Advances and Open Questions in Explainable AI (XAI):

A practical perspective from an HCI researcher

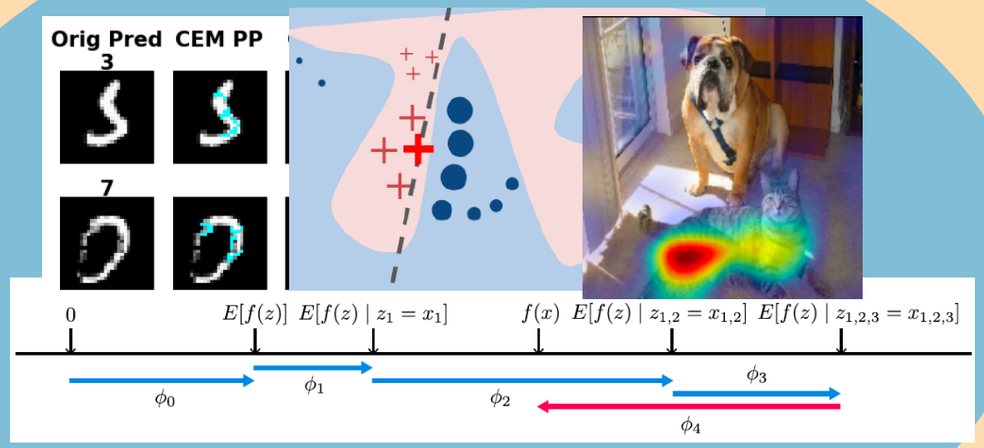
Q. Vera Liao

IBM **Research**

What are the most significant **advances in XAI**?

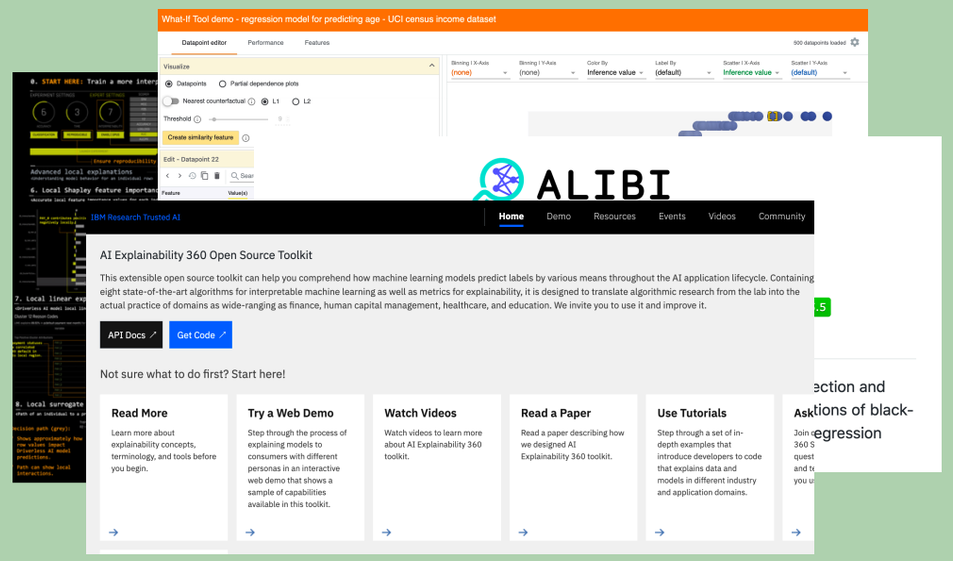
How might they influence **academia vs. industry**?

XAI in Academia



An abundance of XAI algorithms

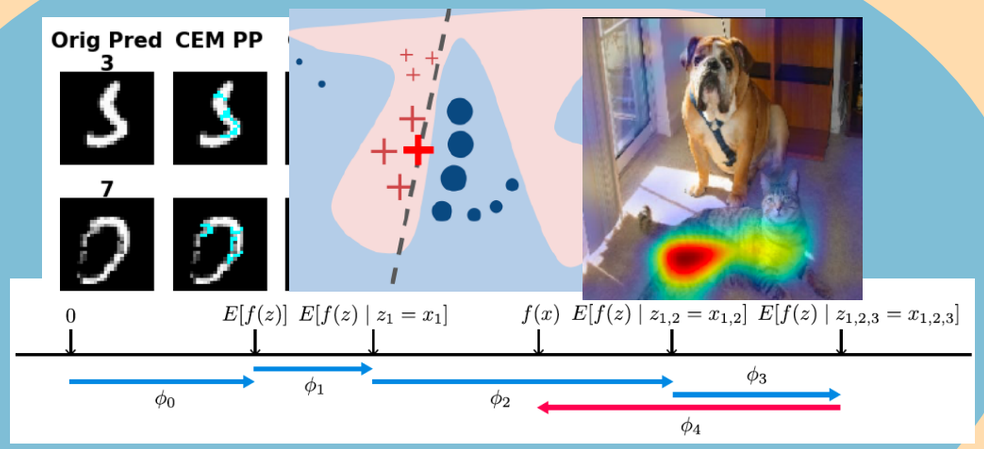
XAI in Practice



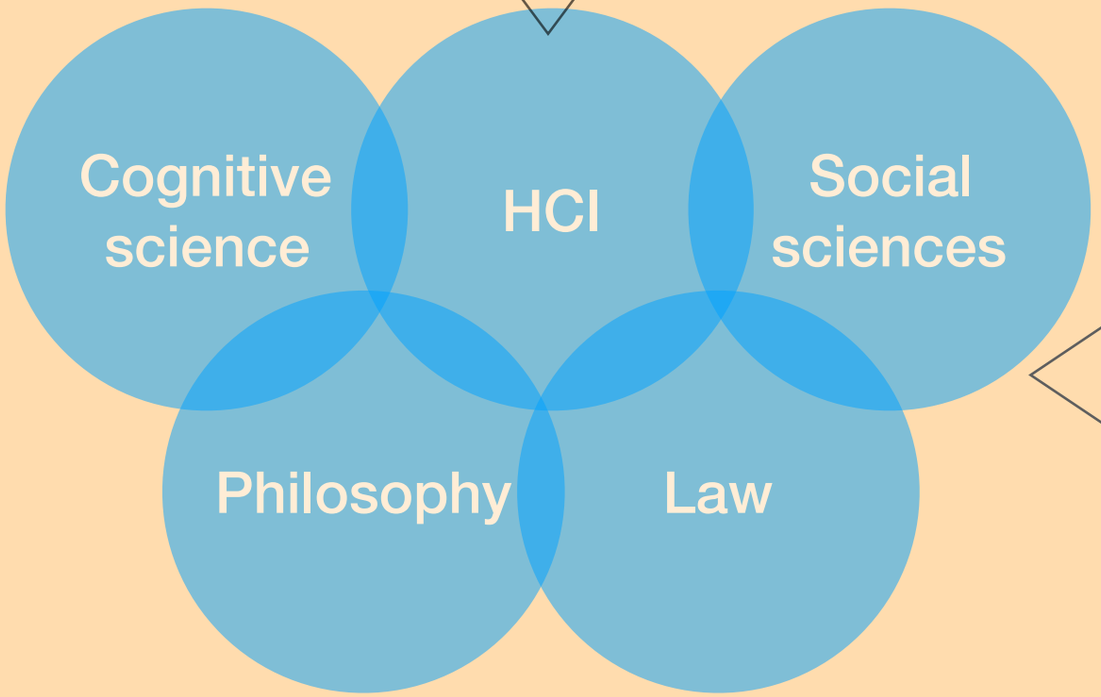
Toolbox of XAI (ML)

Advance 1: From academic research into a practitioners' toolbox

XAI in Academia

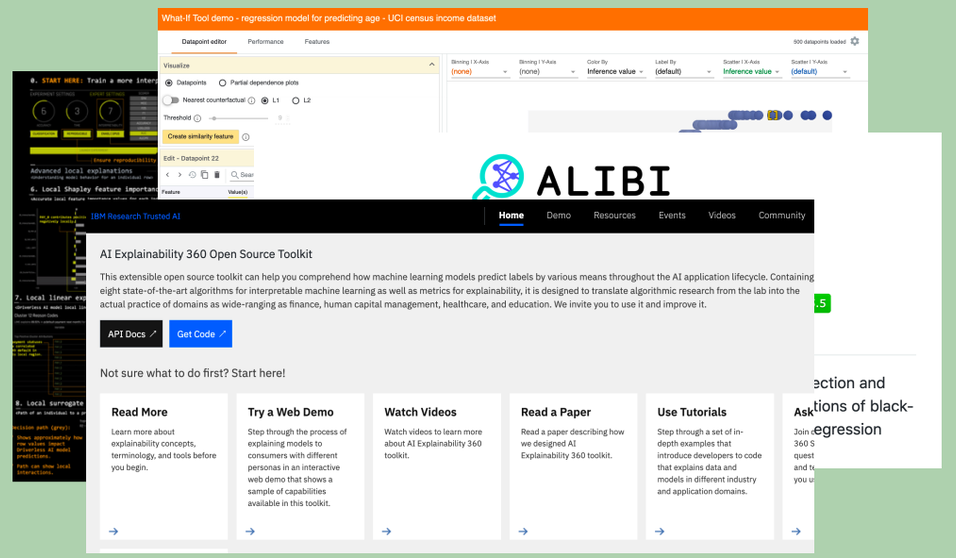


An abundance of XAI algorithms

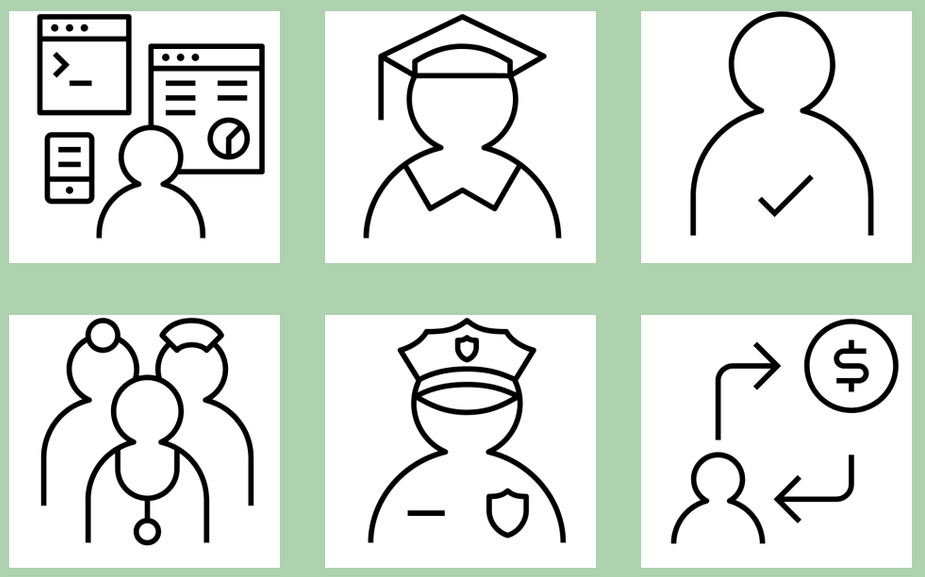


Multi-disciplinary perspectives

XAI in Practice

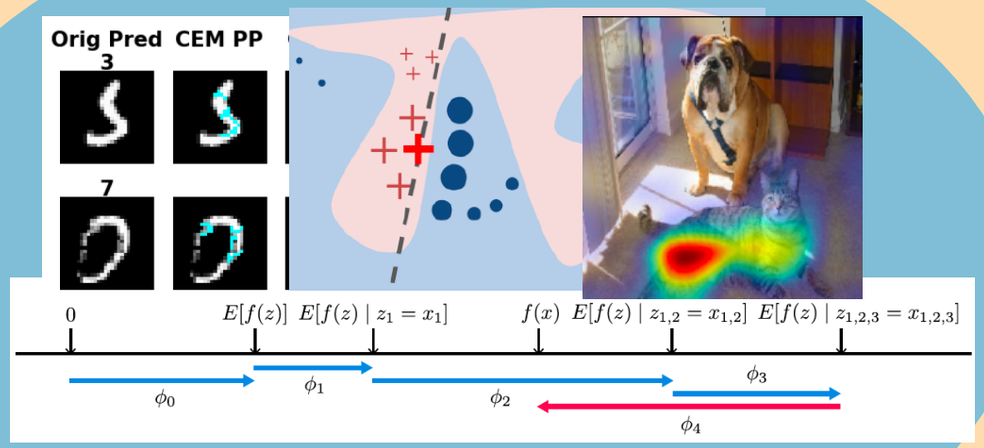


Toolbox of XAI

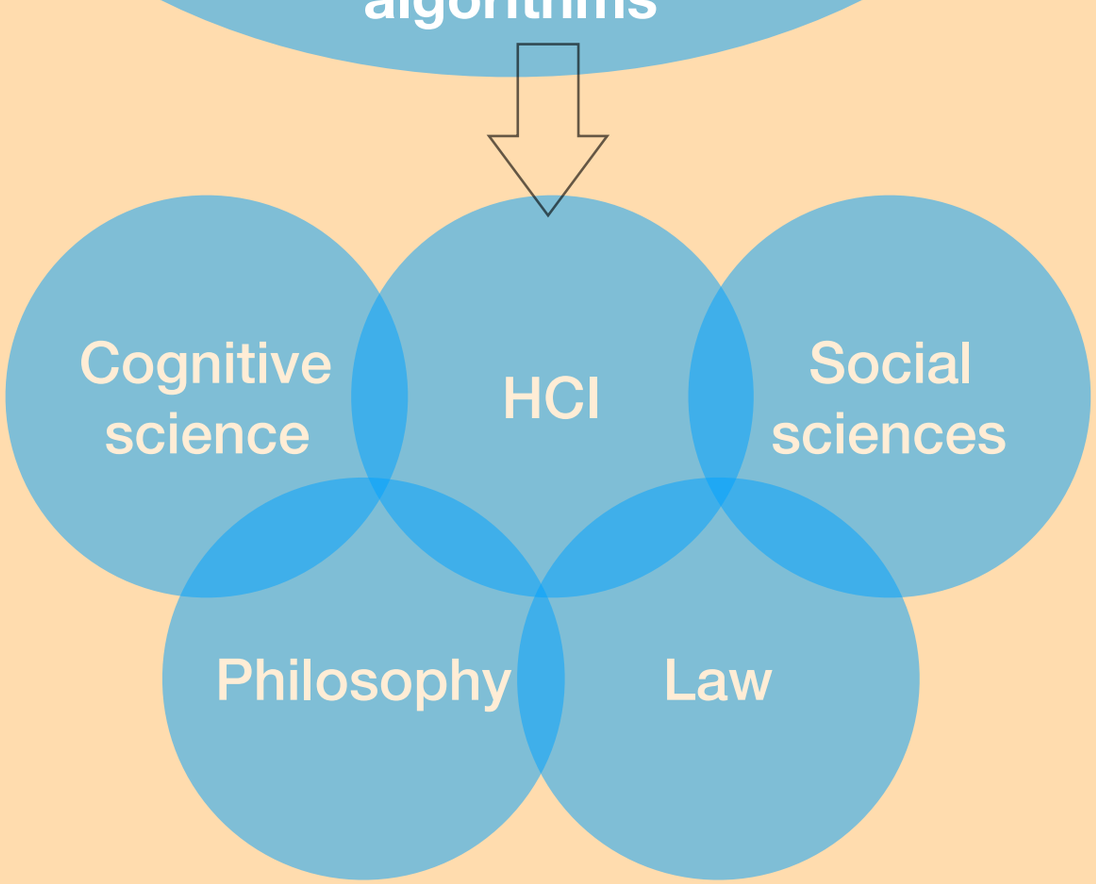


Towards real-world XAI: serving many domains and user groups

XAI in Academia



An abundance of XAI algorithms



Multi-disciplinary perspectives

Advance 2: Towards interdisciplinary perspectives

- The **gaps between XAI output and human explanations:** contrastive, selective, socially interactive (Miller 2019; Mittelstadt et al. 2019)
- The **plurality of motivation** for explanation: diagnosis, predicting the future, sense-making, justification, reconciling dissonance, etc. (Kiel 2006; Lombrozo, 2006)
- Explanatory power is **recipient dependent,** including the question they ask (explanatory relevance) (Hilton, 1990; Walton, 2004)
- More complexities:
  - The plurality of cognitive processes (Petty and Cacioppo, 1986; Kahneman, 2003)
  - Socio-technical systems (Ehsan et al., 2021)

Advances impacting practices

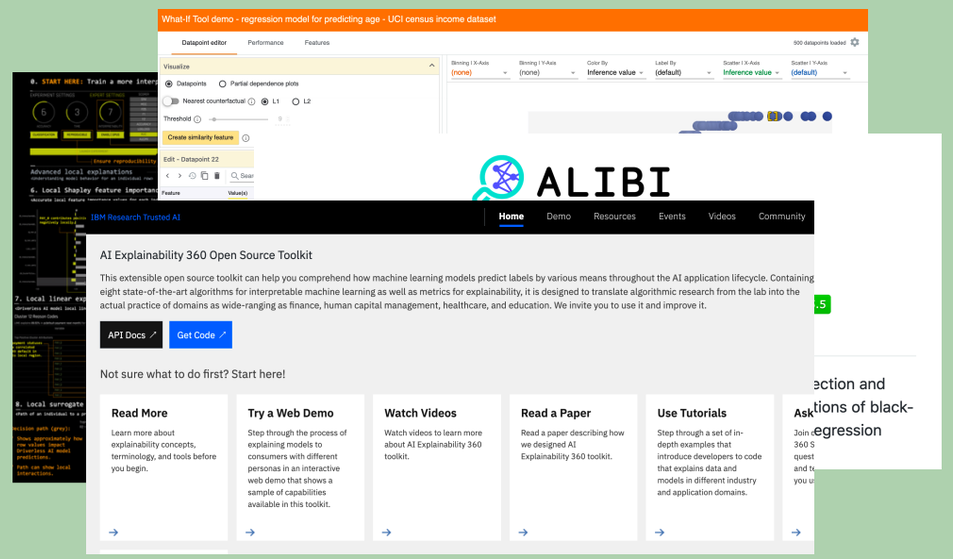
With a toolbox:

How to **select**?

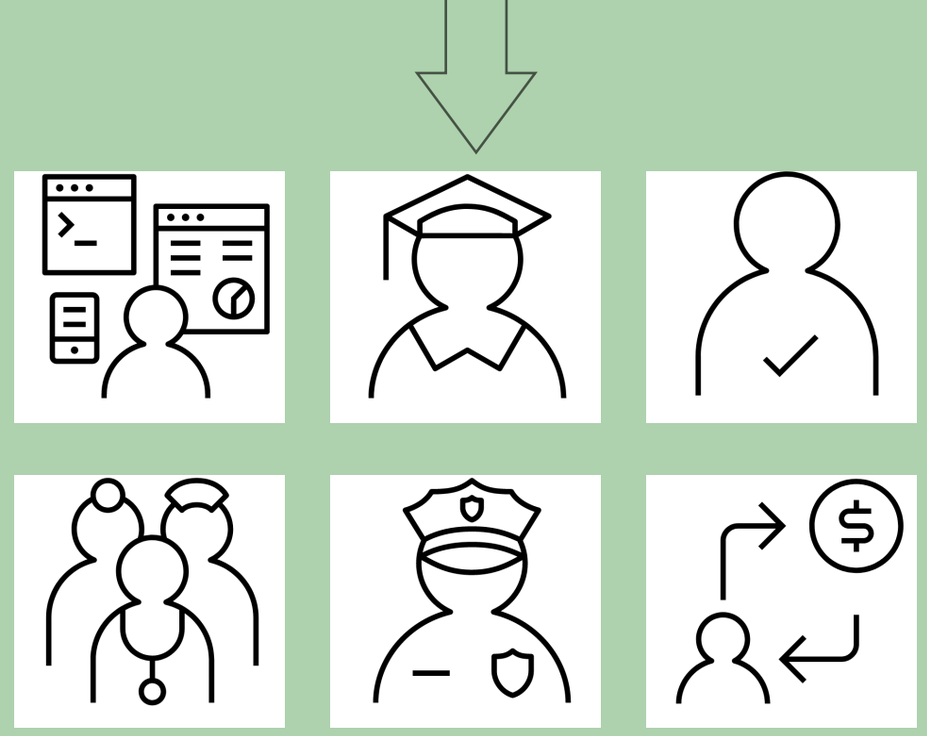
How to **translate**?

- Taxonomies and guidelines
- Empirical design and evaluation

XAI in Practice

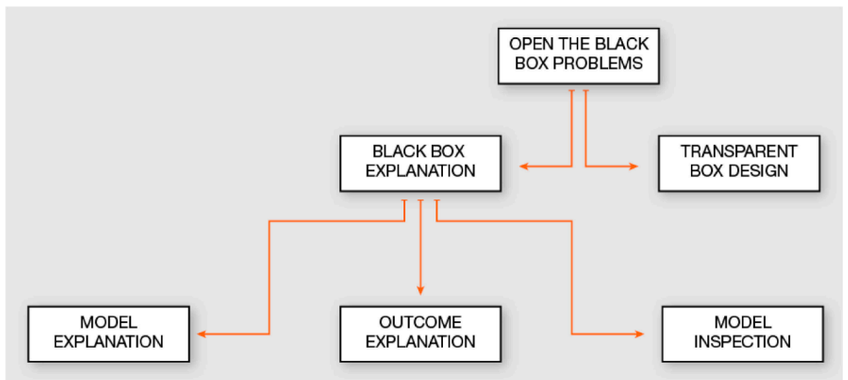


Toolbox of XAI

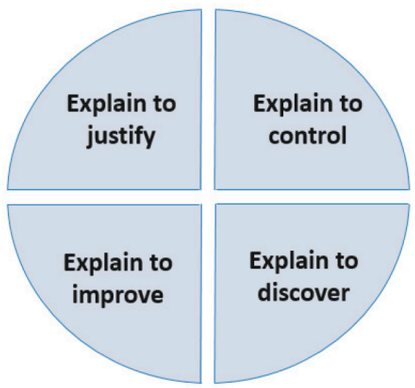


Towards real-world XAI: serving many domains and user groups

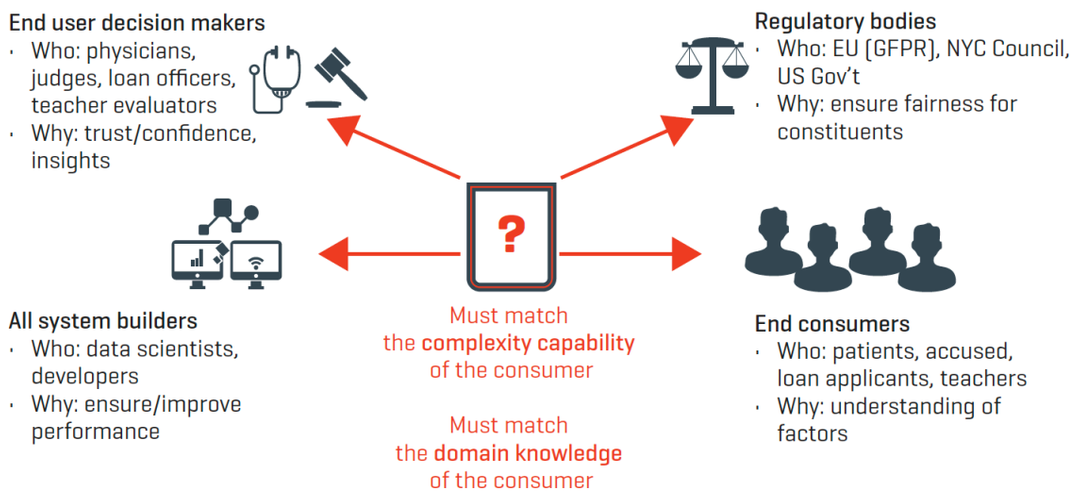
Taxonomies and guidelines



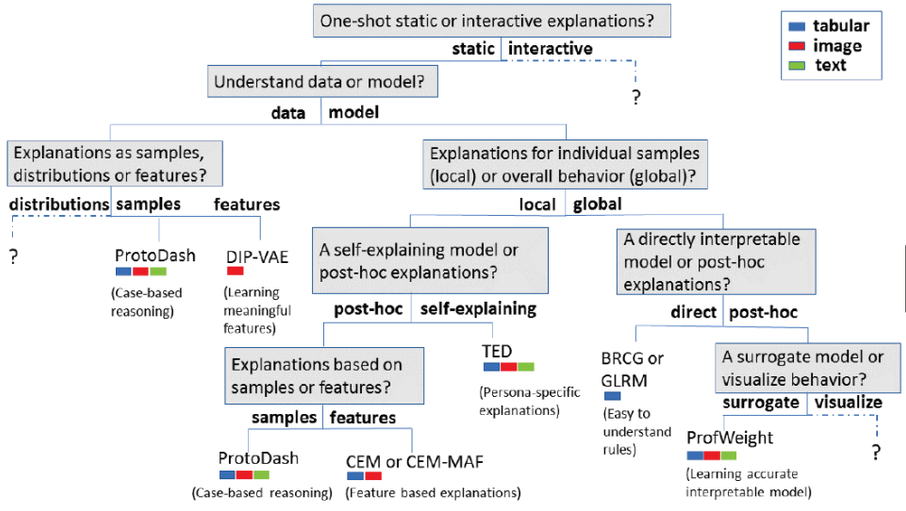
(Guidotti et al, 2018)



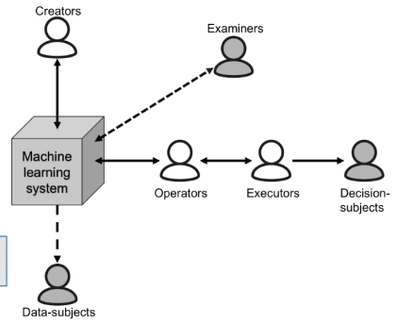
(Adadi et al, 2018)



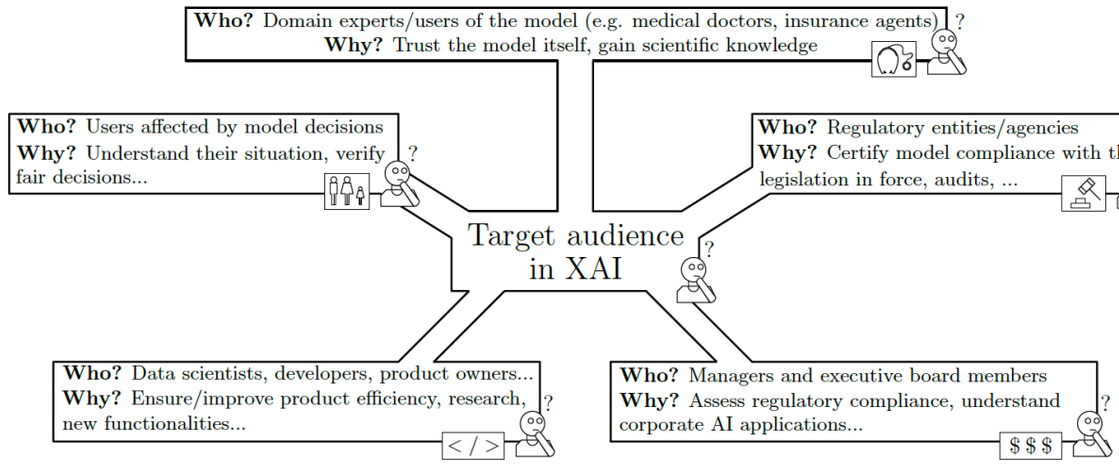
(Hind et al, 2019)



(Arya et al, 2019)



(Tomsett et al, 2019)

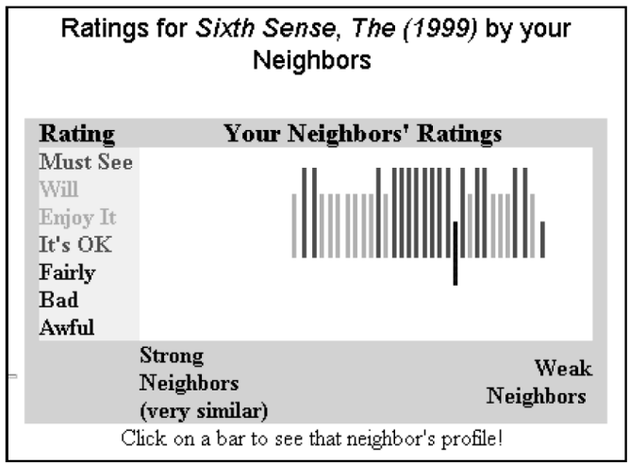


(Arrieta et al, 2019)

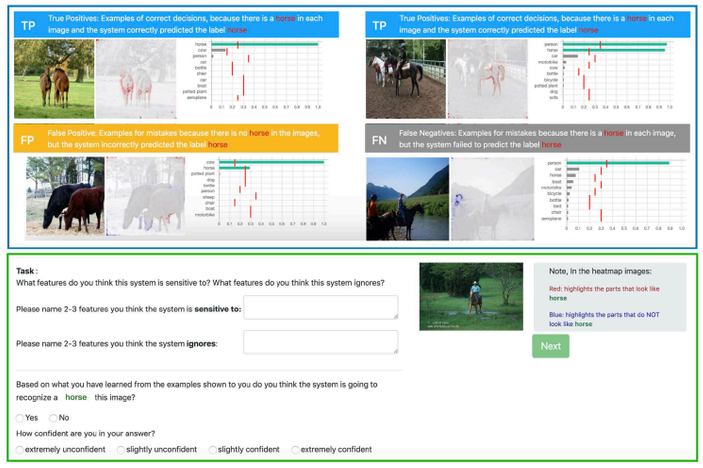
Categorize XAI techniques

Categorize Contexts and User Groups for XAI

Design and evaluation of XAI: HCI work

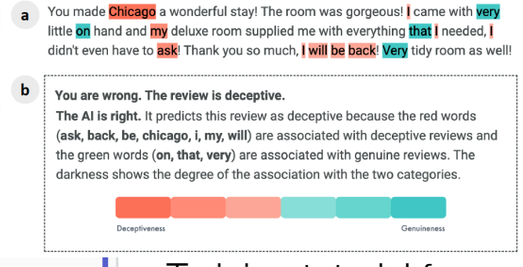


Recommender systems
(Herlocker et al. @ CHI 2000)

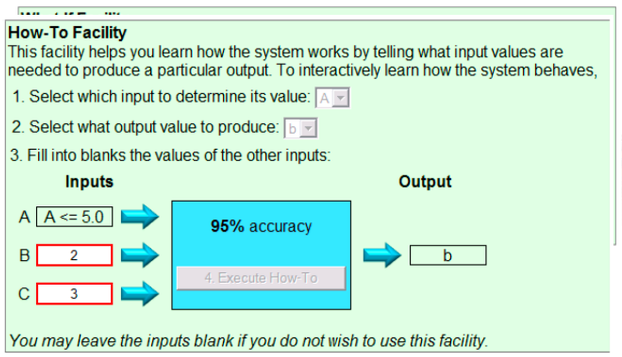


Saliency maps provide limited help for general understanding and simulatability of models
(Alqaraawi et al. @ IUI 2020)

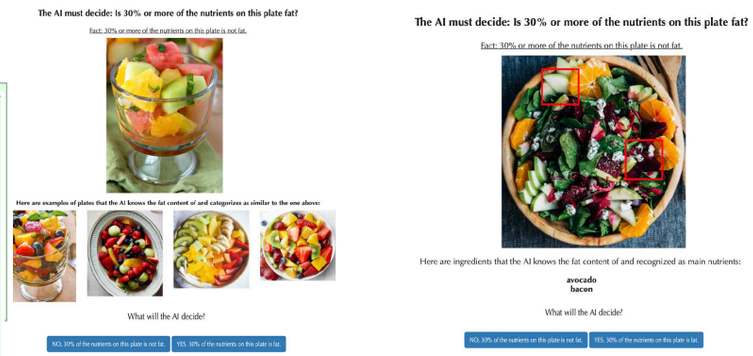
Model evaluation and debugging
(Hohman et al. @ CHI 2019)



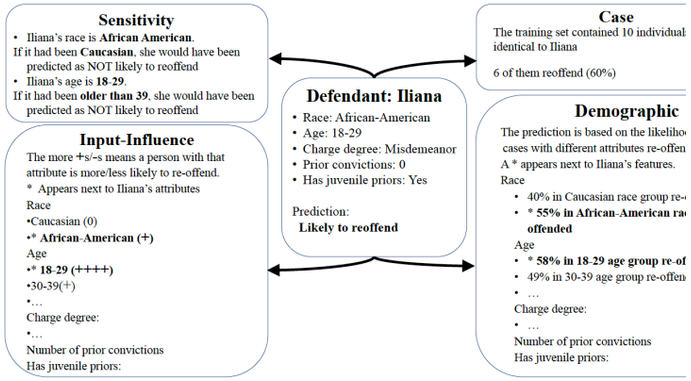
Training tutorial for
content moderation
(Lai at al. @ CHI 2020)



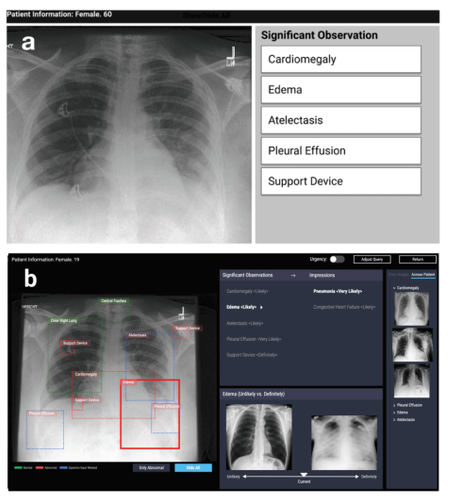
Context-aware systems
(Lim et al. @ CHI 2009)



Better simulatability in "proxy task" may not help decision-making in actual usage
(Buçinca et al. @ IUI 2020)



Model fairness judgment
(Dodge et al. @ IUI 2019)



Medical image analysis
(Xie at al. @ CHI 2020)

Pre-ML

Evaluate and "critique" XAI

Translate and evaluate XAI in specific domains

What are the most significant advances in XAI?
How might they influence industry?

- Advance 1: From academic research to a **practitioners' toolbox**
- Advance 2: Interdisciplinary perspectives inform the **selection** and **translation** of the toolbox

Plurality of motivation

Recipient dependent

Gaps with human explanation

Complexity of cognitive and social processes

Framework of XAI contexts

Framework of XAI user groups

Empirical research to design and evaluate

Not a solved problem yet… more on open questions!

What are the most significant **open questions in XAI**?

How might they influence **academia vs. industry**?

Open Question 1, with a toolbox:

How to **select**?

How to **translate**?

Top-down guidelines are not enough due to limitations in, e.g. *granularity* and *scalability*

Possible paths:

- Allow users to select: **intelligent and interactive XAI**
- Tackle the design process: **user centered design of XAI**

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. (Under review)

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)

How to select: identify user needs for XAI as ***questions***

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

Question	Explanations	Example XAI techniques
Global how	• Describe what algorithm is used and what features are considered, if a user is only interested in a high-level view • Describe the general model logic as feature impact*, rules✣ or decision-trees• (sometimes need to explain with a surrogate simple model)	ProfWeight*, Feature Importance*, PDP*, BRCG✣ , GLRM✣ , Rule List✣ , DT Surrogate•
Why	• Describe what key features of the particular instance determine the model's prediction of it* • Describe rules✣ that the instance fits to guarantee the prediction • Show similar examples• with the same predicted outcome to justify the model's prediction	LIME*, SHAP*, LOCO*, Anchors✣, ProtoDash•
Why not	• Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction* • Show prototypical examples✣ that had the alternative outcome	CEM* , Prototype counterfactual✣ , ProtoDash✣ (on alternative class)
How to be that	• Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction* • Show examples with small differences but had a different outcome than the prediction✣	CEM*, Counterfactuals*, DiCE✣
What if	• Show how the prediction changes corresponding to the inquired change	PDP, ALE, What-if Tool
How to still be this	• Describe 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✣
Performance	• Provide performance metrics of the model • Show confidence information for each prediction • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC Confidence 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

How to translate: support collaborative problem-solving between data scientists and designers with "***boundary objects***"

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

Why is this patient predicted of this risk?

What made him high-risk?

What are his risk factors?

Why

What can be done to reduce the patient's risk?

What worked for other patients with similar profiles?

How to be that

On what types of patient might it work worse?

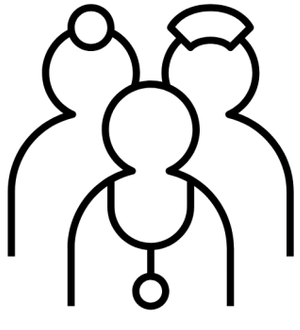
`How well does it work?

Performance

Is the training data similar to my patients?

What is the population of the training data?

Data



Rogers, Steve
MRN: 111111

Age	Sex	Race	Charlson Comorbidity Index
78	M	Black	COPD, PVD, Type 2 DM (2% 10-year survival)

History

Last 12 mo

Admissions	1
Emergency Dept	0
Hospital Acquired Conditions	0

30 day risk of all cause admission

1. Data

View data sources

0% 15% 25%

Low | Moderate | High

Risk score confidence: Good (+/- 2%)

30 day admission risk

5 %

(1 in 20 chance)

Medicare average 16%

average 13%

2. Performance

30 day all cause admissions

Data Sources

Medicare Claims data (2008-2011)

Characteristics of **212, 236** Medicare beneficiarie randomly selected and shared by CMS

Age	
<60	5%
60-69	35%
70-79	45%,
>=80	15%
Gender	
Male	51%
Female	49%
Race	
Caucasian	41%
Black	22%
Hispanic	18%
Other or unidentified	20%

What sources are NOT included?

There is no Medicare Part D (medications) data or any EHR data (labs, physiological data, notes) used in the prediction.

more deta

Factors that contribute to the risk of admission

3. Why

Decreases risk ← → Increases risk

Charleson Comorbidity Index (6 points, 13%)

Mood Disorders (yes)

ED Visits (4)

COPD (true)

Age < 80

Action impact

No pnemonia vaccine

Pnemonia vaccine

People like Steve who had a **pnemonia vaccine** had 3 percent point lower risk.

3 percent point lower risk

4. How to be that

Order

Active smoker

Smoking cessation

People like Steve who **don't smoke** have a 1 percent point lower risk.

1 percent point lower risk

View Programs

* This is made up patient data. No PHI is included

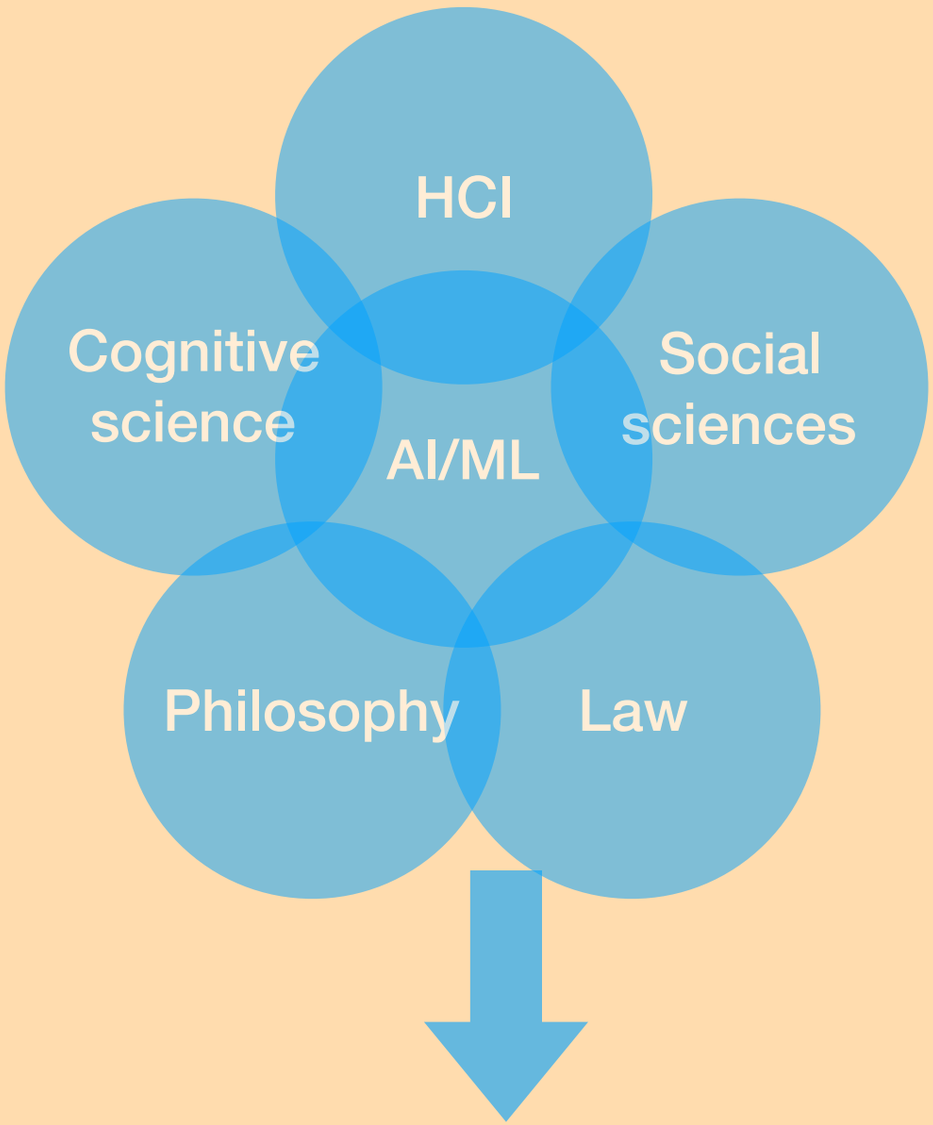
Risk factor to eliminate	Risk improvement
ED visits	-10%
Mood disorders	-9%
Pneumonia risk	-3%
Smoking	

5. How to be that (first version)

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

Open Question 2:

How to **expand the toolbox?**



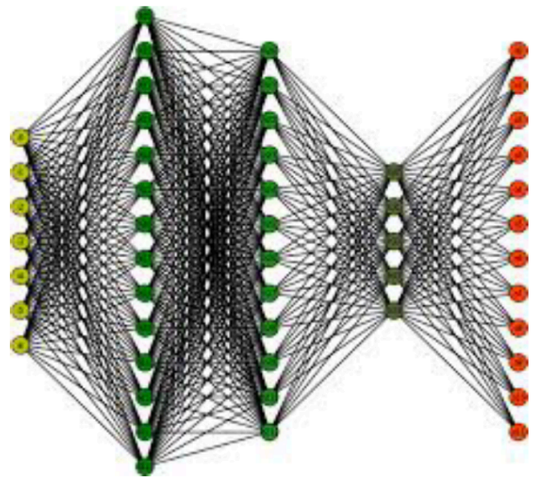
What's missing?

Toolbox of XAI

How to operationalize the missing element and make it an actionable tool?

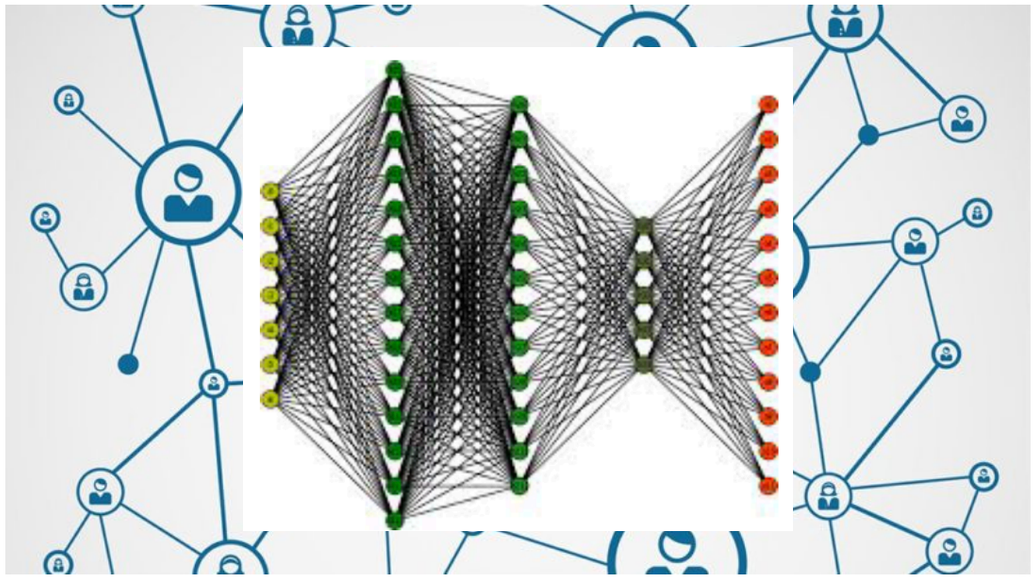
What’s missing? From AI explanations to explainability

Explainability goal:
sense-making for **better decisions**



Algorithmic explanation

Socially situated explainability



"Sense-making is not just about opening the closed box of AI, but also about who is around the box, and the socio-technical factors that govern the use of the AI system and the decision. ***Thus the 'ability' in explainability does not lie exclusively in the guts of the AI system****"*

Ehsan et al. Expanding Explainability: Towards Social Transparency in AI systems.To appear in CHI 2021

Towards socially situated explainability: an actionable framework

Customer: Scout Inc. **Product:** Access Management (SaaS) **Product ID (PID):** 43523X

Recommendation: Sell at $100 per account per month

Justification: the AI system considered the following components

[○] *Quota goals* [○] *Comparative pricing: what similar customers pay* [○] *Cost: $55 /account/month* 1

For this customer, 3 members of your team received pricing recommendations in past sales. However, 1 out 3 have sold at the recommended price. Click to see more details. 2

Nadia M.
Sales Assoc. (AB34)

Action: Reject Recommendation ⇔ **Outcome:** No Sale
Comment: Long-term profitable customer; main revenue from a different vertical ; selling at cost price to maintain relationship
Oct 2, 2019 3

Eric C.
Sales Manager (XZ89)

Action: Accept Recommendation ⇔ **Outcome:** Sale
Comment: Recommended price aligned with profit margins; customer felt the price was fair
Dec 14, 2019 4

4W
What
Who
Why
When

Jess W.
Sales Director (RE43)

Action: Reject Recommendation ⇔ **Outcome:** Sale
Comment: Covid-19 pandemic mode; cannot lose long-term profitable customer; offered 10% below cost price
May 6, 2020 5

Ehsan et al. Expanding Explainability: Towards Social Transparency in AI systems.To appear in CHI 2021

What are the most significant advances and open questions in XAI? How might they influence industry?

- Open question 1: How to actually **select** and **translate** the toolbox into good user experiences?
- Open question 2: How to **expand** the toolbox to support explainability/understanding?

Have answers? Submit to our virtual workshop on **Operationalizing Human-Centered Perspectives in Explainable AI** at **CHI 2021**

https://hcxai.jimdosite.com/ (Deadline February 14)

Thank **YOU!**

Q. Vera Liao
vera.liao@ibm.com
www.qveraliao.com
@QVeraLiao