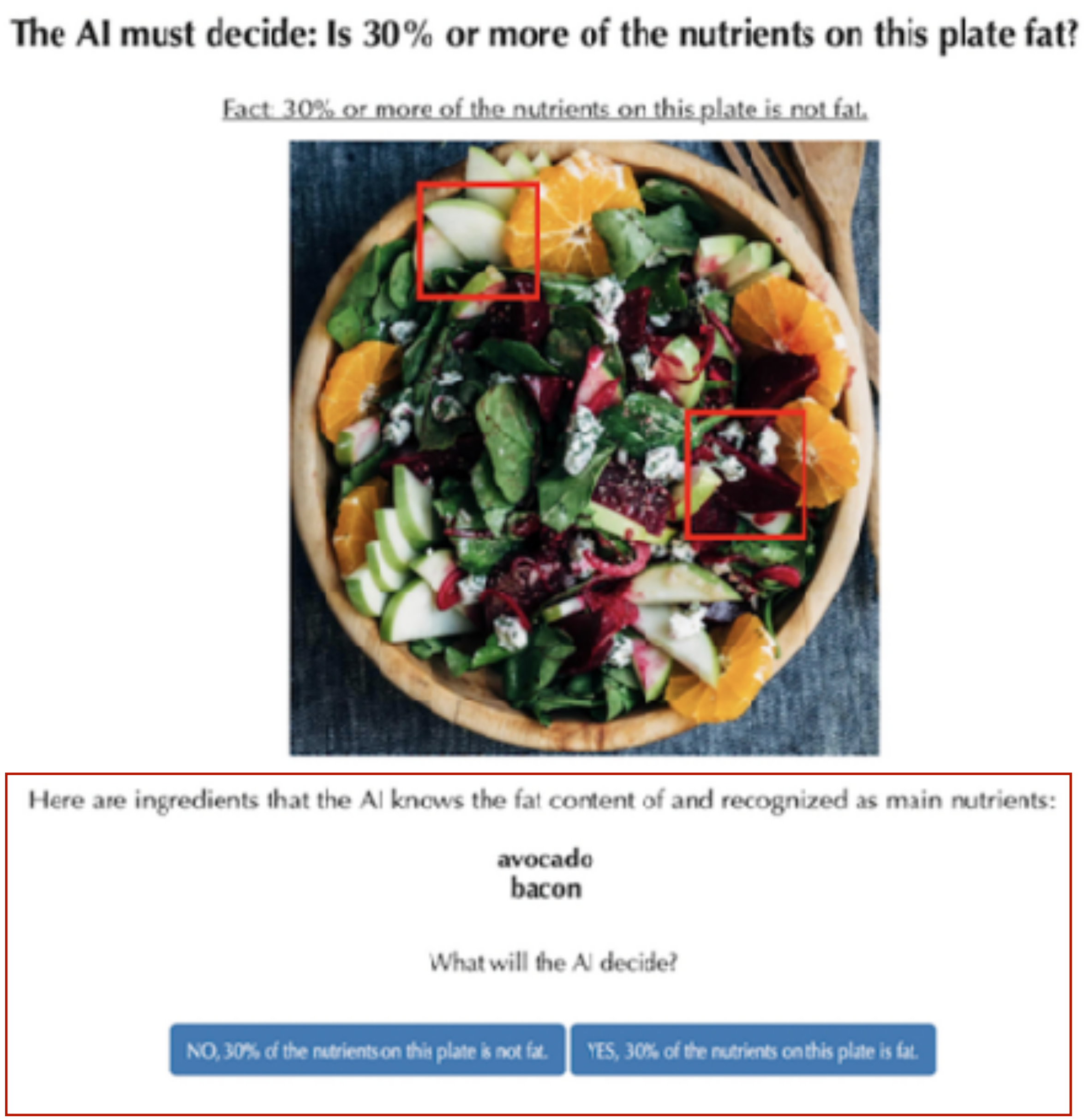
What will the AI decide?

NO, 30% of the nutrients on this plate is not fat. | YES, 30% of the nutrients on this plate is fat.

Proxy task: simulatability test

Buçinca et al. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. IUI 2020

“Proxy evaluation tasks” disconnect with usage contexts and objectives

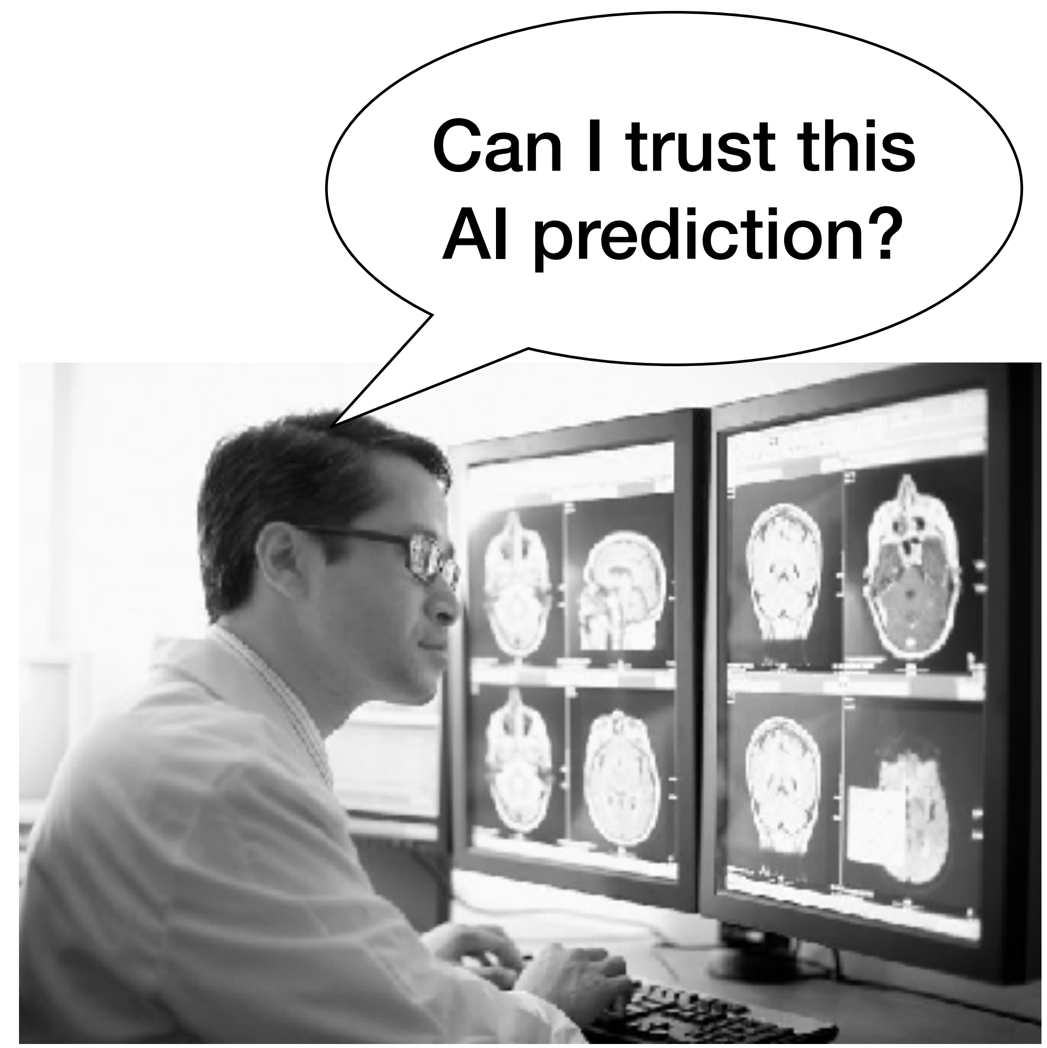


Proxy task: simulatability test

Buçinca et al. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. IUI 2020

“Proxy evaluation tasks” disconnect with usage contexts and objectives

Proxy task: simulatability test



User objective: appropriate reliance

“Proxy evaluation tasks” disconnect with usage contexts and objectives

Proxy task: simulatability test

User objective: seek recourse action

Pitfalls of XAI algorithms

- Disconnect with **user objectives and contexts** in deployment
  - “Explainability” defined in a vacuum v.s. actionable understanding
  - Current proxy evaluation tasks used by AI researchers have limited evaluative power (Buçinca, 2020; Zhang, 2020)
- Disconnect with **cognitive processes** receiving XAI
  - Unwarranted trust and confidence in models
  - Inequality of experiences

XAI can lead to unwarranted trust and confidence

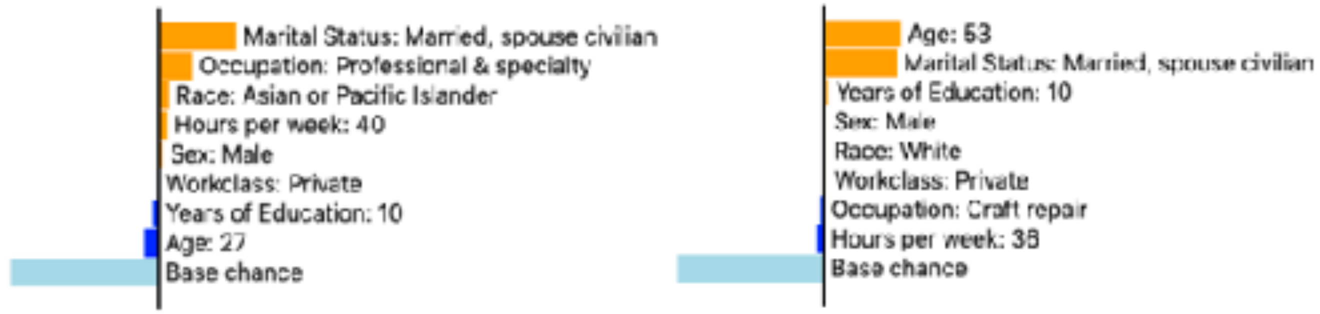


Figure 11: Screenshots of explanation for cases where the model had low confidence.

Showing explanation reduced decision accuracy (Zhang 2020)

Zhang et al. [Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making](). *FAT* 2020*

XAI can lead to unwarranted trust and confidence

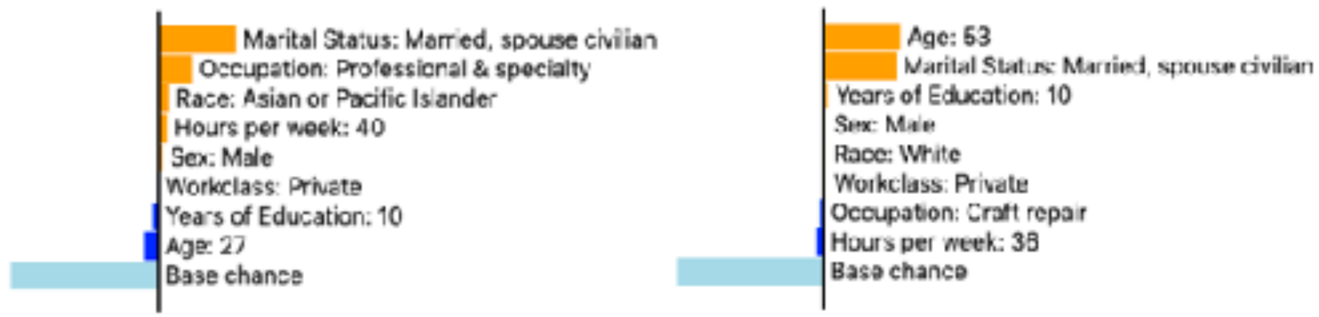
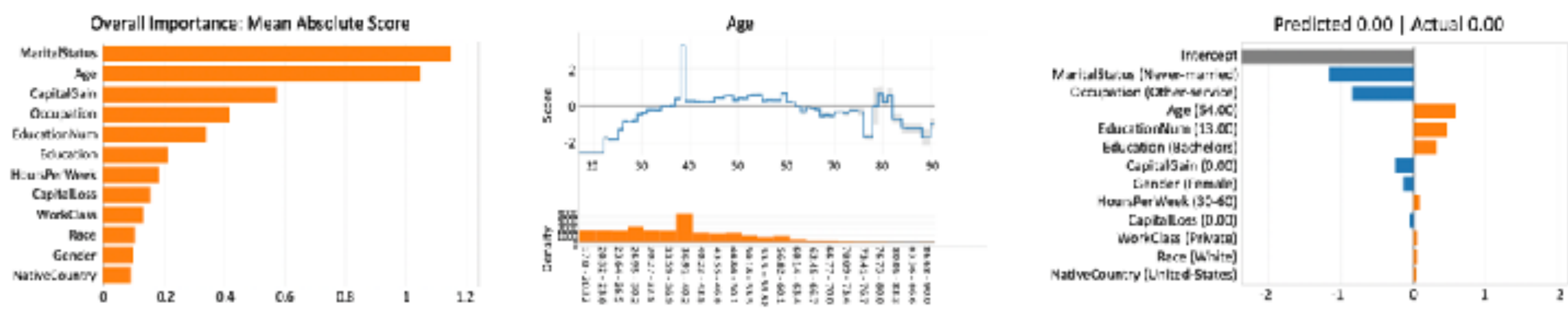


Figure 11: Screenshots of explanation for cases where the model had low confidence.

Showing explanation reduced decision accuracy (Zhang 2020)



“Interpretability tools” for data scientists can lead to over-confidence in readiness for deployment (Kauer 2020)

XAI can lead to unwarranted trust and confidence

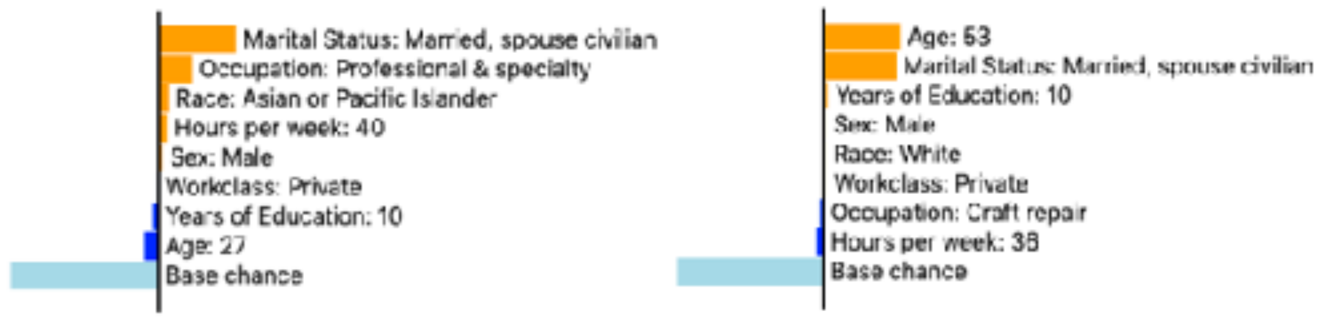
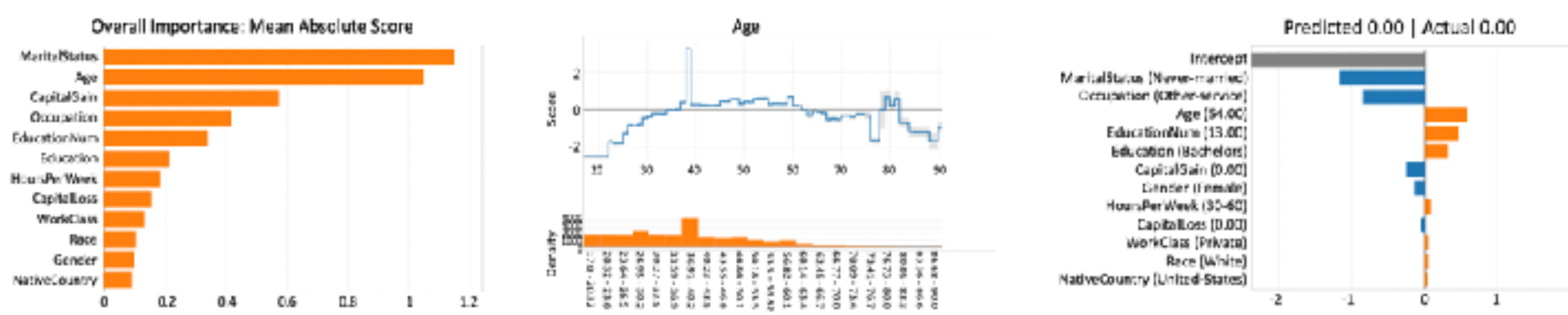


Figure 11: Screenshots of explanation for cases where the model had low confidence.

Showing explanation reduced decision accuracy (Zhang 2020)



"Interpretability tools" for data scientists can lead to over-confidence in readiness for deployment (Kauer 2020)

Placebic Explanation	Real Explanation
We need these details because they are necessary for the algorithm.	Based on this information, the algorithm calculates the need for calories and nutritional values and generates a corresponding nutrition plan so that you can reach your personal goal.

Even "placebic explanations" can increase trust (Einband, 2019)

A blind spot in XAI? Plurality of cognitive processes

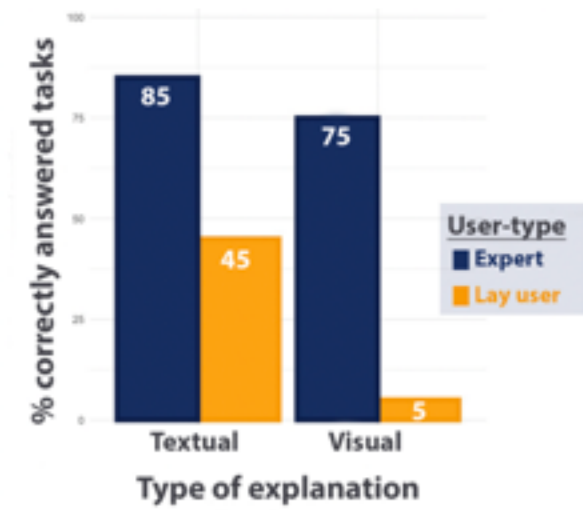
Ideal users assumed by XAI work

Read explanations carefully and able to understand it

Real users interacting with AI systems

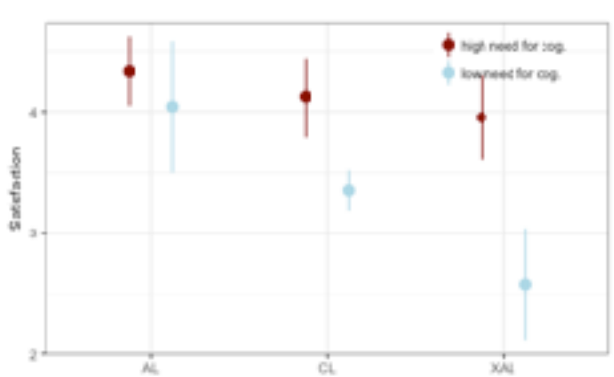
When lacking either **ability** or **motivation**, invoke cognitive heuristics (and biases)

XAI can lead to inequalities of experience



AI novices had less performance gain but more illusory satisfaction (Szymanski, 2021)

Benefited less from why-explanations in **cognitive resource constraint settings** (Robertson, 2021)



Decreased task satisfaction for people with trait of **low Need for Cognition** (Ghai, 2020)

Ghai et al. [Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers](). CSCW 2020

Expand the toolbox: **From algorithmic explanations to actionable understanding**

Paths forward: Cognitively compatible XAI

- Understand what **heuristics** are involved in XAI (Nourani, 2021; Ehsan 2021)
- Cultivate and leverage **warranted heuristics**
- **Interventions** for deeper system 2 processing of XAI (Buçinca, 2021)
- XAI with lower **cognitive workload** (Springer, 2019; Abdul, 2020)
- Developing the design space for **XAI communication**

Paths forward: Sociotechnical approaches to XAI

Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach

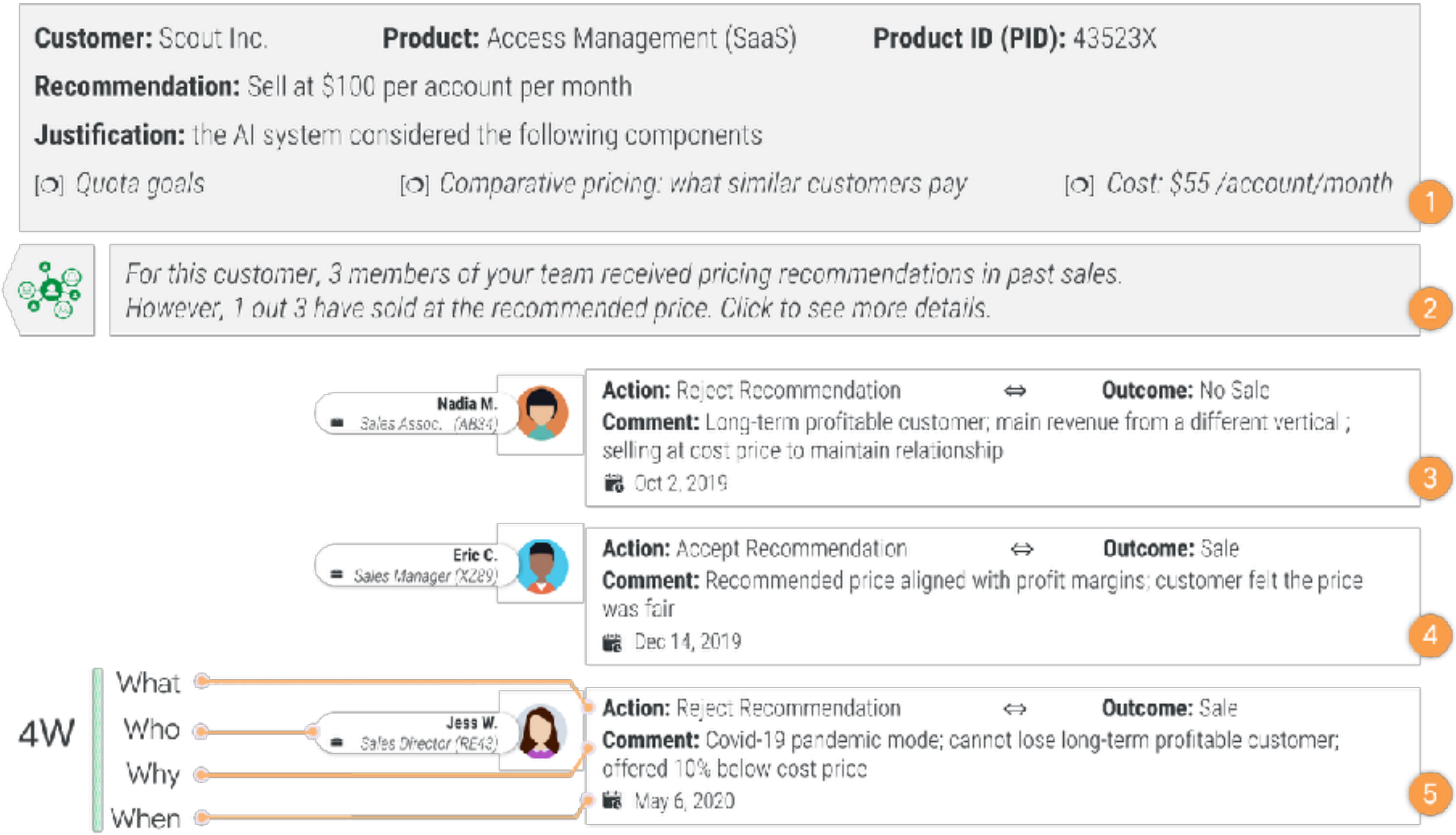
Upol Ehsan and Mark O. Riedl

Georgia Institute of Technology
Atlanta, GA 30308, USA
ehsanu@gatech.edu, riedl@cc.gatech.edu

Abstract. Explanations a form of post-hoc interpretability play an instrumental role in making systems accessible as AI continues to proliferate complex and sensitive sociotechnical systems. In this paper, we introduce Human-centered Explainable AI (HCXAI) as an approach that puts the human at the center of technology design. It develops a holistic understanding of *"who" the human is* by considering the interplay of values, interpersonal dynamics, and the socially situated nature of AI systems. In particular, we advocate for a *reflective sociotechnical* approach. We illustrate HCXAI through a case study of an explanation system for non-technical end-users that shows how technical advancements and the understanding of human factors co-evolve. Building on the case study, we lay out open research questions pertaining to further refining our understanding of "who" the human is and extending beyond 1-to-1 human-computer interactions. Finally, we propose that a *reflective* HCXAI paradigm—mediated through the perspective of Critical Technical Practice and supplemented with strategies from HCI, such as value-sensitive design and participatory design not only helps us understand our intellectual blind spots, but it can also open up new design and research spaces.

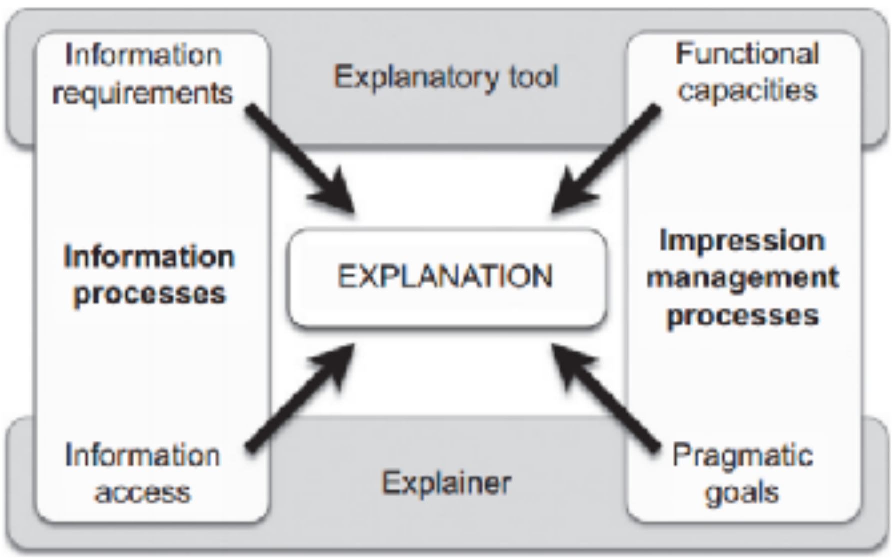
- AI systems are sociotechnical
- The “explainable to whom” and their sense-making process should be socially situated

Paths forward: Sociotechnical approaches to XAI

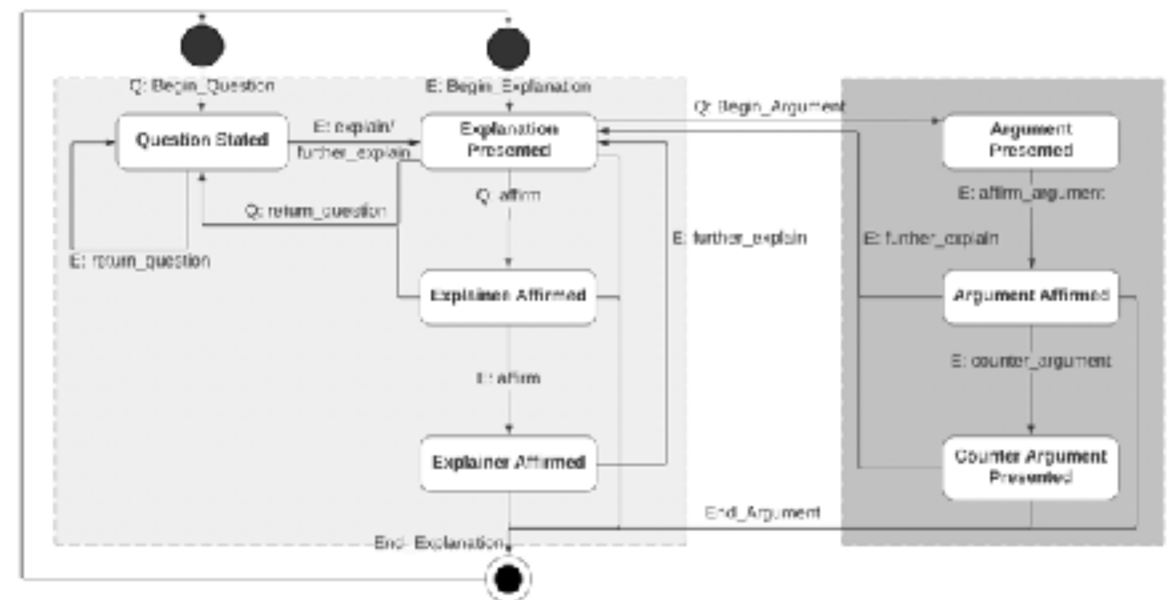


Ehsan et al. [Expanding Explainability: Towards Social Transparency in AI systems](). CHI 2021 🎖

Paths forward: Building on theories of human explanations

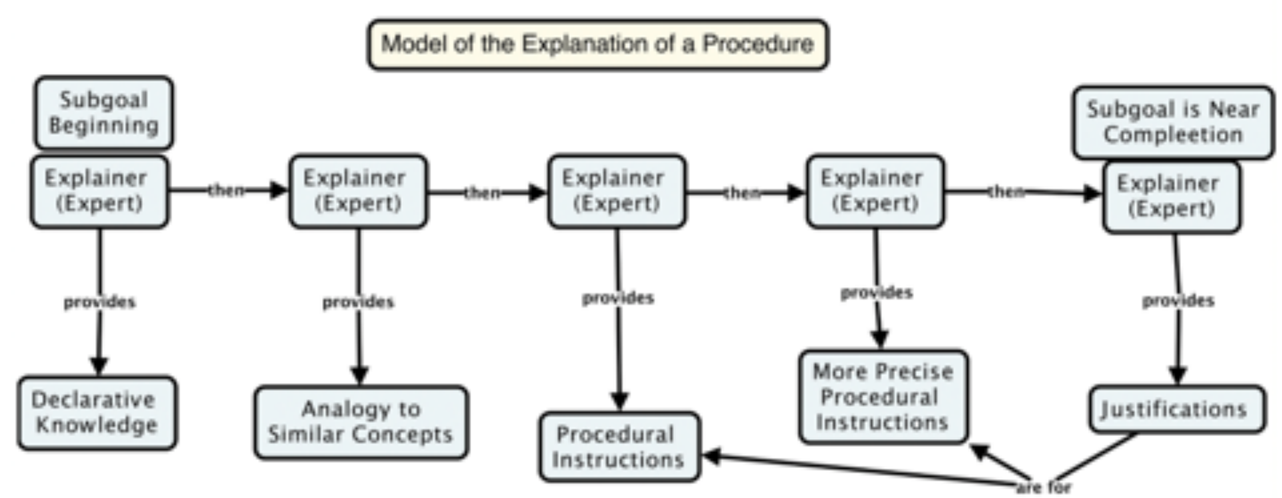


Malle's process model of explanation selection (Miller, 2019)



Explanation dialogue model (Madumal et al., 2019)

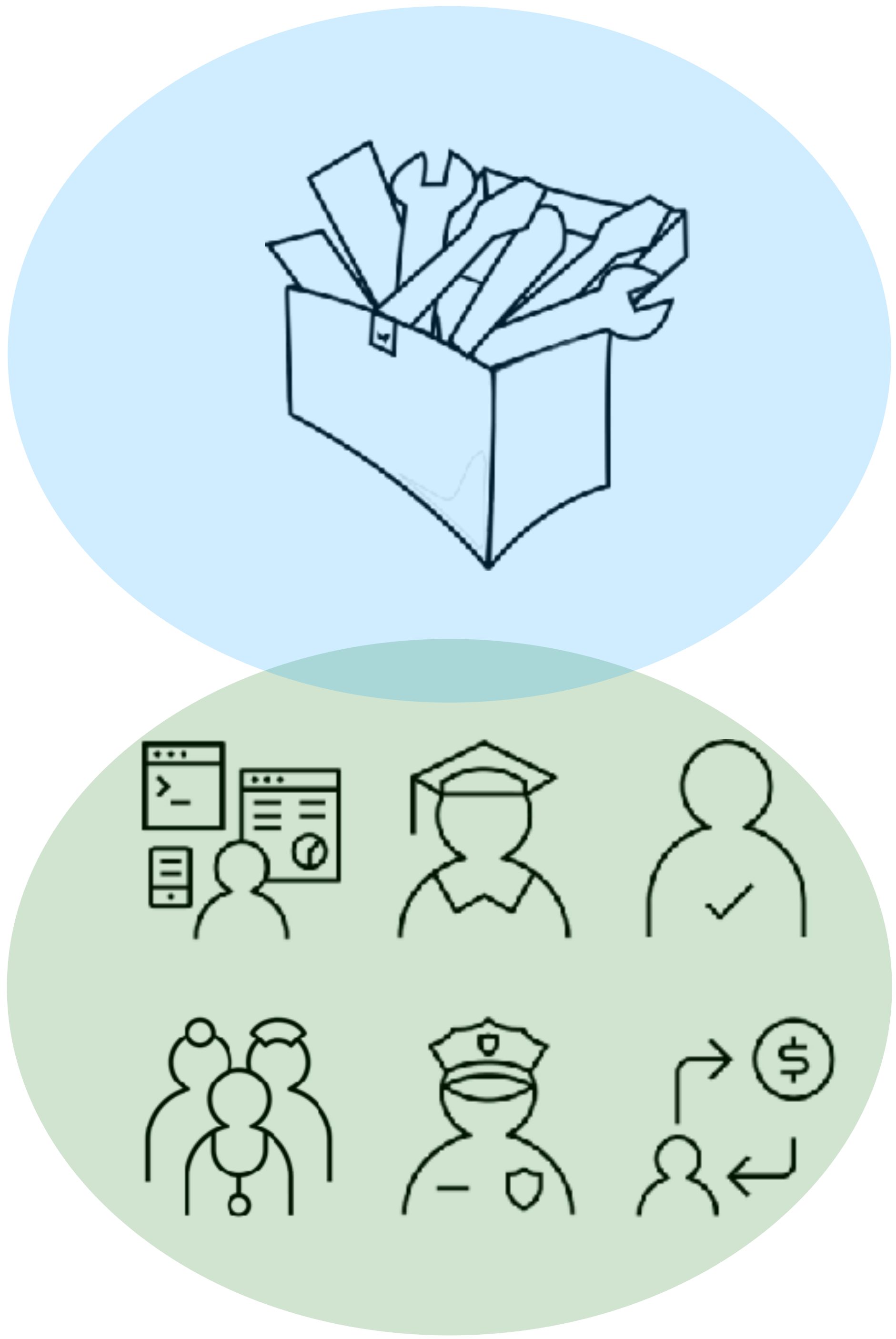
Models of normative and natural reasoning (Wang et al., 2019)



Johnson's model of the collaborative explanation process (Mueller et al., 2019)

Conclusions: HCI research as bridging work

- Human-centered re-framing of technical spaces
- Make responsible use of technical toolboxes
- Expand practitioners' toolbox with "design tools"
- Engage with deployment contexts and people's lived experiences, and bring back into technical development



Thank **YOU!**

…and thanks to

Rachel Bellamy, Amit Dhurandhar, Jonathan Dodge, Casey Dugan, Upol Ehsan, Bhavya Ghai, Werner Geyer, Daniel Gruen, Jaesik Han, Michael Hind, Stephanie Houde, David Piorkowski, Aleksandra Mojsilović, Sarah Miller, Tim Miller, Michael Muller, Shweta Narkar, Milena Pribić, John Richards, Mark Riedl, Daby Sow, Chenhao Tan, Richard Tomsett, Kush Varshney, Justin Weisz, Yunfeng Zhang

HCXAI

HUMAN-CENTERED EXPLAINABLE AI

HCXAI logo made by Upol Ehsan

veraliao@microsoft.com
www.qveraliao.com
@QVeraLiao