Are Explanations Helpful?

A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making

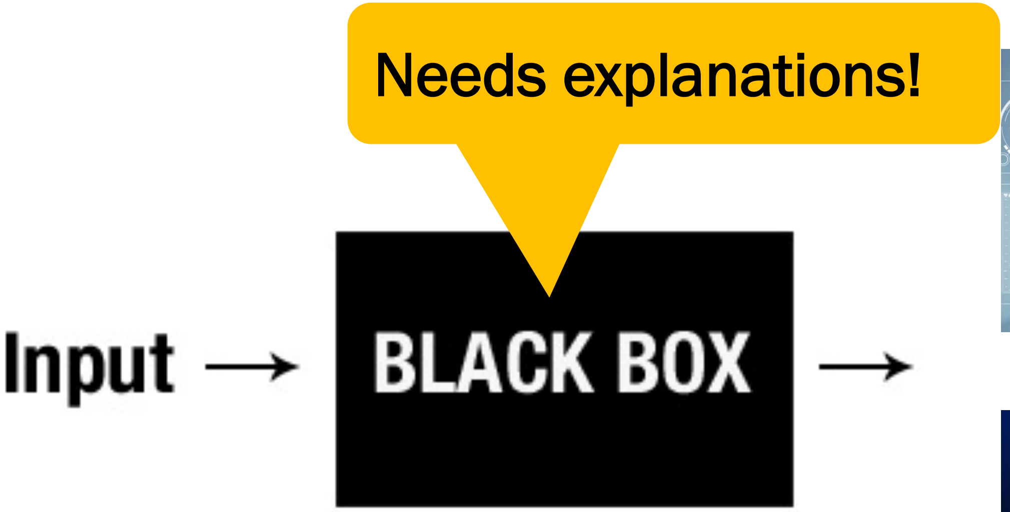
Xinru Wang, Ming Yin

Purdue University

IUI 2021

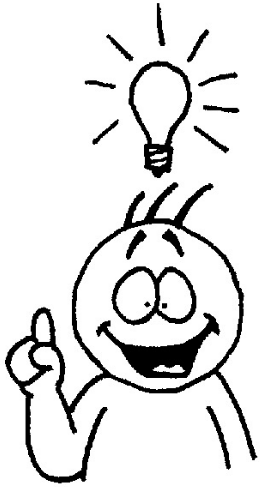
PURDUE UNIVERSITY®

AI-driven decision aids are everywhere…



What constitutes a "good" AI explanation?

Understanding



Uncertainty awareness

Trust calibration

What constitutes a “good” AI explanation?

Understanding

Uncertainty awareness

Trust calibration

Do different types of explanation satisfy these three desiderata?

What's the gap?

Publications	Decision making tasks	AI Explanation methods	Desideratum 1 (Understanding)	Desideratum 2 (Uncertainty awareness)	Desideratum 3 (Trust calibration)
Poursabzi-Sangdeh et al. [59]	house price prediction	intrinsically interpretable model	mixed results	N/A	✗?
Alqaraawi et al. [3]	image classification	saliency map	mixed results	N/A	N/A
Chu et al. [17]	age prediction	saliency map	N/A	N/A	✗?
Cheng et al. [16]	student admission	feature contribution	✓	N/A	N/A
Zhang et al. [71]	income prediction	feature contribution	N/A	✗	✗?
Bansal et al. [6]	sentiment analysis	feature contribution	N/A	N/A	✗
Carton et al. [14]	toxicity content detection	feature contribution	N/A	N/A	✗?
Lai and Tan [42]	deception detection	feature contribution	N/A	N/A	✓?
Lai et al. [41]	deception detection	feature contribution	N/A	N/A	✓?
Cai et al. [13]	drawing recognition	example-based	mixed results	N/A	N/A
Yang et al. [69]	leaf classification	example-based	N/A	N/A	✓

What’s the gap?

Publications	Decision making tasks	AI Explanation methods	Desideratum 1 (Understanding)	Desideratum 2 (Uncertainty awareness)	Desideratum 3 (Trust calibration)
Poursabzi-Sangdeh et al. [59]	house price prediction	intrinsically interpretable model	mixed results	N/A	✗?
Alqaraawi et al. [3]	image classification	saliency map	mixed results	N/A	N/A
Chu et al. [17]	age prediction	saliency map	N/A	N/A	✗?
Cheng et al. [16]	[illegible]	feature contribution	✓	N/A	N/A
Zhang et al. [71]	income prediction	feature contribution	N/A	✗	✗?
Bansal et al. [6]	sentiment analysis	feature contribution	N/A	N/A	✗
Carton et al. [14]	toxicity content detection	feature contribution	N/A	N/A	✗?
Lai and Tan [42]	deception detection	feature contribution	N/A	N/A	✓?
Lai et al. [41]	deception detection	feature contribution	N/A	N/A	✓?
Cai et al. [13]	drawing recognition	example-based	mixed results	N/A	N/A
Yang et al. [69]	leaf classification	example-based	N/A	N/A	✓

few studies on some desiderata

Incomplete and mixed results!

different AI explanations

different tasks

Decision-making tasks

Recidivism prediction

- Will this defendant reoffend within 2 years?

1. Race:	White	2. Gender:	male	3. Age:	45	4. Prior Count:	8
5. Charge Name:	Domestic Violence			6. Charge Degree:	misdemeanor	7. Days in Custody:	11

Forest cover prediction

- Is the primary tree species in this area spruce/fir?

1. Elevation	3086	2. Aspect	32	3. Slope	7	4. Hillshade Index at Noon	224
5. Horizontal Distance to Nearest Surface Water	175	6. Vertical Distance to Nearest Surface Water	2	7. Horizontal Distance to Nearest Roadway	5031	8. Horizontal Distance to Nearest Wildfire ignition Points	1034

Decision-making tasks

Recidivism prediction

- Will this defendant reoffend within 2 years?

1. Race:	White	2. Gender:	male	3. Age:	45	4. Prior Count:	8
5. Charge Name:	Domestic Violence				…meanor	7. Days in	11

83% of participants in a pilot study reported to have more prior knowledge

Forest cover prediction

- Is the primary tree species in this area spruce/fir?

1. Elevation	3086	2. Aspect	32	3. Slope	7	4. Hillshade Index at Noon	224
5. Horizontal Distance to Nearest Surface Water	175	6. Vertical Distance to Nearest Surface Water	2	7. Horizontal Distance to Nearest Roadway	5031	8. Horizontal Distance to Nearest Wildfire ignition Points	1034

Decision-making tasks

Prediction Task (1/33)

Please review the profile below and predict whether the defendant is likely to reoffend in the next two years. *If you don't remember the meaning of an feature, click on the red circle on that feature to view its meaning.*

Defendant Profile:

1. Race:	White	**2. Gender:**	male	**3. Age:**	45	**4. Prior Count:**	8
5. Charge Name:	Domestic Violence			**6. Charge Degree:**	misdemeanor	**7. Days in Custody:**	11

Make Your Prediction:

Do you think this defendant will reoffend within 2 years?

- Yes, I think this defendant **will** reoffend within 2 years.
- No, I think this defendant **will not** reoffend within 2 years.

Machine Learning Prediction:

Our machine learning model predicts that this person **will** reoffend in 2 years.

Make your final prediction:

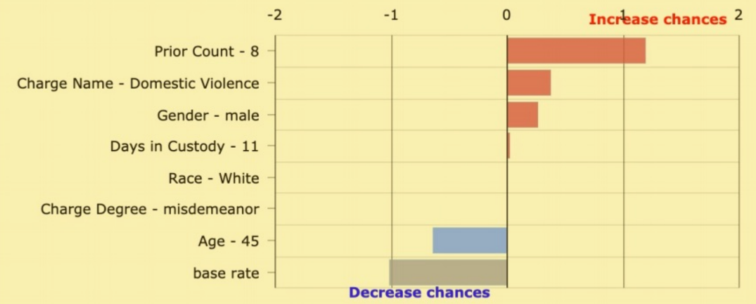
Now, do you think this defendant will reoffend within 2 years?

- Yes, I think this defendant **will** reoffend within 2 years.
- No, I think this defendant **will not** reoffend within 2 years.

Why did our machine learning model make this prediction?

Our machine learning model is trained on many previous defendant profiles for which whether the defendant reoffends is known. Our model has learned from these profiles that for each defendant, each feature of the defendant's profile can increase or decrease the defendant's chance of reoffending, depending on the value of the feature.

The chart below shows for each feature of this defendant's profile, whether it increases (**red bars**) or decreases (**blue bars**) his chance of reoffending, and by how much.



Our model always compares a defendant with the following reference defendant to determine whether each feature increases or decreases the chance of reoffending:

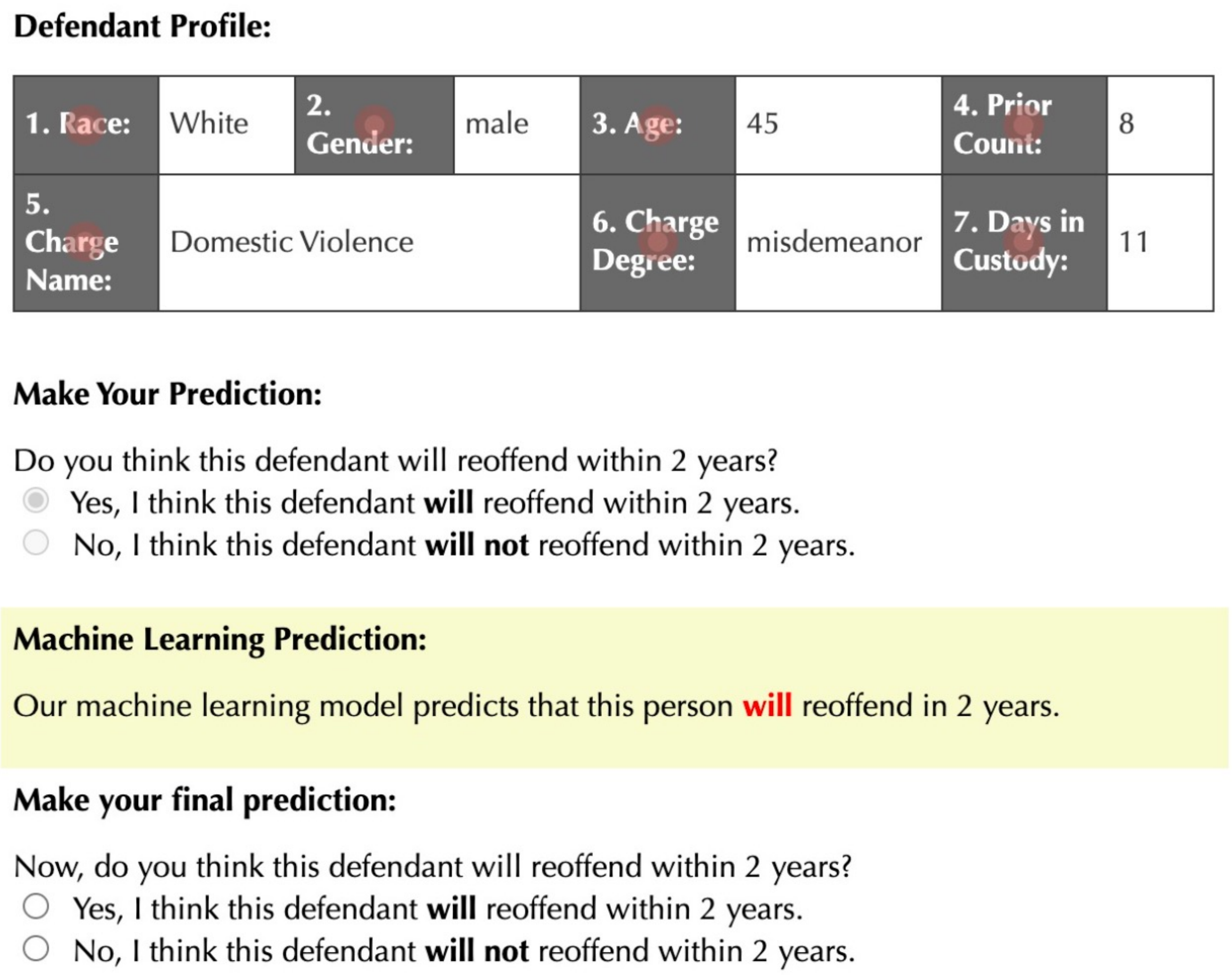
"A White female; aged 31; arrested for a misdemeanor without specific charge; has 0 priors; spent 0 days in custody."

The base chance for the reference defendant is very low, which is shown as the grey bar on the chart above.

Next

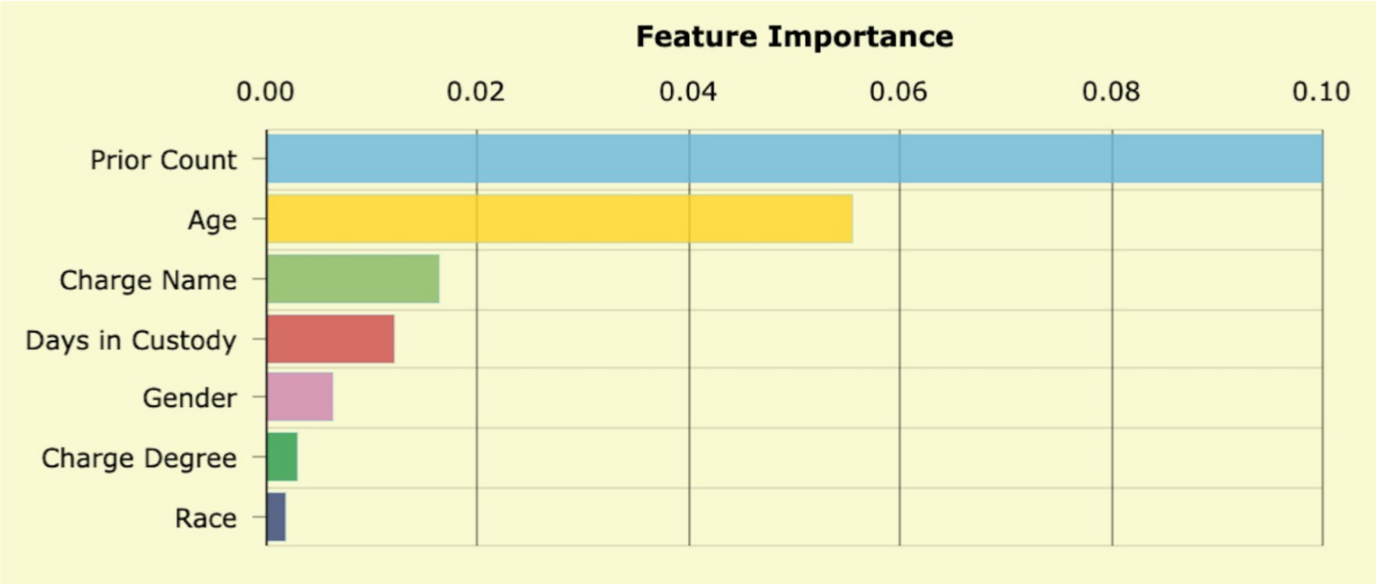
Experimental treatments (Explanations)

- No explanation (control)
- Feature importance
- Feature contribution
- Nearest neighbors
- Counterfactuals



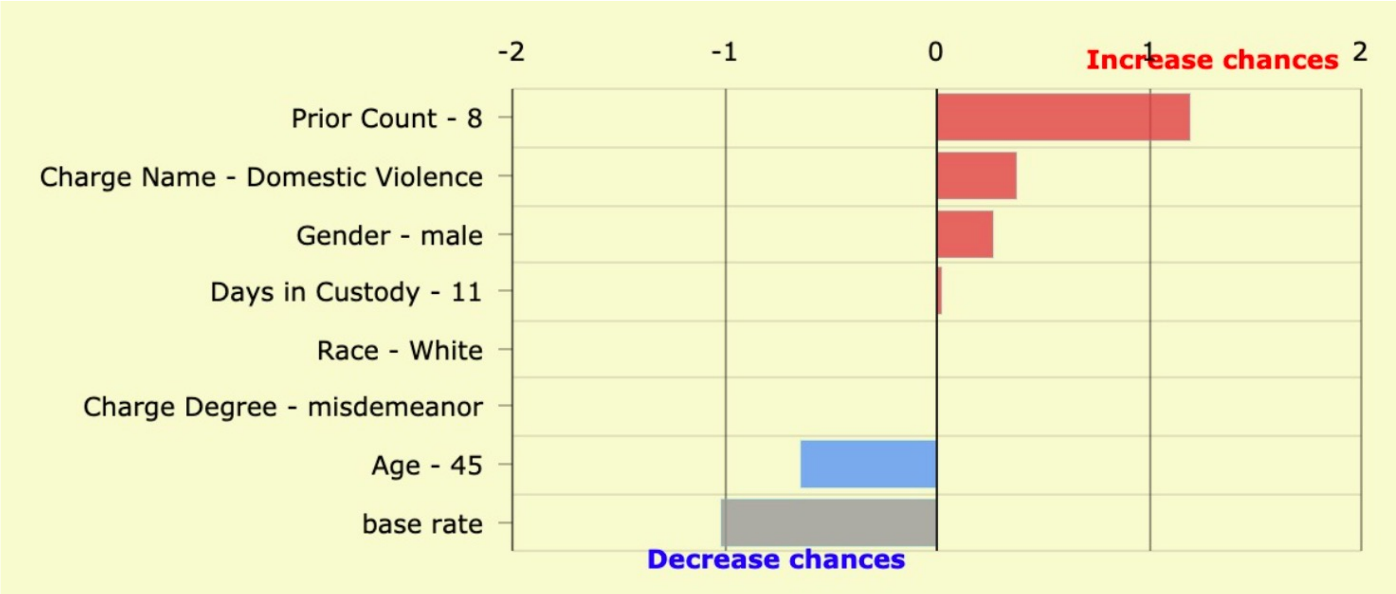
Experimental treatments (Explanations)

- No explanation (control)
- Feature importance
- Feature contribution
- Nearest neighbors
- Counterfactuals



Experimental treatments (Explanations)

- No explanation (control)
- Feature importance
- Feature contribution
- Nearest neighbors
- Counterfactuals



Experimental treatments (Explanations)

- No explanation (control)
- Feature importance
- Feature contribution
- Nearest neighbors
- Counterfactuals

	Current defendant	Defendant A (same)	Defendant B (different)
Machine Learning Prediction	will reoffend	will reoffend	will not reoffend
1. Race:	White	Black	White
2. Gender:	male	male	male
3. Age:	26	26	26
4. Prior Count:	2	2	2
5. Charge Name:	Grand Theft	Grand Theft	arrest case no charge
6. Charge Degree:	felony	felony	felony
7. Days in Custody:	1	1	1

Experimental treatments (Explanations)

- No explanation (control)
- Feature importance
- Feature contribution
- Nearest neighbors
- Counterfactuals

For this defendant, our model would have made the opposite prediction (i.e., predict this defendant "will not reoffend") in the each of following cases:

- **Race**: If the defendant's Race had been **Hispanic** instead of White
- **Gender**: If the defendant's Gender had been **female** instead of male
- **Age**: If the defendant's Age had been **29** instead of 26
- **Prior Count**: If the defendant's Prior Count had been **1** instead of 2
- **Charge Name**: If the defendant's Charge Name had been **Driving with a Suspended License** instead of Grand Theft
- **Charge Degree**: If the defendant's Charge Degree had been **misdemeanor** instead of felony

In contrast, changing the value for each of the following features while keeping other features unchanged would not make our model predict differently:

- Days in Custody

Experimental Procedure

Entry Survey

task familiarity, technical literacy, algorithm literacy, demographic information

Tutorial

32 Tasks - low/high confidence

- ✓ human initial prediction
- ✓ ML prediction w/ or w/o explanation
- ✓ human final prediction

Exit Survey

objective understanding questions, subjective understanding, open-ended feedback

Experimental Procedure

Entry Survey

task familiarity, technical literacy, algorithm literacy, demographic information

Tutorial

32 Tasks - low/high confidence

- ✓ human initial prediction
- ✓ ML prediction w/ or w/o explanation
- ✓ human final prediction

Exit Survey

objective understanding questions, subjective understanding, open-ended feedback

Experimental Procedure

Entry Survey

task familiarity, technical literacy, algorithm literacy, demographic information

Tutorial

32 Tasks - low/high confidence

- ✓ human initial prediction
- ✓ ML prediction w/ or w/o explanation
- ✓ human final prediction

Exit Survey

objective understanding questions, subjective understanding, open-ended feedback

Experimental Procedure

Entry Survey

task familiarity, technical literacy, algorithm literacy, demographic information

Tutorial

32 Tasks - low/high confidence

- ✓ human initial prediction
- ✓ ML prediction w/ or w/o explanation
- ✓ human final prediction

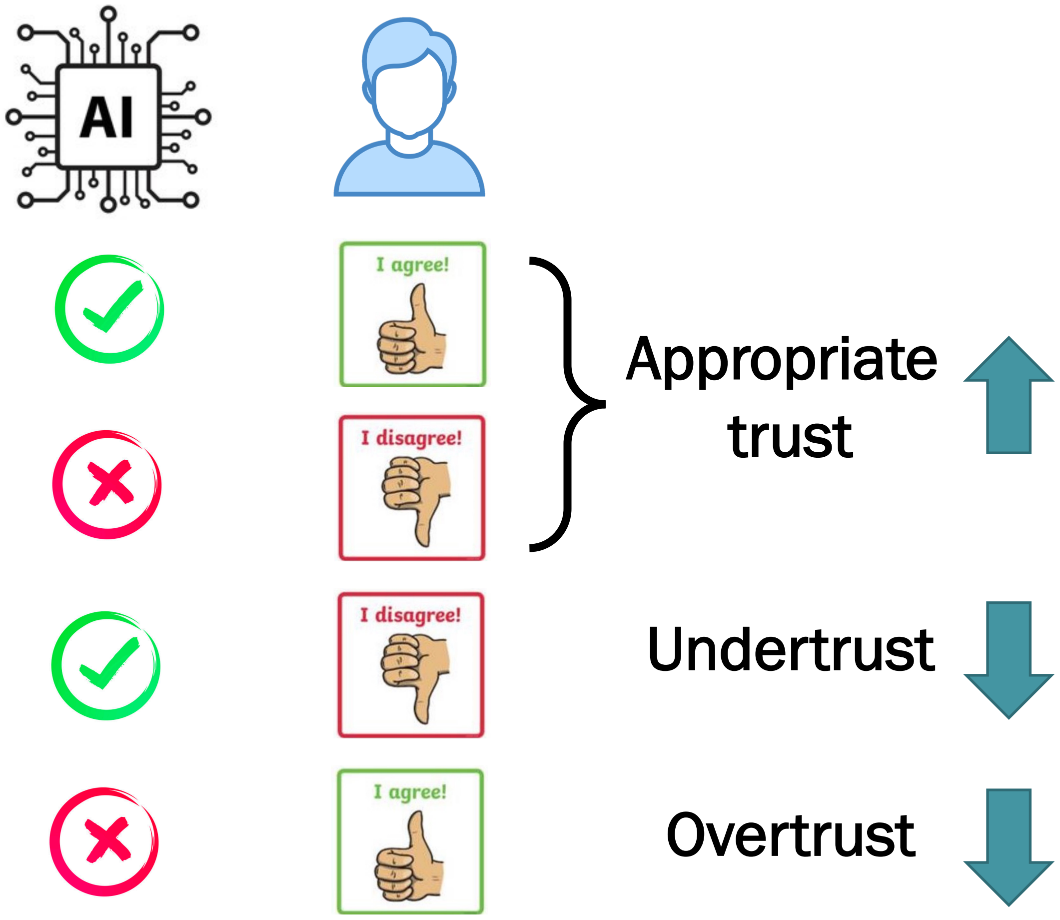
Exit Survey

objective understanding questions, subjective understanding, open-ended feedback

Recidivism prediction: 782 participants

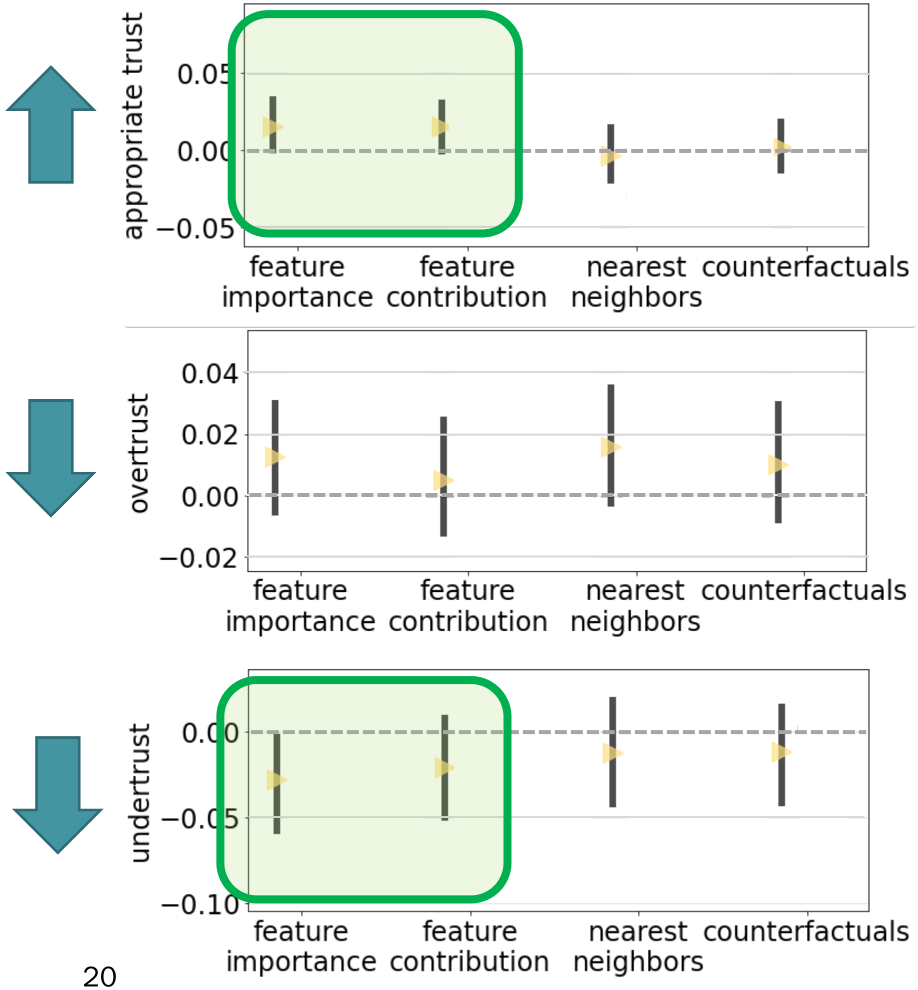
Forest cover prediction: 561 participants

Results: Trust Calibration



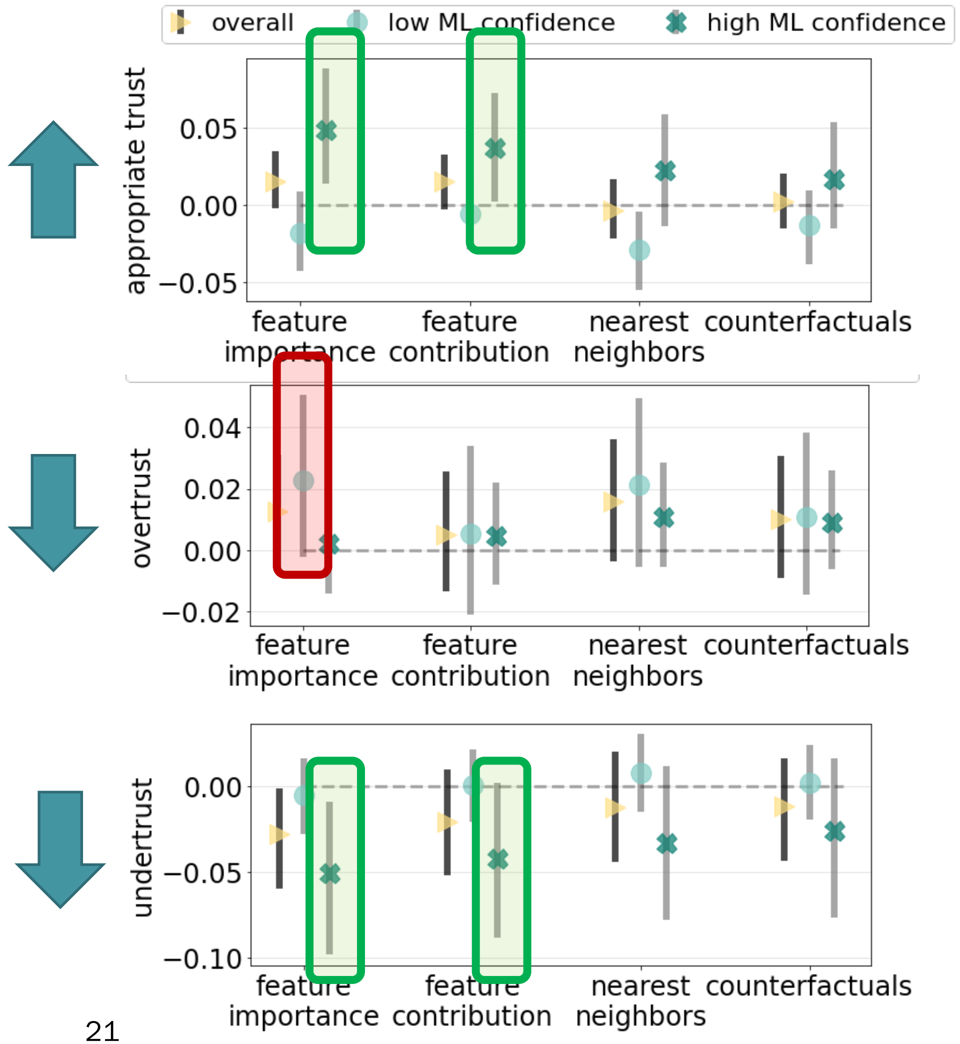
Results: Trust Calibration

Recidivism prediction

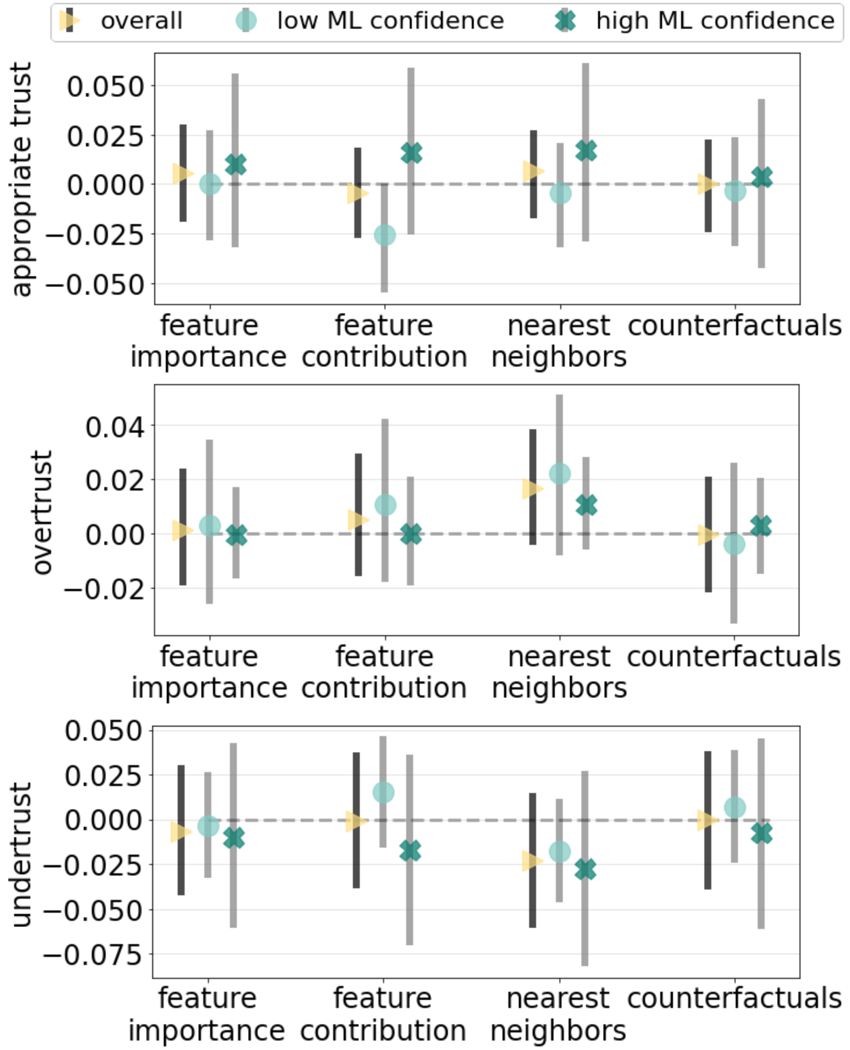


Results: Trust Calibration

Recidivism prediction



Forest cover prediction



Summary

Explanation type	Recidivism prediction: Understanding	Recidivism prediction: Uncertainty Awareness	Recidivism prediction: Trust Calibration	Forest cover prediction: Understanding	Forest cover prediction: Uncertainty Awareness	Forest cover prediction: Trust Calibration
feature importance	✓	✓	✗	✓?	✗	✗
feature contribution	✓?	✓	✓	✓?	✗	✗
nearest neighbor	✓?	✓	✗	✗	✗	✗
counterfactuals	✓	✓	✗	✗	✗	✗

Note: ✓ (or ✗) means our study finds (or does not find) supportive evidence suggesting the explanation method satisfies a desideratum. In the ✓? cases, we only find partial evidence supporting the explanation increases people's understanding of the model (either measured by objective understanding or subjective understanding, but not both).

- The effectiveness of AI explanations are largely different on tasks where people have varying levels of domain expertise in
- Contextual information in empirical results communication
- The right type of explanation for the right purpose?

Thank you!

Are Explanations Helpful?

A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making

Xinru Wang, Ming Yin

Purdue University

Summary

Explanation type	Recidivism prediction			Forest cover prediction		
	Understanding	Uncertainty Awareness	Trust Calibration	Understanding	Uncertainty Awareness	Trust Calibration
feature importance	✓	✓	✗	✓?	✗	✗
feature contribution	✓?	✓	✓	✓?	✗	✗
nearest neighbor	✓?	✓	✗	✗	✗	✗
counterfactuals	✓	✓	✗	✗	✗	✗

Note: ✓ (or ✗) means our study finds (or does not find) supportive evidence suggesting the explanation method satisfies a desideratum. In the ✓? cases, we only find partial evidence supporting the explanation increases people's understanding of the model (either measured by objective understanding or subjective understanding, but not both).

- The effectiveness of AI explanations are largely different on tasks where people have varying levels of domain expertise in.
- Transparency in empirical results communication
- The right type of explanation for the right purpose?