

# Questioning the AI: Toward Human Centered **Interpretable Machine Learning**

Research work 2018-2020

Q. Vera Liao  
IBM **Research**

# HCI research: **Bridging** work

Transfer emerging research or technologies into tangible *tools* and *guidelines* that help product teams navigate the design space



IEEE Access

## Machine Learning Inter Methods and Metrics

Diogo V. Carvalho<sup>1,2,\*</sup>, Eduardo M. Perei<sup>1</sup>

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)

## Explaining Explanations: An Overview of Interpretability of Machine Learning

## A Survey of Methods for Explaining Black Box Models

RICCARDO GUIDOTTI, FRANCO TURINI, KDDL, FOSCA GIANNOTTI, KDI, DINO PEDRESCHI, KDDI

## 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

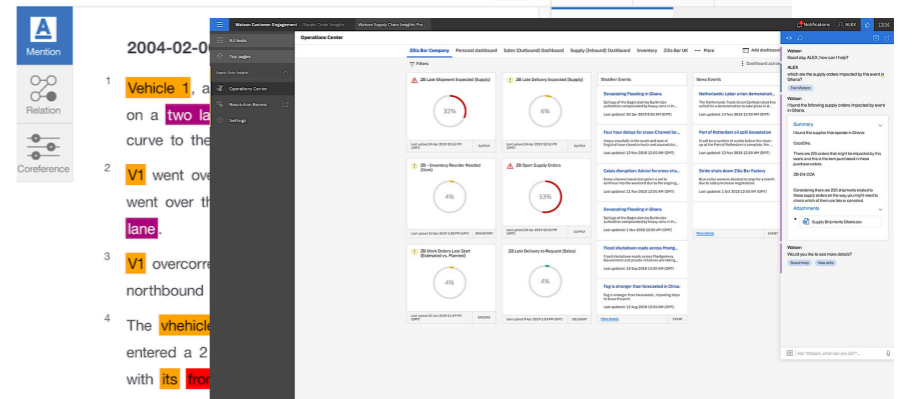
In recent years, many accurat systems that hide their interna ethical issue. The literature rep at the cost of sacrificing accur can be used are various, and ea and, as a consequence, it explic tion. The aim of this article is t respect to the notion of explar box type, and a desired explan

**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 explanation methods is introduced, and finally, the various classes of explanation methods are analyzed to verify if user concerns are justified. Overall, it is clear that (visual) explanations can be given about various aspects of the influence of the input on the output. However, it is noted that explanation methods or interfaces for lay users are missing and an encyclopedia which criteria

Inform usage



Identify gaps



# HCI research: **Bridging** work

Transfer emerging research or technologies into tangible *tools* and *guidelines* that help product teams navigate the design space



IEEE Access

Review  
Machine Learning Inter  
Methods and Metrics

Diogo V. Carvalho<sup>1,2,\*</sup>, Eduardo M. Perei  
<sup>1</sup> Deloitte Portugal, Manuel Bandeira Street, 4

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)

Explaining Explanations: An Overview of Interpretability of Machine Learning

A Survey of Methods for Explaining Black Box Models

RICCARDO GUIDOTTI,  
FRANCO TURINI, KDDLa  
FOSCA GIANNOTTI, KDI  
DINO PEDRESCHI, KDDI

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

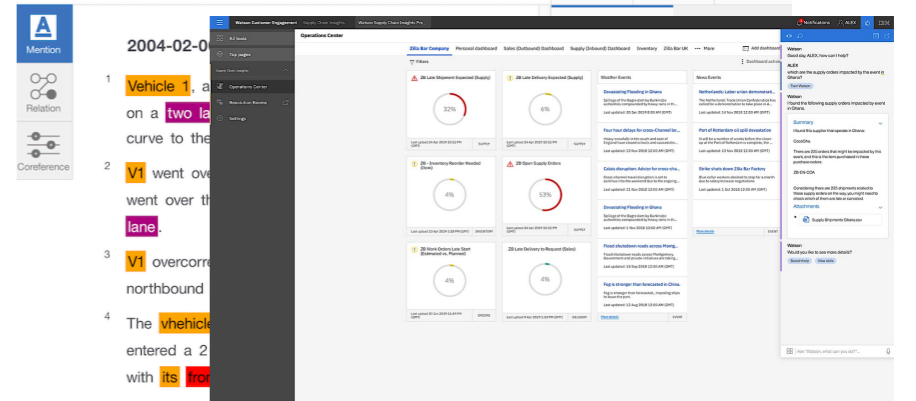
In recent years, many accurat  
systems that hide their interna  
ethical issue. The literature rep  
at the cost of sacrificing accur  
can be used are various, and ea  
and, as a consequence, it explic  
tion. The aim of this article is t  
respect to the notion of explar  
box type, and a desired explan

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 explanation methods is introduced, and finally, the various classes of explanation methods are analyzed to verify if user concerns are justified. Overall, it is clear that (visual) explanations can be given about various aspects of the influence of the input on the output. However, it is noted that explanation methods or interfaces for lay users are missing and an encyclopedia which criteria

Inform usage



Identify gaps



Which explanation technique to use?  
How to design XAI user experiences?

# Terminologies and definitions

Interpretable ML



**Explainable AI (XAI)**

## Narrow definition:

Techniques and methods that make a ML model's decisions understandable by people

## Broader (practitioners') definition:

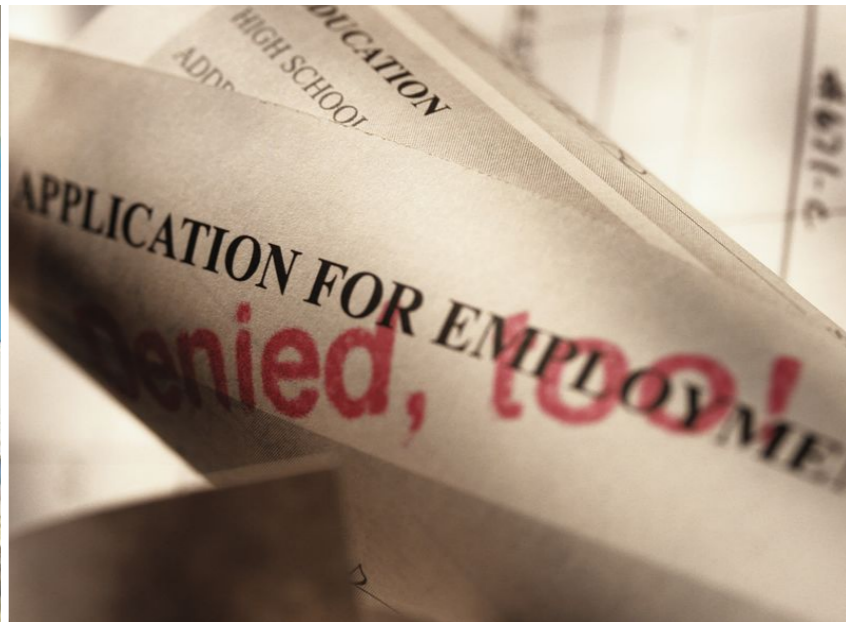
**Everything that makes AI more understandable (e.g., also including data, functions, performance)**

XAI is not just ML (also explainable robotics, planning, etc.), but I will focus on explaining supervised ML

# Towards human centered XAI: Agenda

- **Background and motivation for HCXAI**
- **Research into design**
  - **Question-driven explainable AI ( 🎓 CHI 2020)**
  - Designing social transparency in AI systems (CHI 2021)
- **Research through design and case studies**
  - **XAI for fair ML ( 🎓 IUI 2019)**
  - XAI for AI decision support (FAccT\* 2020)
  - XAI for active learning (CSCW 2020)
  - XAI for autoAI (IUI 2021)

AI is increasingly used in many high-stakes tasks



# The quest for explainable AI (XAI)

**Companies Grapple With AI's Opaque Decision-Making Process**

**We Need AI That Is Explainable, Auditable, and Transparent**

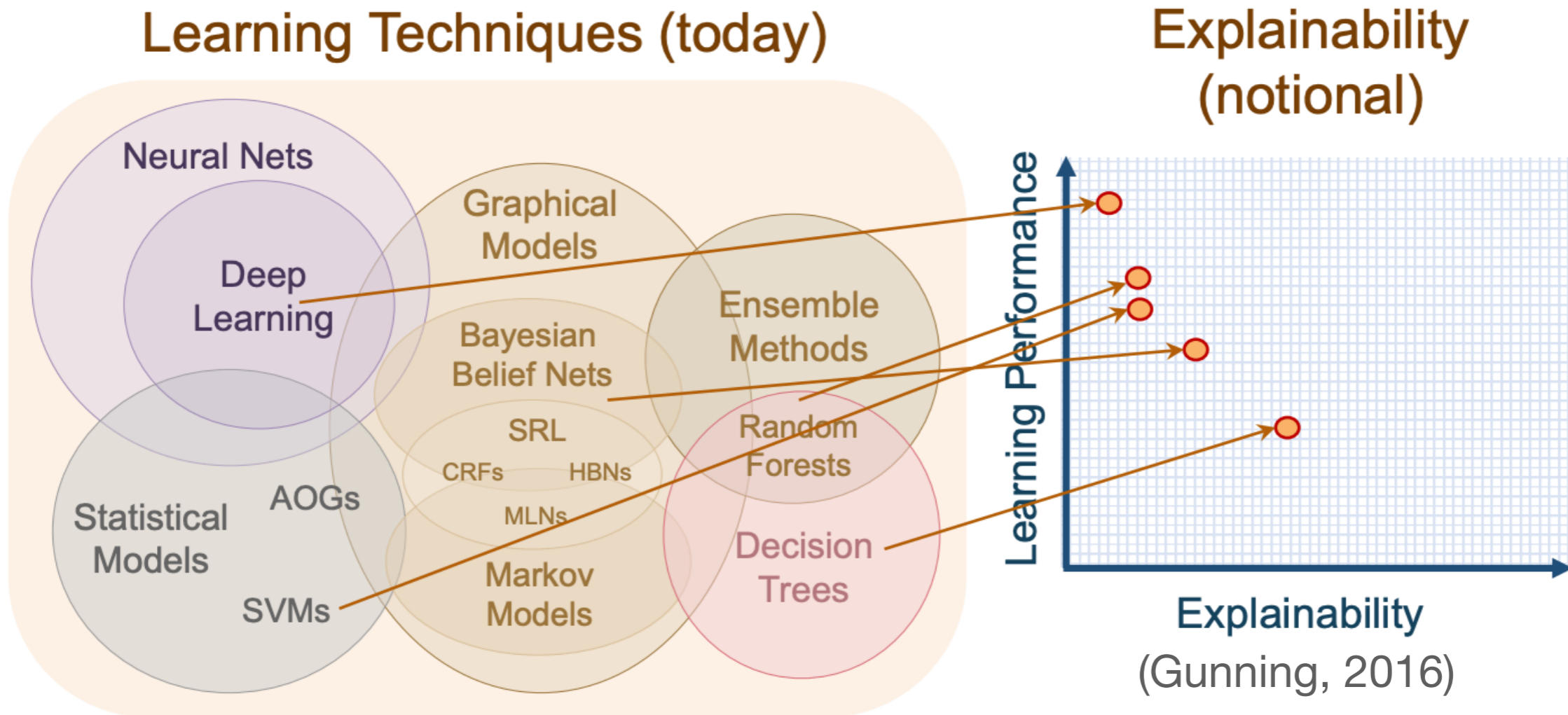
**Why “Explainability” Is A Big Deal In AI**

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

**How Explainable AI Is Helping Algorithms Avoid Bias**

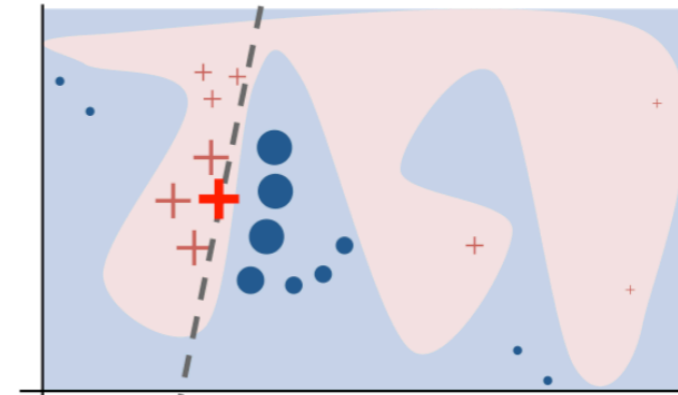
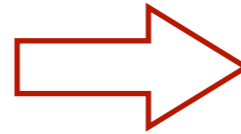
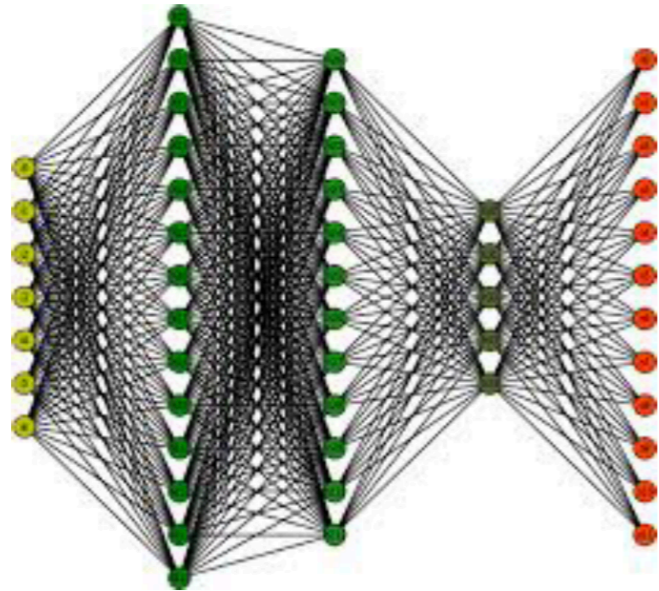


XAI is hard: it is technical





XAI is hard: it is technical



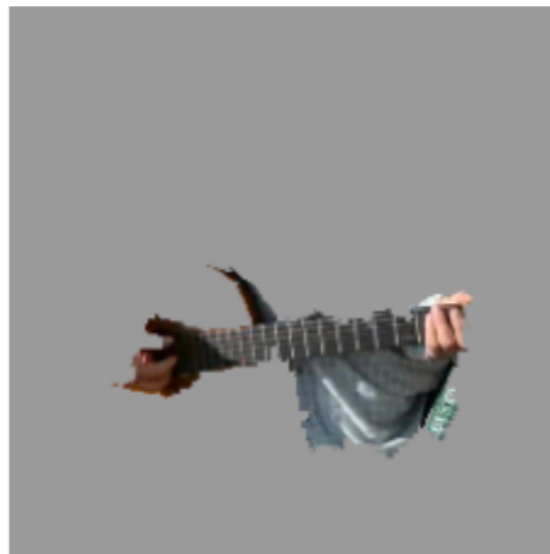
LIME (Ribeiro et al. 2016)

Neural network, not directly explainable

Use a *post-hoc* XAI technique



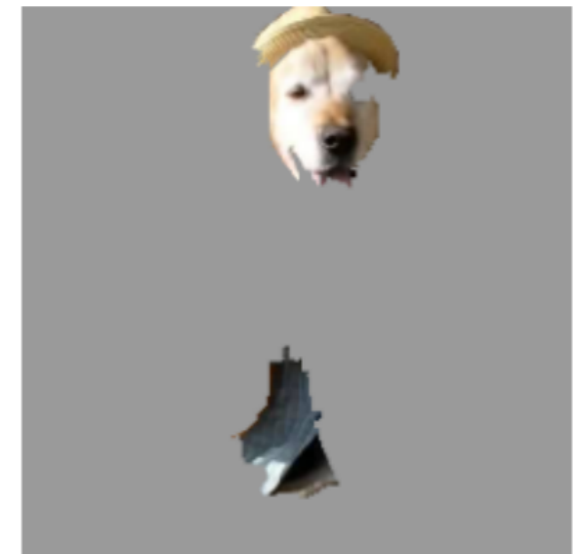
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

Review
Machine Learning Interpretability: A Survey on Methods and Metrics

Diogo V. Carvalho 1,2,\* and Eduardo M. Pereira 1 and
1 Deloitte Portugal, Manuel Bandeira Street, 43, 4150-47
2 Faculty of Engineering, University of Porto, Dr. Rober
3 INESC TEC, Dr. Roberto Frias Street, 4200-465 Porto, l
\* Correspondence: diocarvalho@deloitte.pt

Received: 21 June 2019; Accepted: 24 July 2019; Published

Abstract: Machine learning systems are becoming in
has been expanding, accelerating the shift toward
algorithmically informed decisions have greater po
most of these accurate decision support systems rem
logic and inner workings are hidden to the user

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

Abstract—There has recently been a surge of work in ex-
planatory artificial intelligence (XAI). This research area tackles
the important problem that complex machines and algorithms

As a first step towards creating explanation mechanism:
there is a new line of research in interpretability, loosely
defined as the science of comprehending what a model did (c
le models and learning method
les include visual cues to fin
networks in image recognition

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)

ABSTRACT At the dawn of the fourth industrial revolution, we are witnessing a fast and widespread
adoption of artificial intelligence (AI) in our daily life, which contributes to accelerating the shift towards a
more algorithmic society. However, even with such unprecedented advancements, a key impediment to the
use of AI-based systems is that they often lack transparency. Indeed, the black-box nature of these systems
allows powerful predictions, but it cannot be directly explained. This issue has triggered a new debate on
explainable AI (XAI). A research field holds substantial promise for improving trust and transparency of

A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems

SINA MO...
ERIC D...

The need for
intelligence
reasoning
to define, c
on differen

A growing collection of XAI techniques

challenges for identifying appropriate design and evaluation methodology and consolidating knowledge from
across efforts. To this end, this paper presents a survey and framework intended to share knowledge and
experiences of XAI design and evaluation methods across multiple disciplines. Aiming to support diverse
design goals and evaluation method in XAI research, after a thorough review of XAI related papers in the
fields of machine learning, visualization, and human-computer interaction, we present a categorization of

ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France
University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain
Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Bizkaia, Spain
Segula Technologies, Parc d'activité de Pissaloup, Trappes, France
Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, France
Computational Intelligence, University of Granada, 18071 Granada, Spain
nica, 28050 Madrid, Spain

A Survey of Methods for Explaining

RICCARDO GUIDOTTI, ANNA MONREALE, SALV
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
systems that hide their internal logic to the user. This lack of ex
ethical issue. The literature reports many approaches aimed at c
at the cost of sacrificing accuracy for interpretability. The appli
can be used are various, and each approach is typically develop
and, as a consequence, it explicitly or implicitly delineates its ov
tion. The aim of this article is to provide a classification of the m
respect to the notion of explanation and the type of black box
box type, and a desired explanation, this survey should help the

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, expla-
nations 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 con-
cerns 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
explanation methods is introduced, and finally, the various classes of explanation methods are
analyzed to verify if user concerns are justified. Overall, it is clear that (visual) explanations can
be given about various aspects of the influence of the input on the output. However, it is noted
that explanation methods or interfaces for lay users are missing and we speculate which criteria

(AI) has achieved a notable momentum that, if harnessed
tions over many application sectors across the field. For this
ire community stands in front of the barrier of explainability,
brought by sub-symbolism (e.g. ensembles or Deep Neural
type of AI (namely, expert systems and rule based models).
in the so-called eXplainable AI (XAI) field, which is widely
ictical deployment of AI models. The overview presented in
id contributions already done in the field of XAI, including a
r this purpose we summarize previous efforts made to define
ing a novel definition of explainable Machine Learning that
th a major focus on the audience for which the explainability
propose and discuss about a taxonomy of recent contributions

# Skater

Skater is a unified framework to enable Model Interpretation for all forms of model to help one build an Interpretable machine learning system often needed for real world use-cases (\*\* we are actively working towards to enabling faithful interpretability for all forms models). It is an open source python library designed to demystify the learned structures of a black box model both globally(inference on the basis of a complete data set) and locally(inference about an individual prediction).

The project was started as a research idea to find ways to enable better interpretability(preferably human interpretability) to predictive "black boxes" both for researchers and practioners. The project is still in beta phase.

## Install Skater

```
pip install -U skater
```

Option 1: without rule lists and without deepinterpreter  
pip install -U skater



What-If Tool demo - regression model for predicting age - UCI census income dataset

Visualize

Datapoints  Partial dependence plots

Nearest counterfactual  L1  L2

Threshold

Create similarity feature

Edit - Datapoint 22

Search features Descending

Feature	Value(s)	Attribution value(s)
marital-status	Divorced	0.1367
capital-gain	0	0.027

Microsoft Azure | Machine learning

Run 109

Select Explanation

Global explanation on classification model trained on IBM employee attrition dataset

Explainer: shap\_kernel Comment: Global explanation on classification model trained on IBM employee attrition dataset

Top K Features: 8

# A growing number of toolkits making XAI techniques accessible for practitioners

Alibi is an open source library on black-

Documentation

If you're interested detect.

## Goals

Provide high c

moDel A

build passing

## Overview

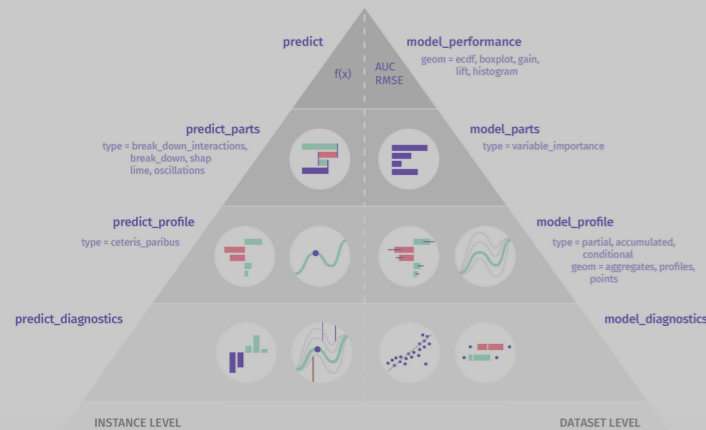
Unverified black box model is the path to the failure. Opaqueness leads to distrust. Distrust leads to ignorance. Ignorance leads to rejection.

The DALEX package xrays any model and helps to explore and explain its behaviour, helps to understand how complex models are working. The main function `explain()` creates a wrapper around a predictive model. Wrapped models may then be explored and compared with a collection of local and global explainers. Recent developments from the area of Interpretable Machine Learning/eXplainable Artificial Intelligence.

The philosophy behind DALEX explanations is described in the [Explanatory Model Analysis](#) e-book. The DALEX package is a part of [DrWhy.AI](#) universe.

If you work with `scikitlearn`, `keras`, `H2O`, `mljar` or `mlr`, you may be interested in the DALEXtra package. It is an extension pack for DALEX with easy to use connectors to models created in these libraries.

## DALEX: moDel Agnostic Language for Exploration and eXplanation



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 te you u:



## 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 a Question

Join our AI Explainability 360 Slack Channel to ask questions, make comments, and tell stories about how you use the toolkit.



### View Notebooks

Open a directory of Jupyter notebooks in GitHub that provide working examples of explainability in sample datasets. Then share your own notebooks!



### Contribute

You can add new algorithms and metrics in GitHub. Share Jupyter notebooks showcasing how you have enabled explanations in your machine learning application.



Learn how to put this toolkit to work for your application or industry problem. Try these tutorials.

### Credit Approval

See how to explain credit approval models using the

### Medical Expenditure

See how to create

### Dermoscopy

See how to explain dermoscopic image datasets

### Health and Nutrition Survey

See how to quickly

### Proactive Retention

See how to explain predictions of a model that

## 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

### Read More

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



### View Notebooks

Open a directory of Jupyter notebooks in GitHub that provide working examples of explainability in sample datasets. Then share your own notebooks!



- State-of-the-art XAI algorithms
- Comprehensive technical and educational resources
- Support a community of users and contributors

Website	<a href="http://aix360.mybluemix.net/">http://aix360.mybluemix.net/</a>
Repository	<a href="https://github.com/IBM/AIX360/">https://github.com/IBM/AIX360/</a>

application.



### Ask a Question

Join our AI Explainability 360 Slack Channel to ask questions, make comments, and tell stories about how you use the toolkit.



Learn how to put this toolkit to work for your application or industry problem. Try these tutorials.

### Credit Approval

See how to explain credit approval models using the

### Medical Expenditure

See how to create

### Dermoscopy

See how to explain dermoscopic image datasets

### Health and Nutrition Survey

See how to quickly

### Proactive Retention

See how to explain predictions of a model that

XAI is hard: it has to be user-centered

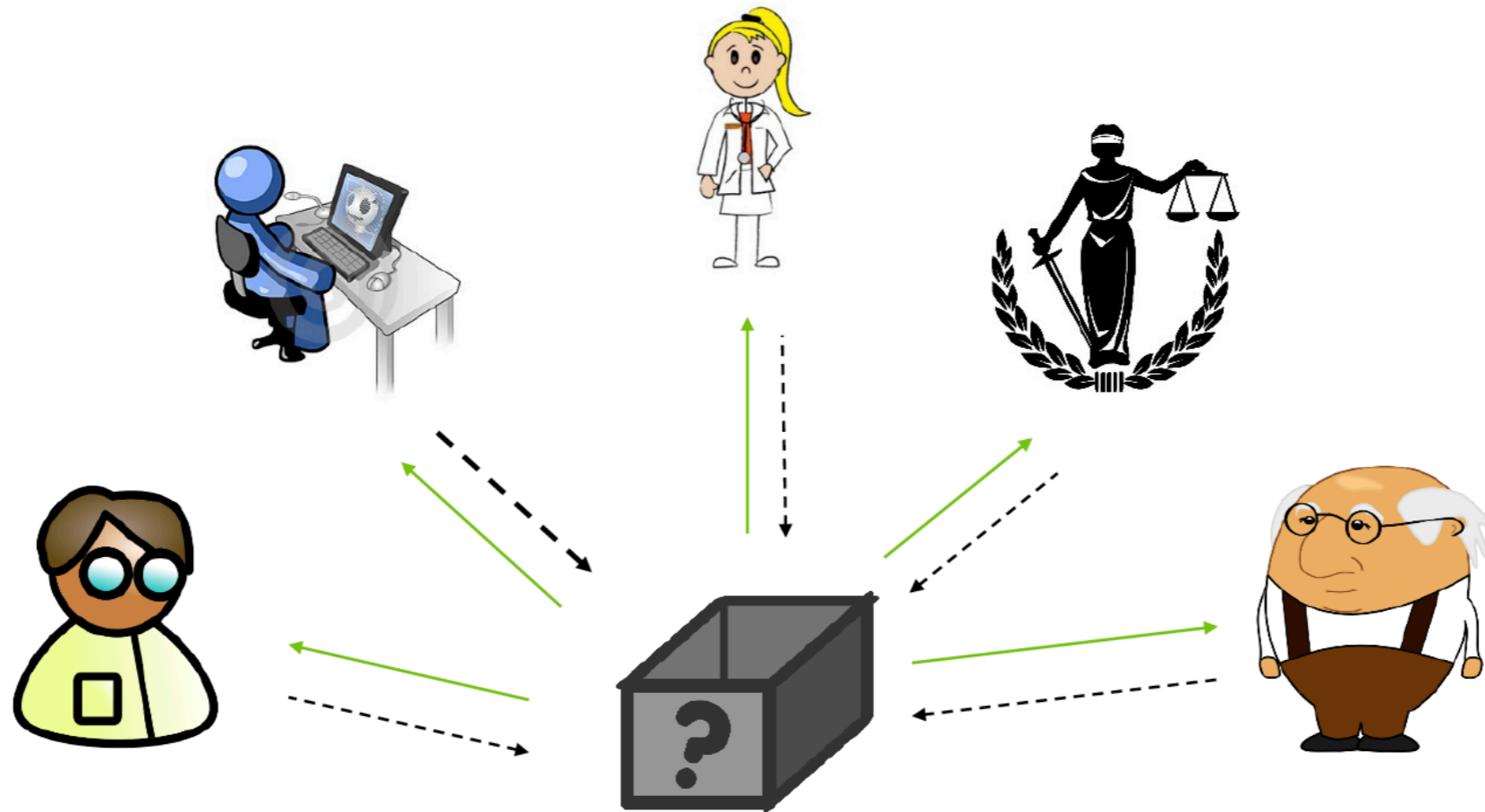
### The General Data Protection Regulation (GDPR)

- Limits to **decision-making** based solely on **automated processing** and profiling (Art.22)
- Right to be provided with **meaningful information** about the **logic** involved in the decision ( Art.13 (2) f. and 15 (1) h)

**“meaningful” ???**

(Nemitz, 2018)

XAI is hard: it has to be user-centered



(Hind et al., 2019)

Which explanation technique to use?  
How to design XAI user experiences?

# Motivation: Research into **XAI Design Practices**

## Why AI design practitioners?

- Bridging roles connecting user needs and XAI techniques
  - ➔ Develop design methods to support creating HCXAI
- Understanding real-world user needs for XAI
  - ➔ Inform future directions of XAI





# Machine Learning Interpretability: A Survey on Methods and Metrics

Diogo V. Carvalho<sup>1,2,\*</sup>, Eduardo M. Pereira<sup>1</sup> and  
<sup>1</sup> Deloitte Portugal, Manuel Bandeira Street, 43, 4150-47  
<sup>2</sup> Faculty of Engineering, University of Porto, Dr. Rober  
<sup>3</sup> INESC TEC, Dr. Roberto Frias Street, 4200-465 Porto, I  
\* Correspondence: diocarvalho@deloitte.pt

Received: 21 June 2019; Accepted: 24 July 2019; Published

**Abstract:** Machine learning systems are becoming in has been expanding, accelerating the shift toward algorithmically informed decisions have greater po most of these accurate decision support systems rem logic and inner workings are hidden to the user ratic

## 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

**Abstract**—There has recently been a surge of work in explanatory artificial intelligence (XAI). This research area tackles the important problem that complex machines and algorithms

As a first step towards creating explanation mechanisms: there is a new line of research in interpretability, loosely defined as the science of comprehending what a model did (c le models and learning method les include visual cues to fin networks in image recognition

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

AMINA ADADI<sup>1</sup> AND MOHAMMED BERRADA  
Computer and Interdisciplinary Physics Laboratory, Sidi Mohammed Ben Abdellah University, Fez 30050, Morocco  
Corresponding author: Amina Adadi (amina.adadi@gmail.com)

**ABSTRACT** At the dawn of the fourth industrial revolution, we are witnessing a fast and widespread adoption of artificial intelligence (AI) in our daily life, which contributes to accelerating the shift towards a more algorithmic society. However, even with such unprecedented advancements, a key impediment to the use of AI-based systems is that they often lack transparency. Indeed, the black-box nature of these systems allows powerful predictions, but it cannot be directly explained. This issue has triggered a new debate on explainable AI (XAI). A research field holds substantial promise for improving trust and transparency of

## A Multidisciplinary Survey and Framework for Design and Evaluation

SINA MOHAMED  
ERIC D. RAY  
The need for intelligence and reasoning be to define, des on different c challenges fo across effort experiences design goals fields of machine learning, visualization, and human-computer

# A technical space people are not quite in there yet... how to talk about it?

## A Survey of Methods for Explaining

RICCARDO GUIDOTTI, ANNA MONREALE, SALV 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 systems that hide their internal logic to the user. This lack of ethical issue. The literature reports many approaches aimed at c at the cost of sacrificing accuracy for interpretability. The appli can be used are various, and each approach is typically develop and, as a consequence, it explicitly or implicitly delineates its ov tion. The aim of this article is to provide a classification of the m respect to the notion of explanation and the type of black box box type, and a desired explanation, this survey should help the

## 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 explanation methods is introduced, and finally, the various classes of explanation methods are analyzed to verify if user concerns are justified. Overall, it is clear that (visual) explanations can be given about various aspects of the influence of the input on the output. However, it is noted that explanation methods or interfaces for lay users are missing and we speculate which criteria

computational Intelligence, University of Granada, 18071 Granada, Spain  
nica, 28050 Madrid, Spain

(AI) has achieved a notable momentum that, if harnessed tions over many application sectors across the field. For this ire community stands in front of the barrier of explainability, brought by sub-symbolism (e.g. ensembles or Deep Neural type of AI (namely, expert systems and rule based models). in the so-called *eXplainable* AI (XAI) field, which is widely ictical deployment of AI models. The overview presented in id contributions already done in the field of XAI, including a r this purpose we summarize previous efforts made to define ring a novel definition of explainable Machine Learning that th a major focus on the audience for which the explainability propose and discuss about a taxonomy of recent contributions

# Study probe: algorithm informed **XAI Questions**

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model <b>(Global)</b>	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	<b>How</b>
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	<b>How, Why, Why not, What if</b>
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	<b>How, Why, Why not, What if</b>
Explain a prediction <b>(Local)</b>	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	<b>Why</b>
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	<b>Why, How to still be this</b>
<b>Inspect counterfactual</b>	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	<b>What if, How to be that, How to still be this</b>
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	<b>Why, Why not, How to be that</b>
<b>Example based</b>	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	<b>Why, How to still be this</b>
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	<b>Why, Why not, How to be that</b>

- User needs for XAI are represented as **prototypical questions**
- A **question** can be answered by one or multiple **XAI methods**
- An **XAI method** can be implemented by one or multiple **XAI algorithms**

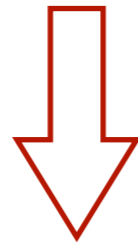


*An explanation is an answer to a question (Wellman, 2011; Miller 2018)*

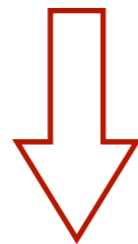
*The effectiveness of an explanation depends on the question asked (Bromberger, 1992)*



**Question:** Why is this husky classified as wolf?



**XAI method:** local feature (pixels) contribution



**XAI algorithms:**

- LIME (Ribeiro et al. 2016)
- SHAP (Lundberg and Lee 2017)
- ...

# Study probe: algorithm informed **XAI Questions**

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model <b>(Global)</b>	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	<b>How</b>
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	<b>How, Why, Why not, What if</b>
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	<b>How, Why, Why not, What if</b>
Explain a prediction <b>(Local)</b>	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	<b>Why</b>
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	<b>Why, How to still be this</b>
<b>Inspect counterfactual</b>	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	<b>What if, How to be that, How to still be this</b>
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	<b>Why, Why not, How to be that</b>
<b>Example based</b>	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	<b>Why, How to still be this</b>
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	<b>Why, Why not, How to be that</b>

+

**Input (data), output, performance**

(Lim et al., 2009)

# Methodology

- Interviewed **20 participants**
  - **16 AI products** in IBM
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?

# Methodology

- Interviewed **20 participants**
  - **16 AI products** in IBM
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 question bank

## Data

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

## Output

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

## Performance

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

## How (global)

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

## Why

## Why not

## What If

## How to be that

## How to still be this

## Others

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

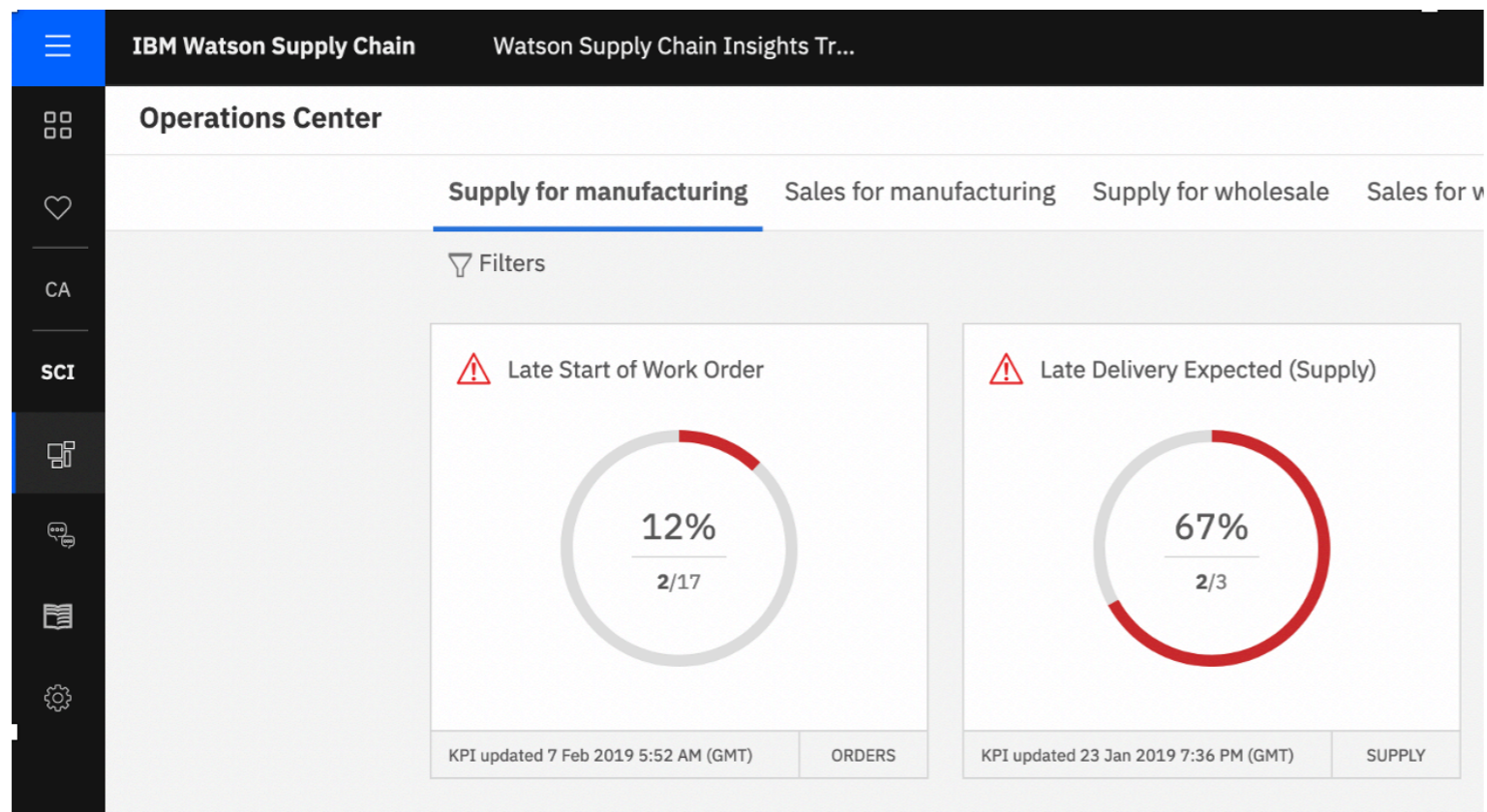
# XAI design challenge 1: Variability of XAI needs

## **Diverse end goals for explainability**

- To gain further insights for the decision
- To appropriately evaluate AI's capability
- To adapt usage or interaction
- To improve AI performance
- Ethical responsibilities of AI products



To gain further insights for the decision



**Why**  
**How to be that**

“ Users need to know why the system is saying this will be late because the reason is going to determine what their next action is...If it's because of a weather event, so no matter what you do you're not going to improve this number, versus something small, if you just make a quick call, you can get that number down (1-5)

To appropriately evaluate AI's capability



**Performance  
How**

“ There is a calibration of trust, whether people will use it over time. But also saying hey, we know this fails in this way (1-6)

# XAI design challenge 1: Variability of XAI needs

## **Diverse end goals for explainability**

- To gain further insights for the decision
- To appropriately evaluate AI's capability
- To adapt usage or interaction
- To improve AI performance
- Ethical responsibilities of AI products

Also varying XAI needs: **User group, usage point, algorithm and data type, decision context**

# XAI design challenge 2: Gaps between algorithmic output and human explanations

Human explanations are

- **Selective**
- **Contrastive**
- **Interactive**
- **Tailored for recipients**



Design attempt to mimic how people, especially domain experts, explain

## XAI design challenge 3: “in the dark” design process

- **Challenge navigating the technical capabilities**
- **Communication barriers** between designers, data scientists and other stakeholders
- **Cost of time and resource** impeding buy-in

“ 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. (1-8)

# XAI Question Bank

## Data

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

## Output

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

## Performance

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

## How (global)

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

## Why

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

## Why not

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

## What If

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

## How to be that

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

## How to still be this

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

## Others

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

- **A checklist representing the space of user needs for XAI**
- Understand real-world user questions to derive design guidelines
- New questions (with \*) inform gaps in XAI technical work

# XAI Question Bank

## Data

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

## Output

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

## Performance

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

## How (global)

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

## Why

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

## Why not

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

## What If

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

## How to be that

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

## How to still be this

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

## Others

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

- A checklist representing the space of user needs for XAI
- **Understand real-world user questions and how to address them**
- New questions (with \*) inform gaps in XAI technical work

# Understand XAI questions and desired solutions

**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 limitation of the AI and make it actionable

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

**Local decision:** Provide resources for “why not”

**Counterfactual:** Consider opportunities as utility features for analytics or system exploration



# XAI Question Bank

## Data

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

## Output

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

## Performance

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

## How (global)

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

## Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?
- **Why/how is this instance NOT predicted...?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

## Why not

## What If

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

## How to be that

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

## How to still be this

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

## Others

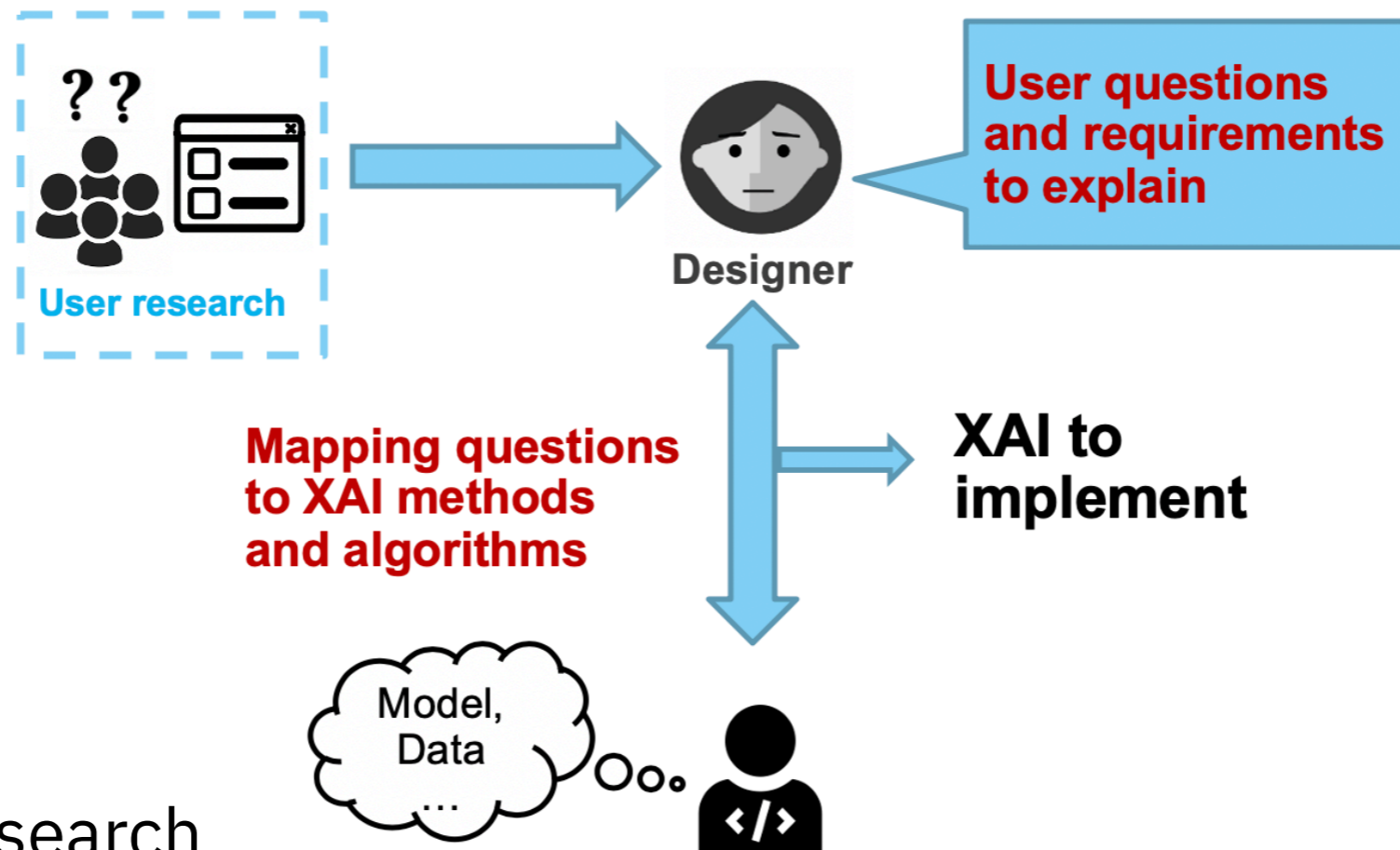
- \* How/what/why will the system change/adapt/improve/drift over time? (change)
- \* How to improve the system? (change)
- \* Why using or not using this feature/rule/data? (follow-up)
- \* What does [ML terminology] mean? (terminological)
- \* What are the results of other people using the system?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

- A checklist representing the space of user needs for XAI
- Understand real-world user questions and how to address them
- **New questions (with \*) inform gaps in XAI technical work**

# Opportunities for future technical XAI work

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

# Supporting the process: **question-driven XAI design**

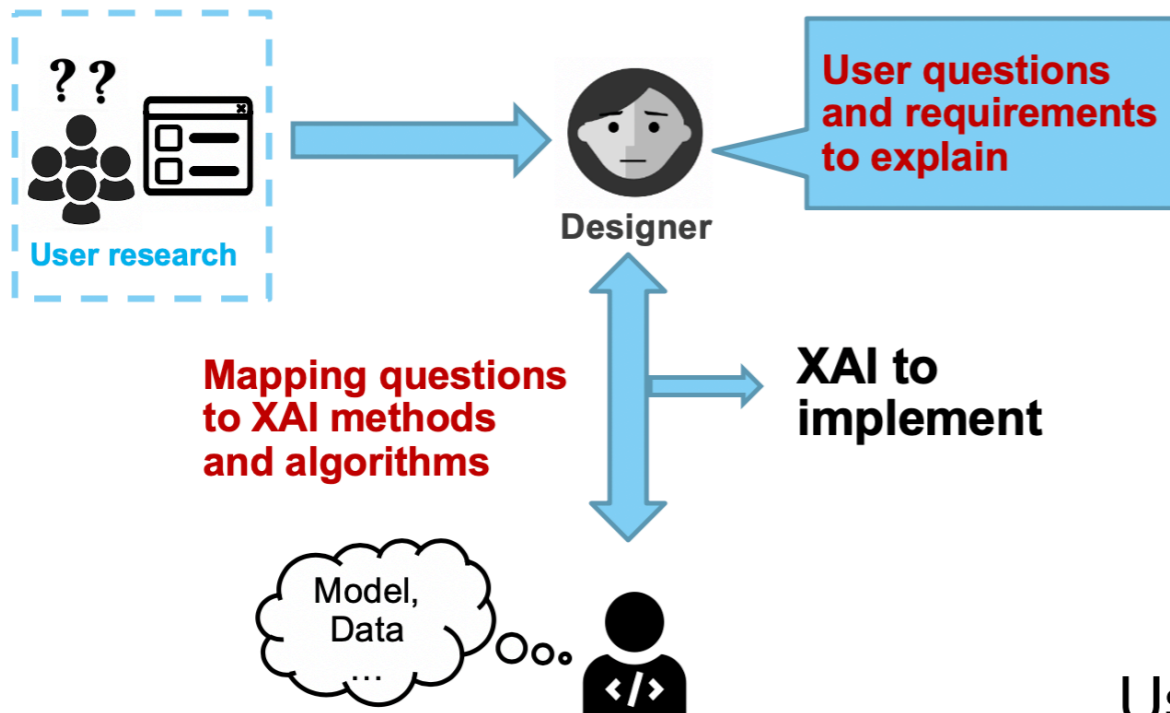


Through user research

- **Questions** elicitation
- Identify user **requirements** to address the *questions*

Working with data scientists and the team

- Map the *questions* to **XAI technique(s)**
- **Iteratively** evaluate by the user requirements and fill the gaps



### XAI Question Bank

<b>Data</b>	<ul style="list-style-type: none"> <li>• What kind of data does the system learn from?</li> <li>• What is the source of the data?</li> <li>• How were the labels/ground-truth produced?</li> <li>• * What is the sample size?</li> <li>• * What data is the system NOT using?</li> <li>• * What are the limitations/biases of the data?</li> <li>• * How much data [like this] is the system trained on?</li> </ul>	<b>Why</b>	<ul style="list-style-type: none"> <li>• Why/how is this instance given this prediction?</li> <li>• What feature(s) of this instance leads to the system's prediction?</li> <li>• Why are [instance A and B] given the same prediction?</li> </ul>
<b>Output</b>	<ul style="list-style-type: none"> <li>• What kind of output does the system give?</li> <li>• What does the system output mean?</li> <li>• How can I best utilize the output of the system?</li> <li>• * What is the scope of the system's capability? Can it do...?</li> <li>• * How is the output used for other system component(s)?</li> </ul>	<b>Why not</b>	<ul style="list-style-type: none"> <li>• Why/how is this instance NOT predicted...?</li> <li>• Why is this instance predicted P instead of Q?</li> <li>• Why are [instance A and B] given different predictions?</li> </ul>
<b>Performance</b>	<ul style="list-style-type: none"> <li>• How accurate/precise/reliable are the predictions?</li> <li>• How often does the system make mistakes?</li> <li>• In what situations is the system likely to be correct/incorrect?</li> <li>• * What are the limitations of the system?</li> <li>• * What kind of mistakes is the system likely to make?</li> <li>• * Is the system's performance good enough for...?</li> </ul>	<b>What if</b>	<ul style="list-style-type: none"> <li>• What would the system predict if this instance changes to...?</li> <li>• What would the system predict if this feature of the instance changes to...?</li> <li>• What would the system predict for [a different instance]?</li> </ul>
<b>How (global)</b>	<ul style="list-style-type: none"> <li>• How does the system make predictions?</li> <li>• What features does the system consider? <ul style="list-style-type: none"> <li>• * Is [feature X] used or not used for the predictions?</li> </ul> </li> <li>• What is the system's overall logic? <ul style="list-style-type: none"> <li>• How does it weigh different features?</li> <li>• What rules does it use?</li> <li>• How does [feature X] impact its predictions?</li> <li>• * What are the top rules/features it uses?</li> </ul> </li> <li>• * What kind of algorithm is used?</li> <li>• * How are the parameters set?</li> </ul>	<b>How to be that</b>	<ul style="list-style-type: none"> <li>• How should this instance change to get a different prediction?</li> <li>• How should this feature change for this instance to get a different prediction?</li> </ul>
		<b>How to still be this</b>	<ul style="list-style-type: none"> <li>• What is the scope of change permitted to still get the same prediction?</li> <li>• What is the [highest/lowest/... ] feature(s) one can have to still get the same prediction?</li> <li>• What is the necessary feature(s) present or absent to guarantee this prediction?</li> <li>• What kind of instance gets this prediction?</li> </ul>
		<b>Others</b>	<ul style="list-style-type: none"> <li>• * How/what/why will the system change/adapt/improve/drift over time? (change)</li> <li>• * How to improve the system? (change)</li> <li>• * Why using or not using this feature/rule/data? (follow-up)</li> <li>• * What does [ML terminology] mean? (terminological)</li> <li>• * What are the results of other people using the system? (social)</li> </ul>

Use XAI question bank to guide question elicitation

Category of Methods	Explanation Method	Definition	Algorithm Examples	Question Type
Explain the model <b>(Global)</b>	Global feature importance	Describe the weights of features used by the model (including visualization that shows the weights of features)	[41, 60, 69, 90]	<b>How</b>
	Decision tree approximation	Approximate the model to an interpretable decision-tree	[11, 47, 52]	<b>How, Why, Why not, What if</b>
	Rule extraction	Approximate the model to a set of rules, e.g., if-then rules	[26, 93, 102]	<b>How, Why, Why not, What if</b>
Explain a prediction <b>(Local)</b>	Local feature importance and saliency method	Show how features of the instance contribute to the model's prediction (including causes in parts of an image or text)	[61, 74, 83, 85, 101]	<b>Why</b>
	Local rules or trees	Describe the rules or a decision-tree path that the instance fits to guarantee the prediction	[39, 75, 99]	<b>Why, How to still be this</b>
<b>Inspect counterfactual</b>	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature (often in a visualization format)	[8, 33, 36, 51]	<b>What if, How to be that, How to still be this</b>
	Contrastive or counterfactual features	Describe the feature(s) that will change the prediction if perturbed, absent or present	[27, 91, 100]	<b>Why, Why not, How to be that</b>
<b>Example based</b>	Prototypical or representative examples	Provide example(s) similar to the instance and with the same record as the prediction	[13, 48, 50]	<b>Why, How to still be this</b>
	Counterfactual example	Provide example(s) with small differences from the instance but with a different record from the prediction	[37, 55, 66]	<b>Why, Why not, How to be that</b>

A guide to mapping questions to XAI techniques for supervised ML

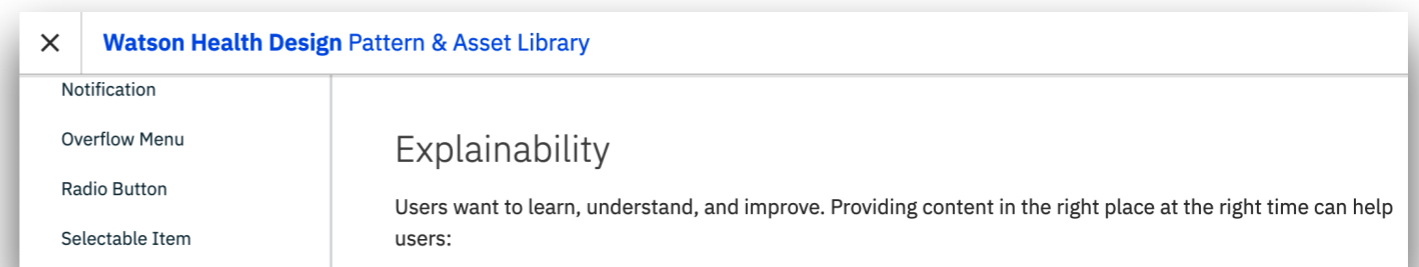
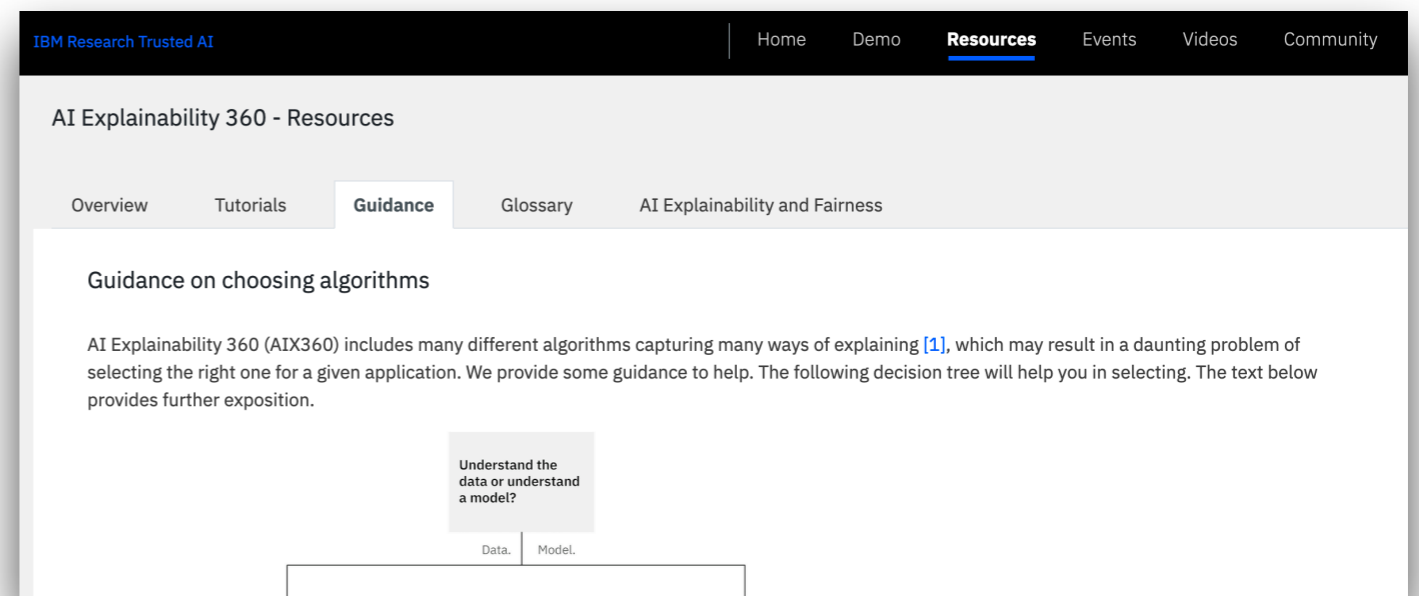
# Summary, and putting research into practices...

## Real-world user needs for **nine categories of AI explainability**

- Guidelines to address them
- Opportunities for future algorithmic work

## Challenges faced by design practitioners

- **XAI Question Bank**
- **Question-driven XAI design process**



# Research through design and case studies

Use case	Fair ML	AI-assisted decision	Active learning	Auto AI/ML
User	Regulator, impacted group	Decision-maker	Annotator	Model builder
Key RQ	How do different styles of explanation impact <b>fairness judgment</b> ?	Can local explanation improve <b>decision outcomes</b> ?	Can local explanation improve <b>model training</b> and <b>annotator experience</b> ?	Can interactive explanation support <b>model selection</b> ?

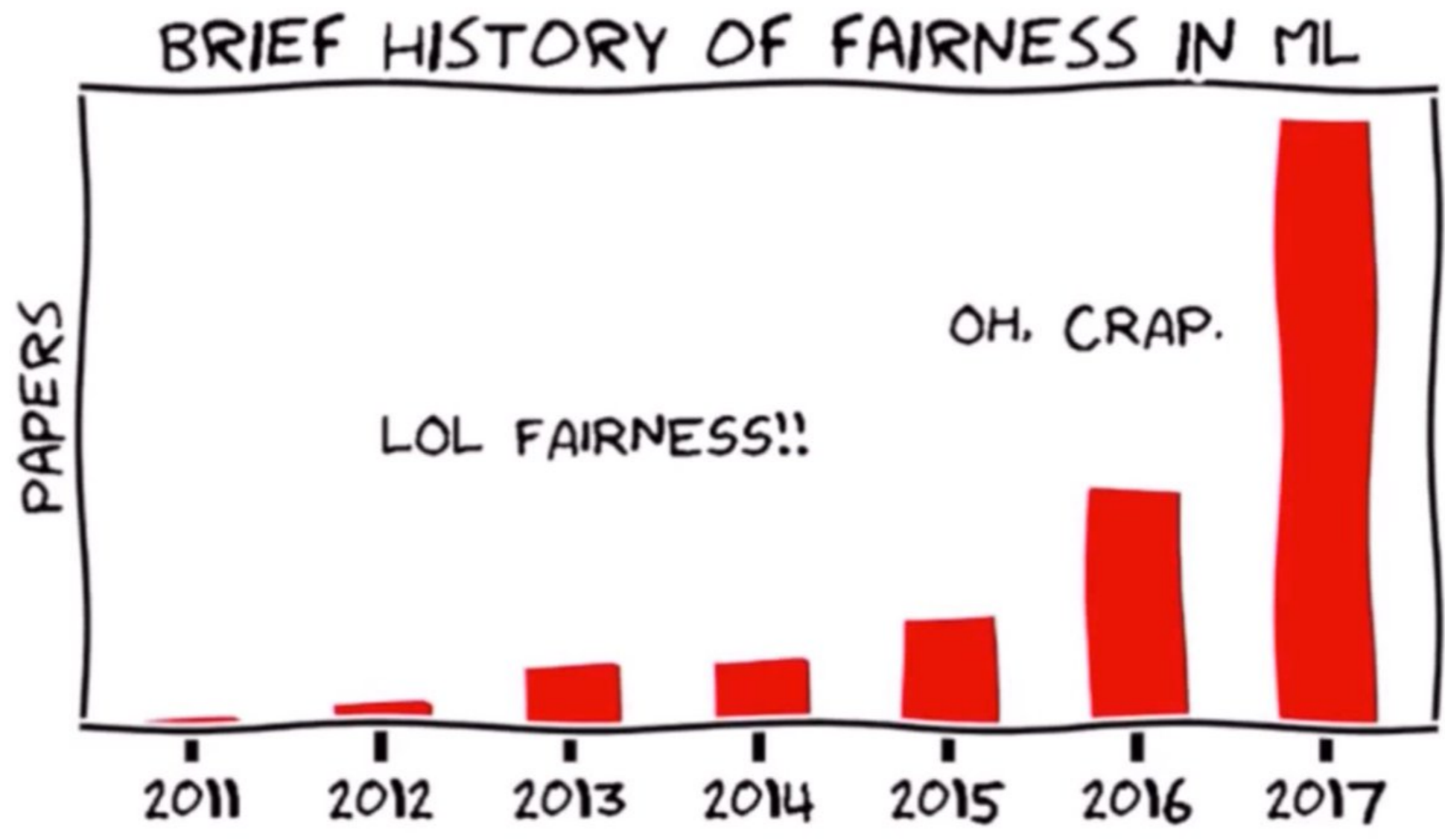
# Fair ML: What is unwanted bias?



Discrimination becomes objectionable when it places certain **unprivileged** groups at a systematic disadvantage

Illegal in certain contexts

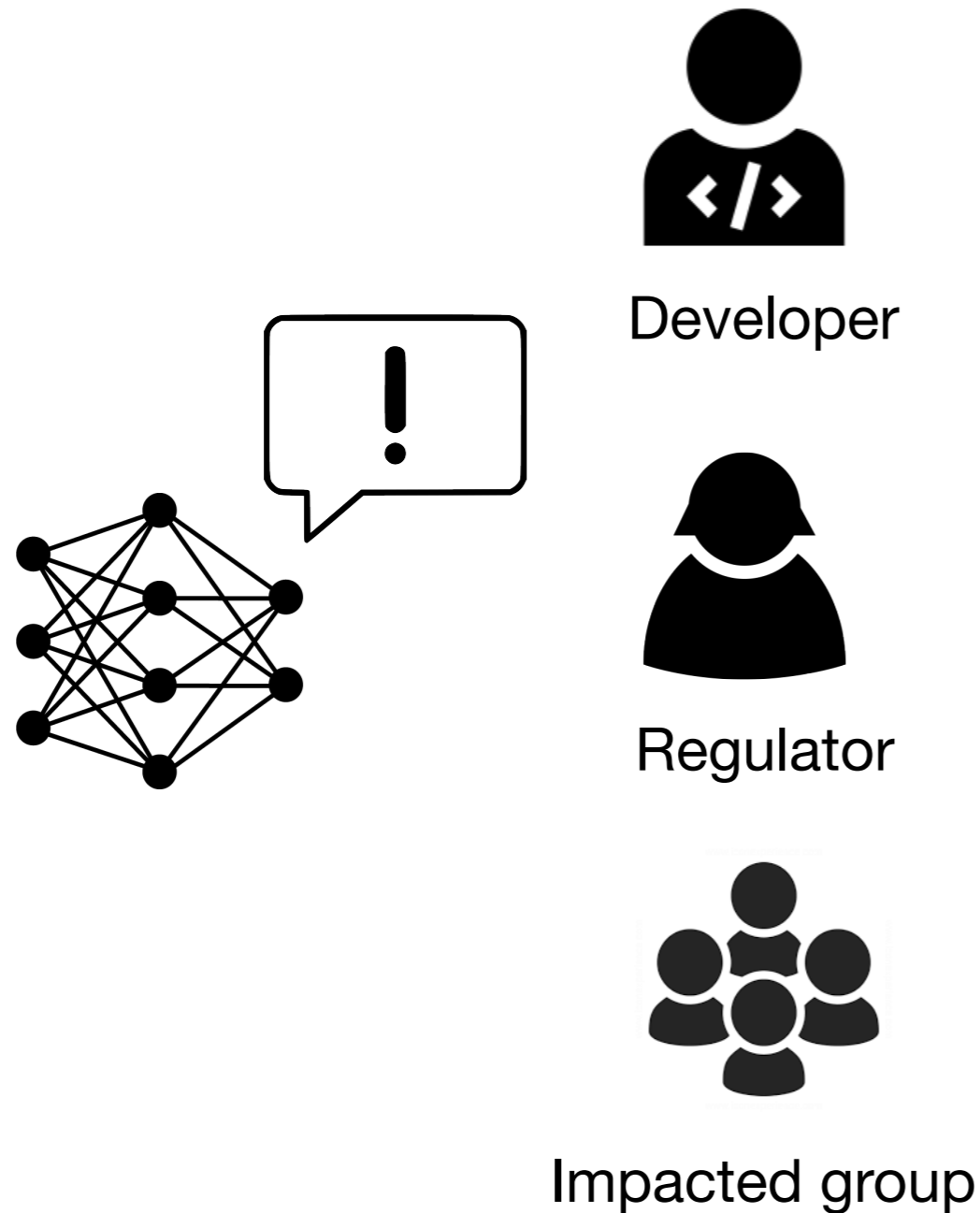
(Barocas and Selbst, 2017)



(Hardt, 2017)



# XAI as interfaces for scrutinizing model biases



End goal of XAI:  
Fairness **calibration**

If the model is **biased**

➔ **Help identify the bias**

If the model is **unbiased**

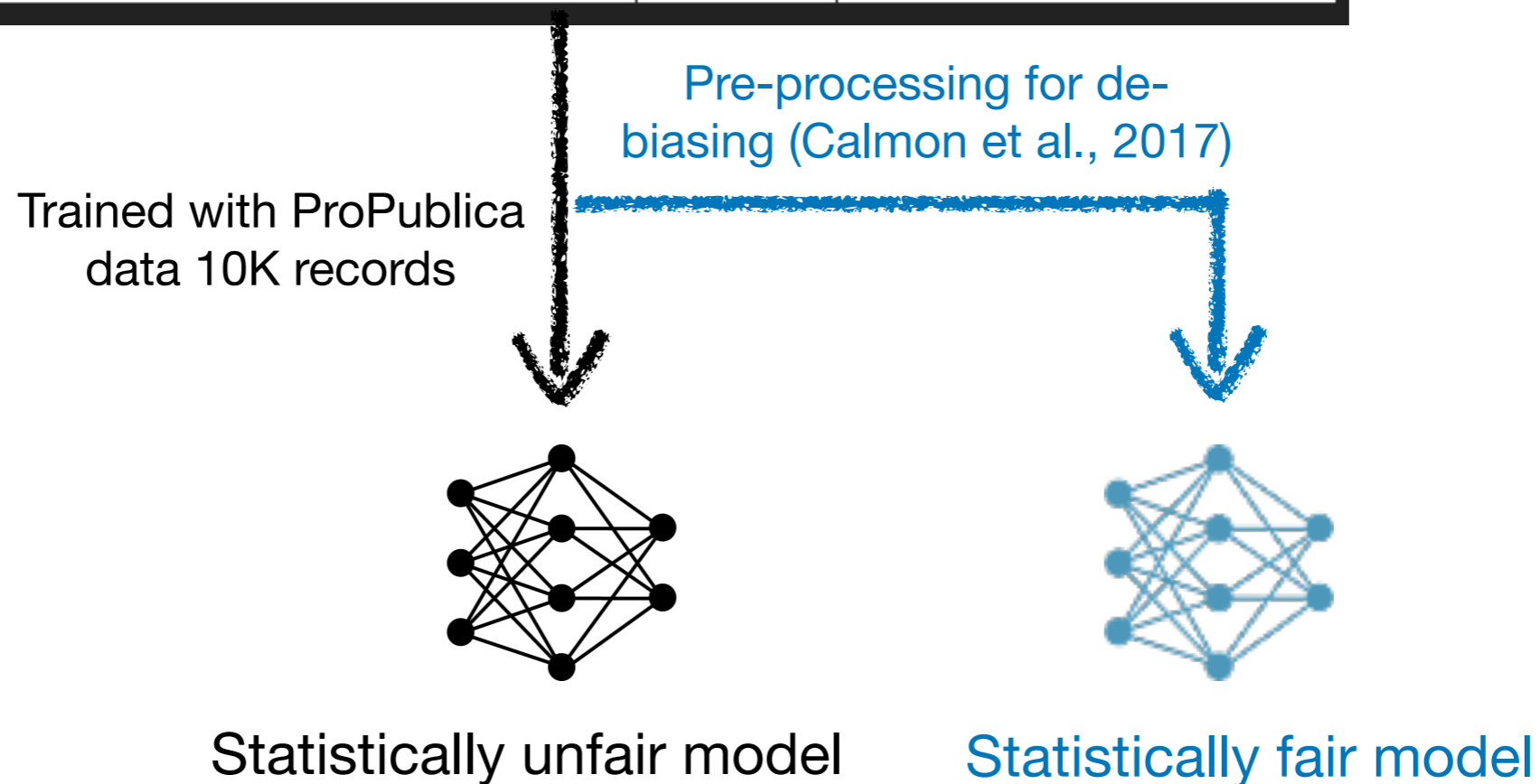
➔ **Help foster trust and confidence**

# Prototype and use case: explaining COMPAS

COMPAS is a software used to assess the recidivism risk of a defendant who posts a bail. Widely criticized as biased.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

(Larson et al.  
ProPublica, 2016)



### Sensitivity

- Iliana's race is **African American**. If it had been **Caucasian**, she would have been predicted as NOT likely to reoffend
- Iliana's age is **18-29**. If it had been **older than 39**, she would have been predicted as NOT likely to reoffend

### Input-Influence

The more +s/-s means a person with that attribute is more/less likely to re-offend.

\* Appears next to Iliana's attributes

Race

- Caucasian (0)
- \* **African-American (+)**

Age

- \* **18-29 (++++)**
- 30-39(+)
- ...

Charge degree:

- ...

Number of prior convictions

Has juvenile priors:

### Defendant: Iliana

- Race: African-American
- Age: 18-29
- Charge degree: Misdemeanor
- Prior convictions: 0
- Has juvenile priors: Yes

Prediction:

**Likely to reoffend**

### Case

The training set contained 10 individuals identical to Iliana

6 of them reoffend (60%)

### Demographic

The prediction is based on the likelihood of previous cases with different attributes re-offended or not. \* Appears next to Iliana's

Race

- 40% in Caucasian race group re-offended
- \* **55% in African-American race group re-offended**

Age

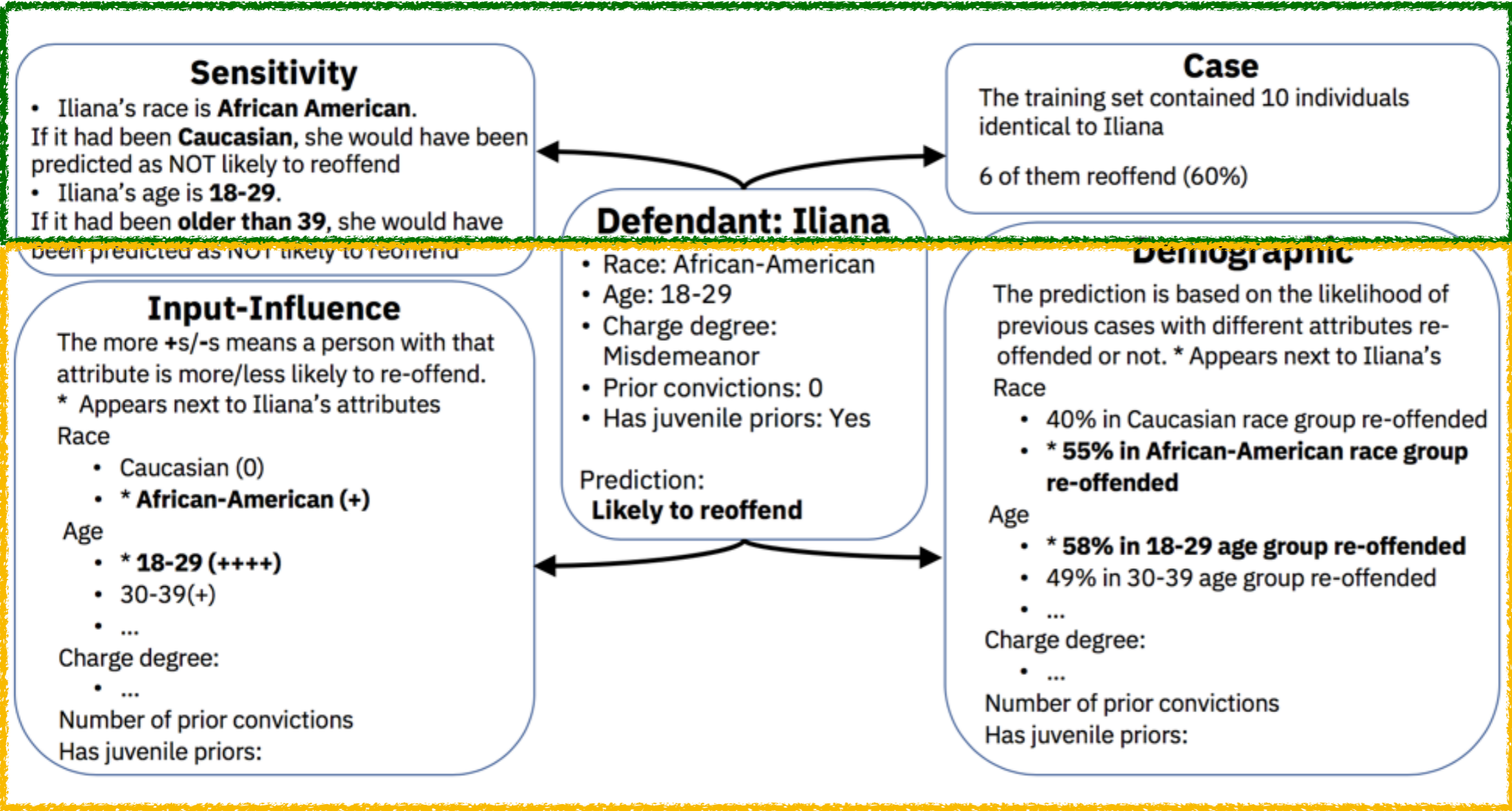
- \* **58% in 18-29 age group re-offended**
- 49% in 30-39 age group re-offended
- ...

Charge degree:

- ...

Number of prior convictions

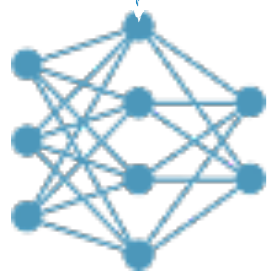
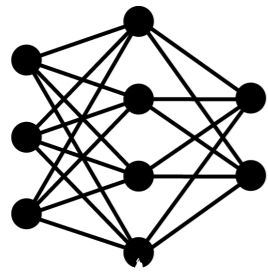
Has juvenile priors:



**Local explanations**  
**Why**

**Global explanations**  
**How**

# Research questions



Fairness  
**calibration**

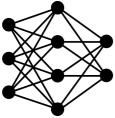

How do different styles of explanation impact fairness judgment?

- **Fairness calibration?**
- Surfacing **individual fairness** issue—**similar individuals receiving different treatment?**
- Perceived **inherently less fair?**

How do individual factors mediate the impact?

# Experimental design

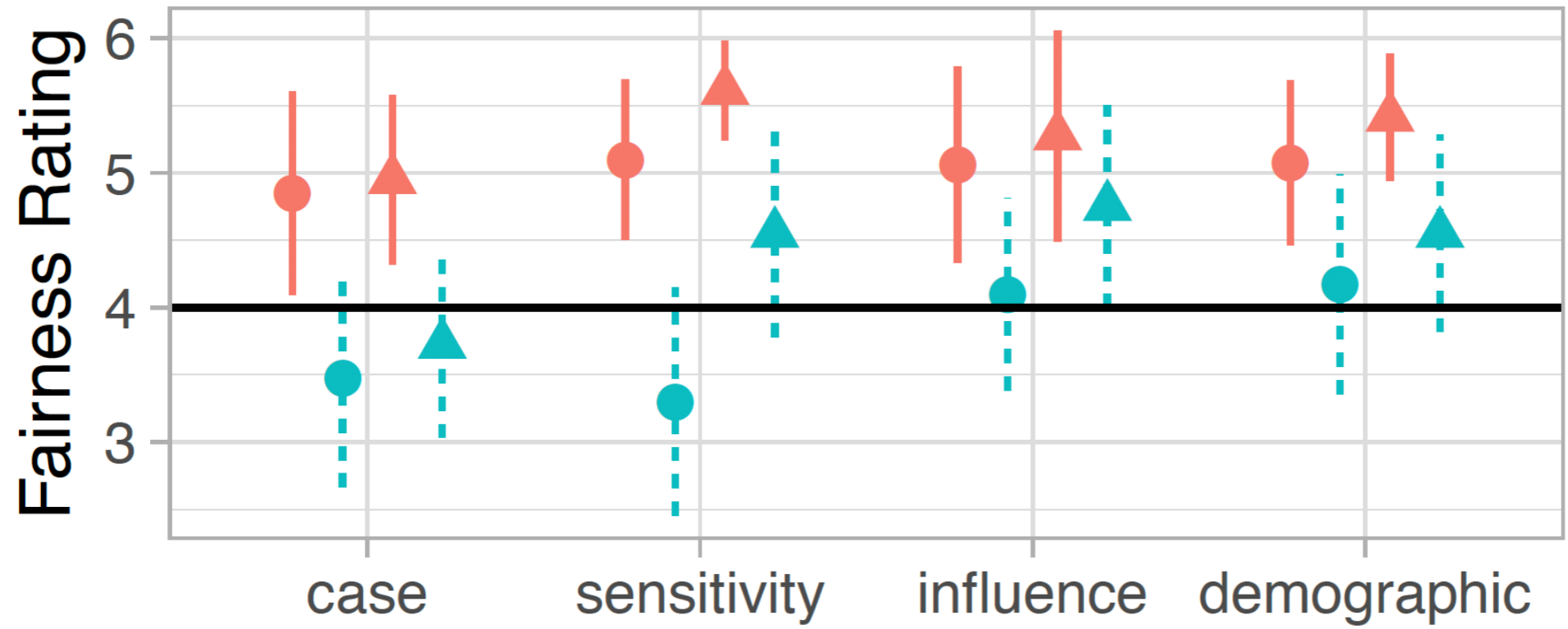
160 MTurk participants

	Input influence	Data demographic	Sensitivity	Case based
	20	20	20	20
	20	20	20	20

Sampled 6 instances from test data, oversampled 1/3 disparately impacted individuals (*individual unfairness*)

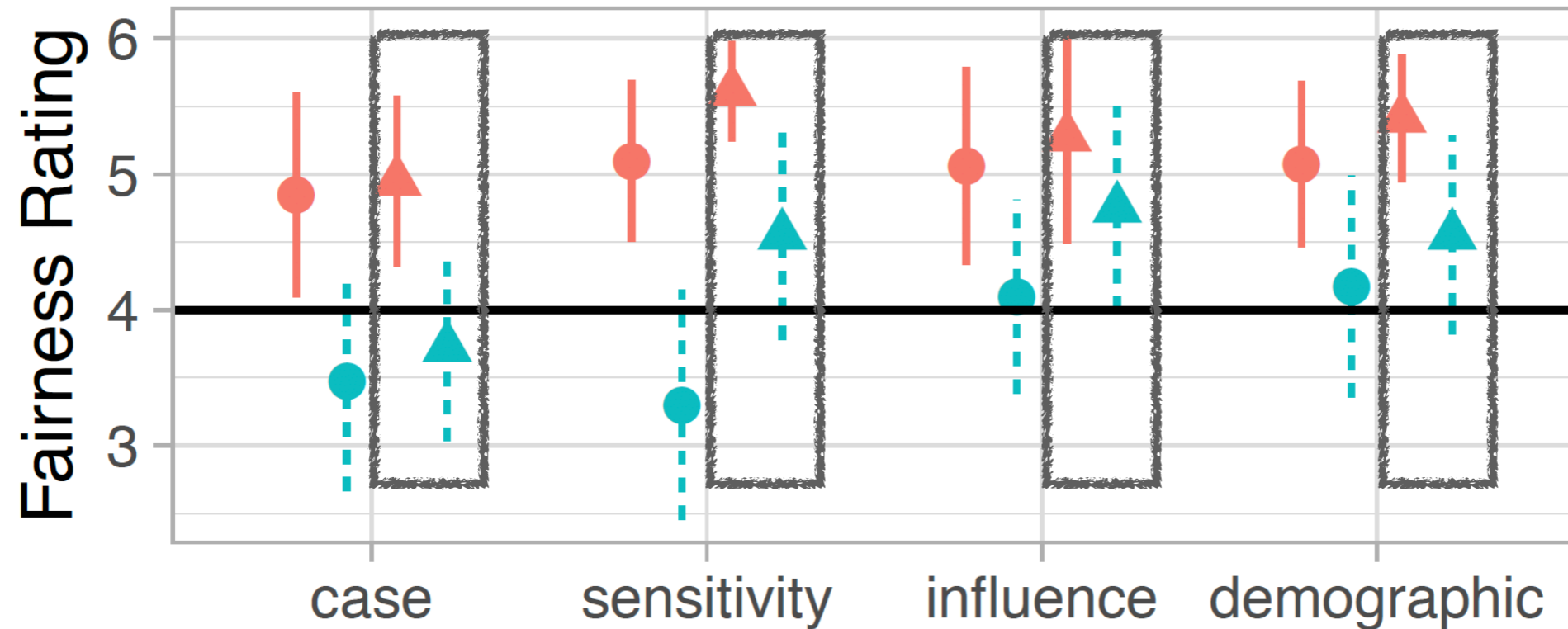
Present model prediction and explanation  
Rate “**how the software made the prediction was fair**” and explain the rating

Survey: demographic, prior position on general ML fairness, fairness of race feature, cognitive styles



Legend: data process (*raw*=●, *processed*=▲), and sample group (*impacted*=blue dashed lines, *non-impacted*=red solid lines)

# Fairness calibration?

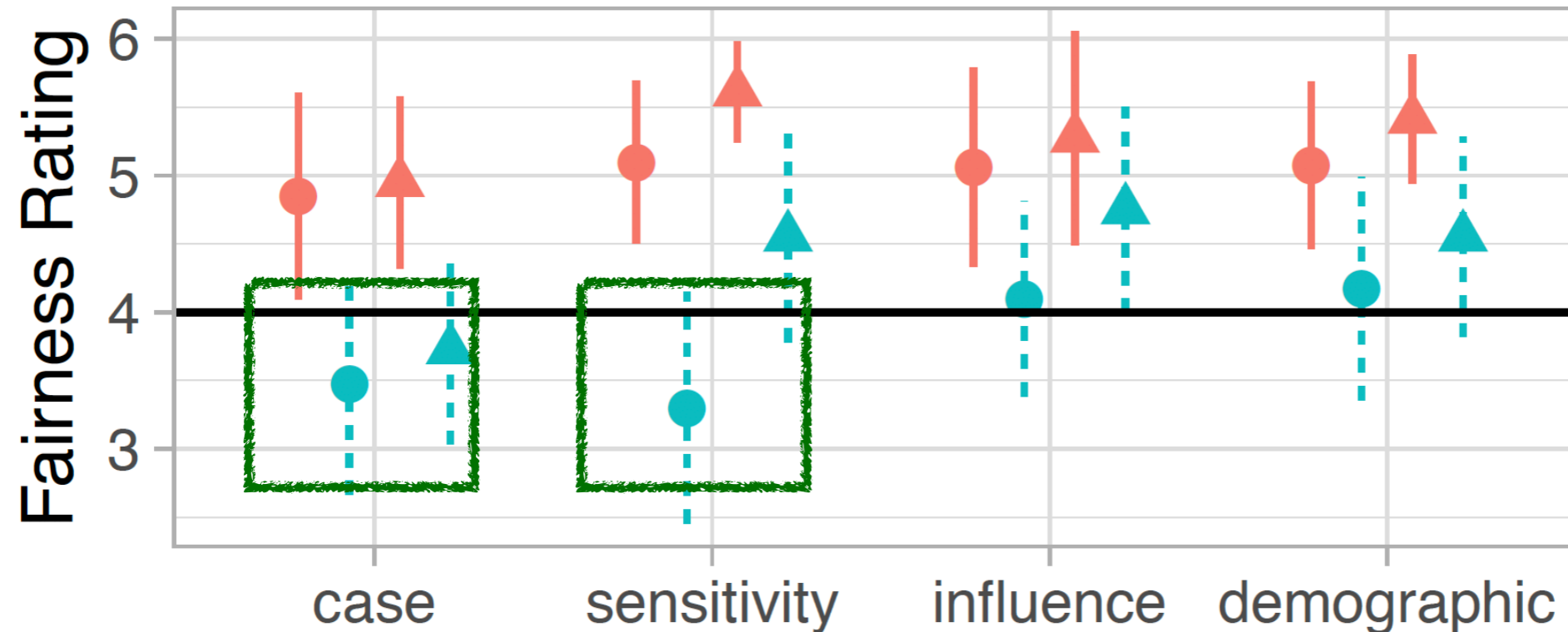


Legend: data process (*raw*=●, *processed*=▲), and sample group (*impacted*=blue dashed lines, *non-impacted*=red solid lines)

**All styles of explanation supported fairness calibration**



# Surfacing individual fairness issue?



Legend: data process (*raw*=●, *processed*=▲), and sample group (*impacted*=blue dashed lines, *non-impacted*=red solid lines)

**Local (why) explanations are more effective in surfacing individual fairness issue**

### Sensitivity

- Iliana's race is **African American**.  
If it had been **Caucasian**, she would have been predicted as NOT likely to reoffend
- Iliana's age is **18-29**.  
If it had been **older than 39**, she would have been predicted as NOT likely to reoffend

### Input-Influence

The more +s/-s means a person with that attribute is more/less likely to re-offend.

\* Appears next to Iliana's attributes

Race

- Caucasian (0)
- \* **African-American (+)**

Age

- \* **18-29 (++++)**
- 30-39(+)
- ...

Charge degree:

- ...

Number of prior convictions

Has juvenile priors:

### Case

The training set contained 10 individuals identical to Iliana

6 of them reoffend (60%)

### Defendant: Iliana

- Race: African-American
- Age: 18-29
- Charge degree: Misdemeanor
- Prior convictions: 0
- Has juvenile priors: Yes

Prediction:

**Likely to reoffend**

### Demographic

The prediction is based on the likelihood of previous cases with different attributes re-offended or not. \* Appears next to Iliana's

Race

- 40% in Caucasian race group re-offended
- \* **55% in African-American race group re-offended**

Age

- \* **58% in 18-29 age group re-offended**
- 49% in 30-39 age group re-offended
- ...

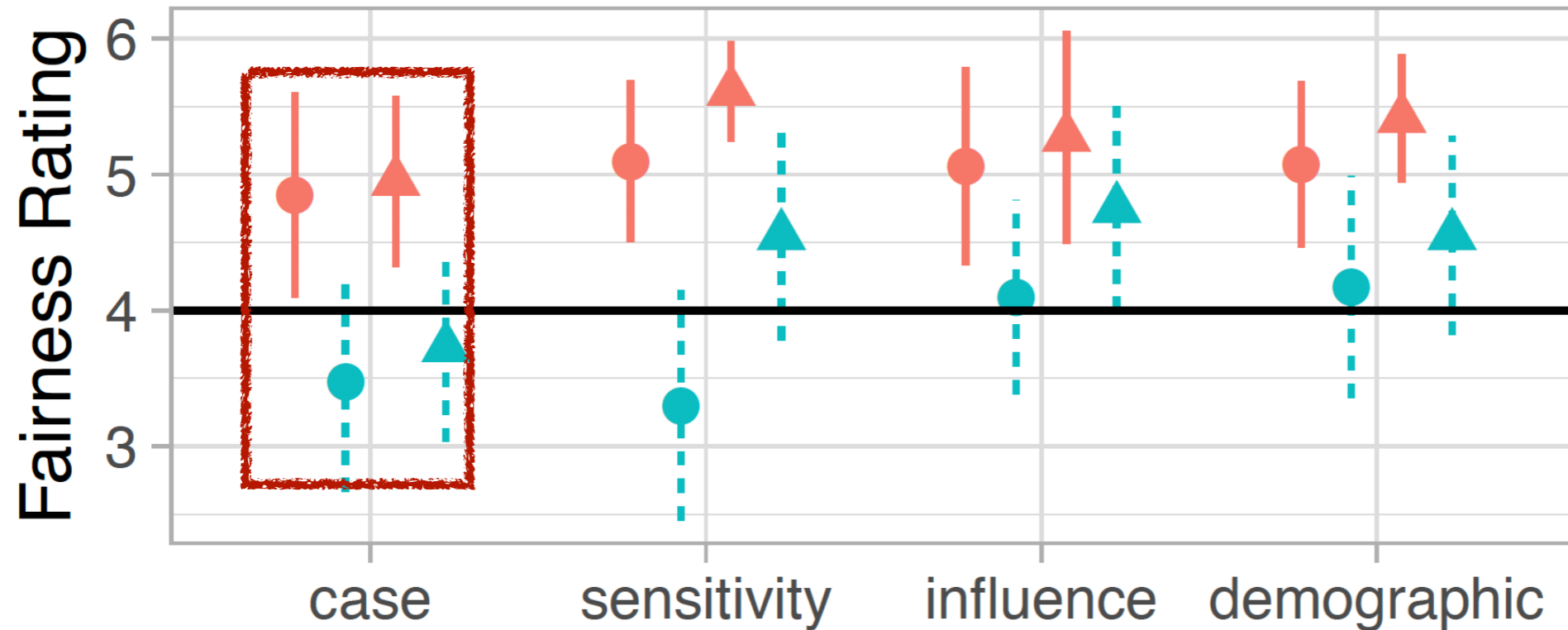
Charge degree:

- ...

Number of prior convictions

Has juvenile priors:

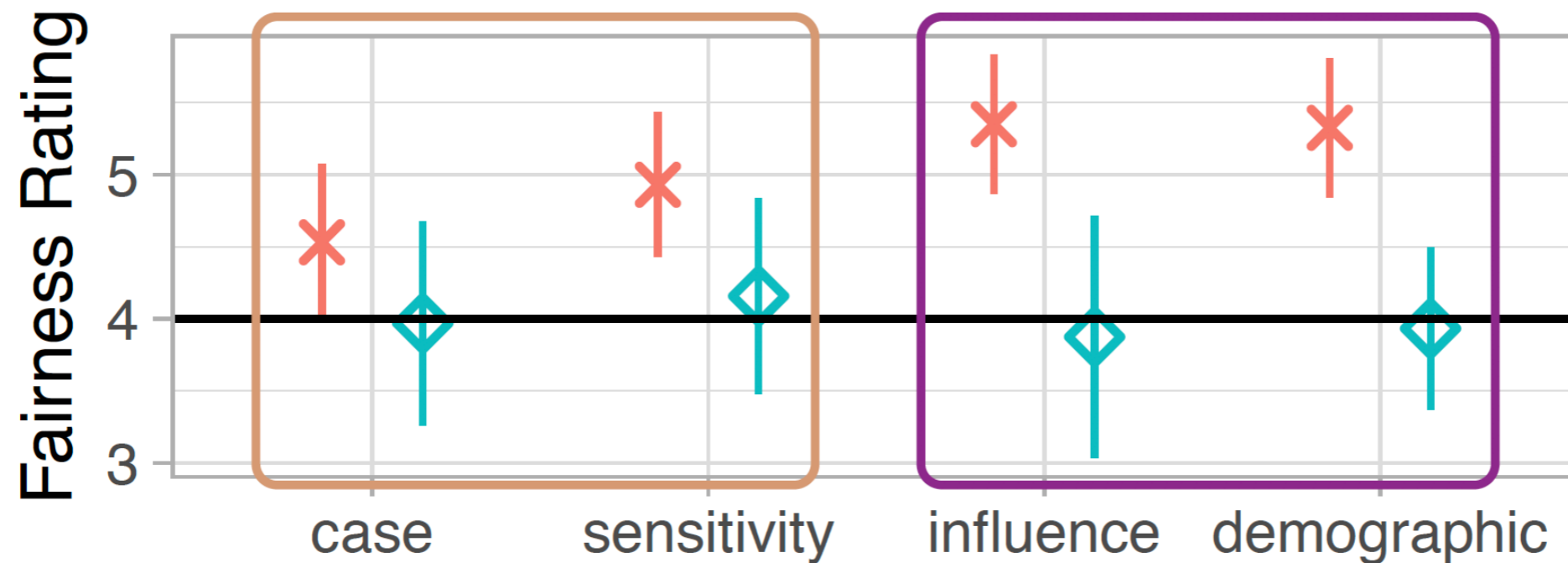
# Inherently less fair?



Legend: data process (*raw*=●, *processed*=▲), and sample group (*impacted*=blue dashed lines, *non-impacted*=red solid lines)

**Case-based explanation is perceived to be inherently less fair**

# Individual differences: prior position on ML fairness



Participants who consider “*ML fair to use*” (x) rated the system to be fairer when presented with **global explanations**

# Design guidelines: XAI supporting model scrutinization

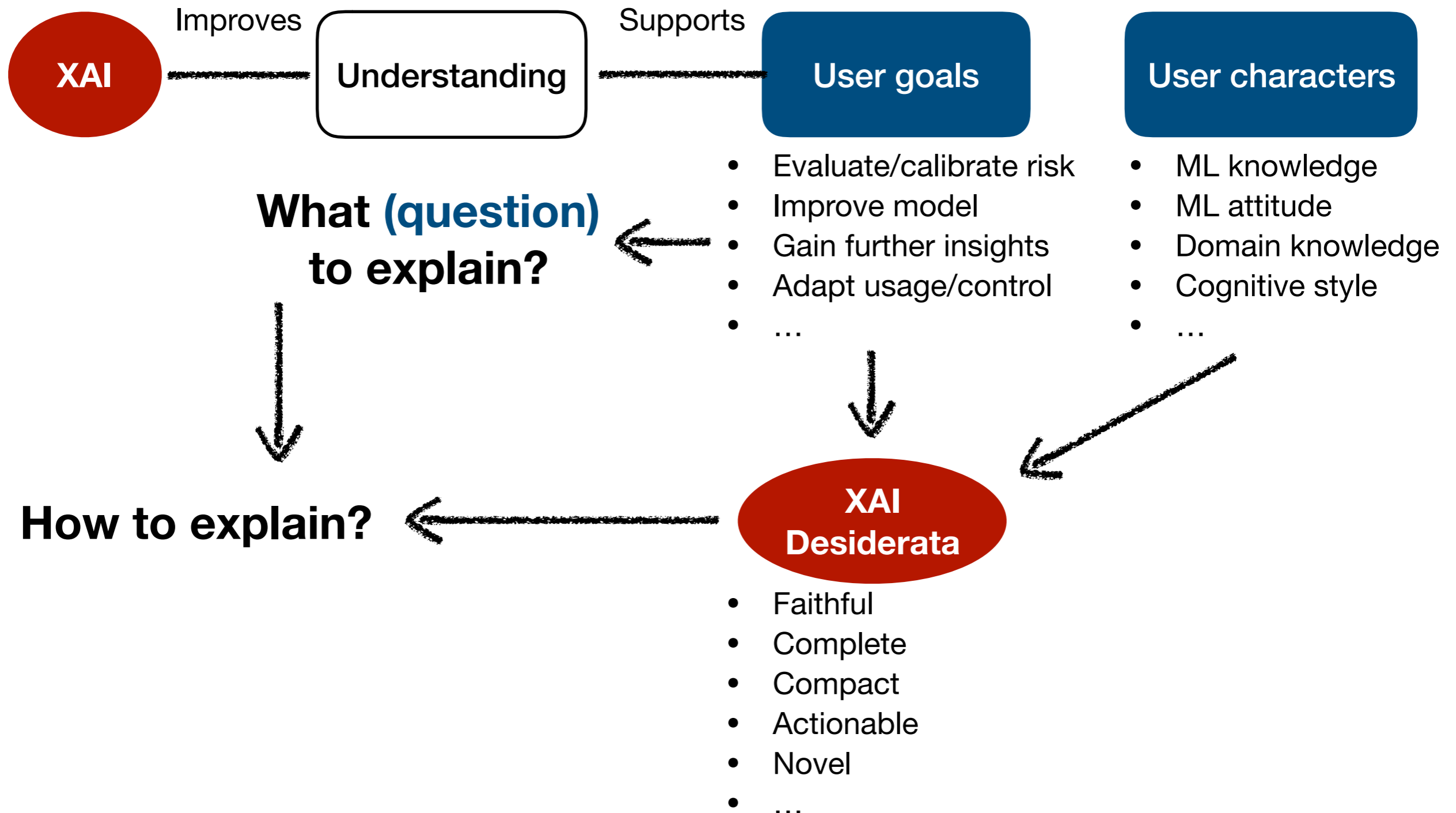
Design and evaluate with the **goal of calibration**

- Start with “ground truth” of model biases/problems

## **No one-size-fits-all**

- Types of fairness problems
- Offsetting v.s. accommodating individual difference
- Fine-grained scrutinization: Data, feature fairness, feature importance, feature interaction, procedural fairness

# Concluding remarks: toward **contextualized** and **actionable** human-centered XAI



# Thank **YOU!**

...and thanks to

Rachel Bellamy, Amit Dhurandhar, Jonathan Dodge, Bhavya Ghai, Werner Geyer, Daniel Gruen, Michael Hind, Stephanie Houde, David Millen, Aleksandra Mojsilović, Sarah Miller, Michael Muller, John Richards, Chenhao Tan, Richard Tomsett, Kush Varshney, Justin Weisz, Yunfeng Zhang

Q. Vera Liao  
[vera.liao@ibm.com](mailto:vera.liao@ibm.com)  
[www.qveraliao.com](http://www.qveraliao.com)  
@QVeraLiao