Watch Out For Updates:

Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making

Xinru Wang, Ming Yin
Purdue University

PURDUE UNIVERSITY

AI supports human decision making.

investment choice

medical diagnostic

toxic content detection

AI supports human decision making.

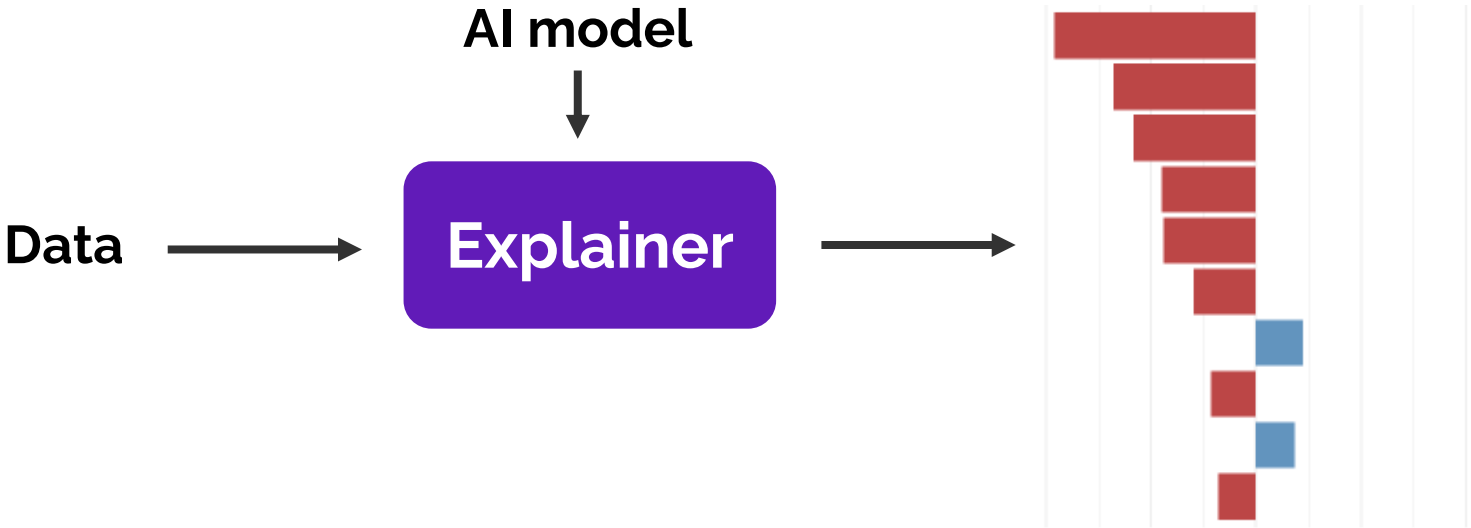
investment choice

toxic content detection

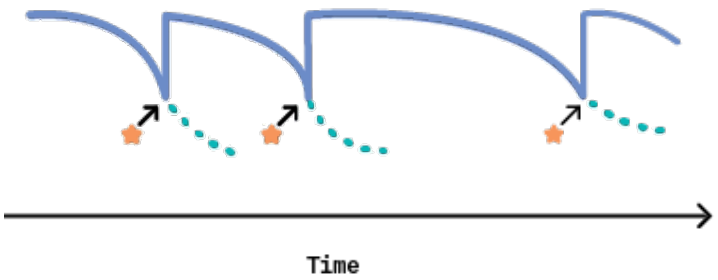
medical diagnostic

Explanations!

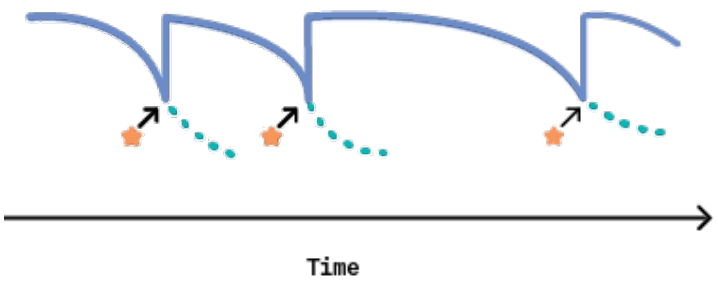
E.g., feature importance explanations

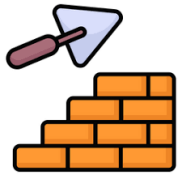


AI gets updated over time.



AI gets updated over time.



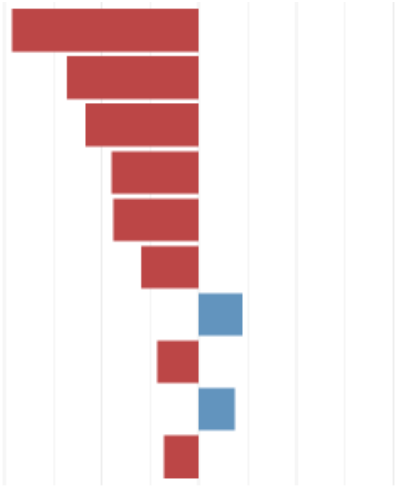


new training data

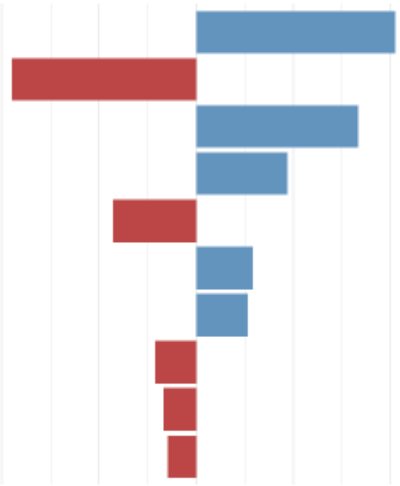
new AI algorithms

new regulations

AI updates can result in changes in its explanations.

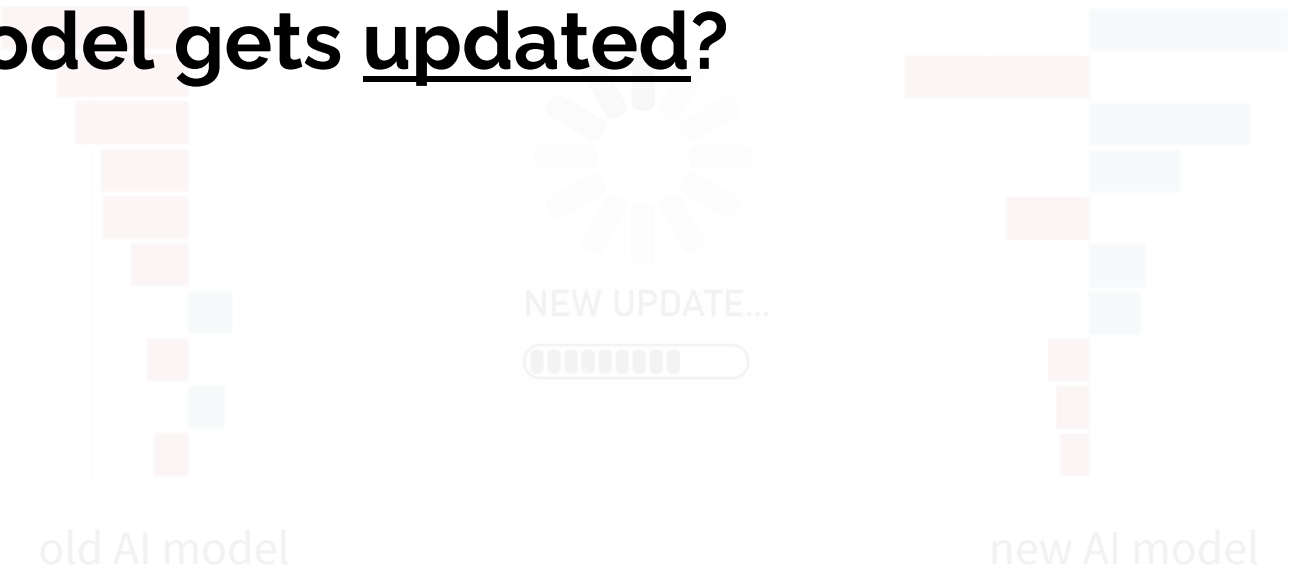


old AI model



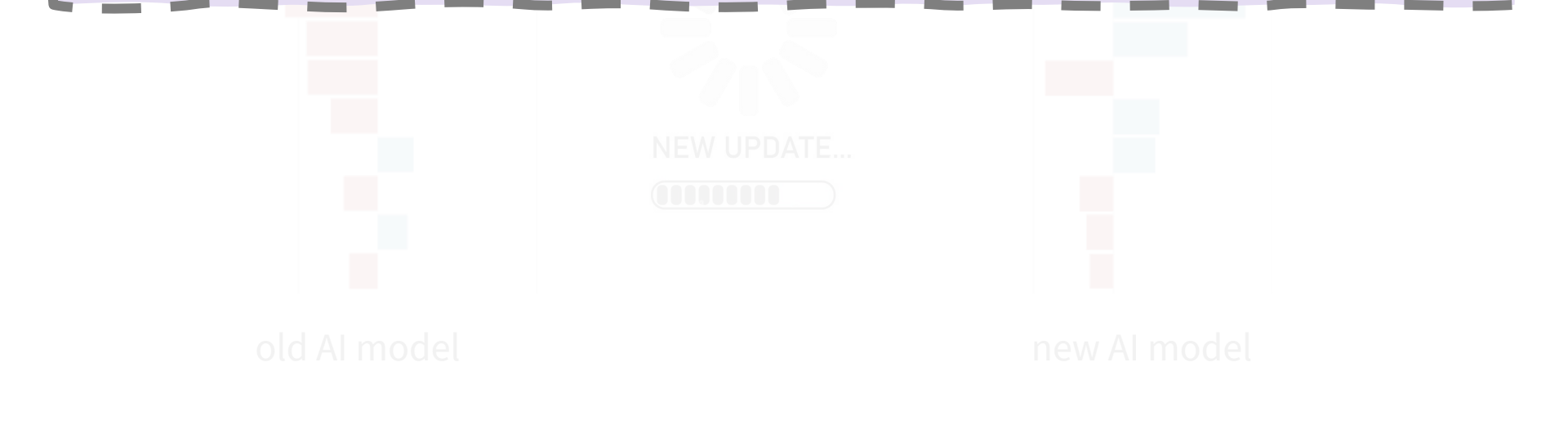
new AI model

How do end-users of the AI-driven decision aid react to changes in AI explanations, as the AI model gets updated?



Specifically:

1. Can end-users perceive the changes ?



Specifically:

1. Can end-users perceive the changes ?

2. Will the changes affect end-users' trust and satisfaction with the AI?

Specifically:

1. Can end-users perceive the changes ?

2. Will the changes affect end-users' trust and satisfaction with the AI?

3. What are the underlying mechanisms of these effects?

Task (1/30)

Please review the profile below and predict whether this applicant will default on the loan. *If you don't remember the meaning of a feature, click on the red circle on that feature to view its meaning.*

Profile of this loan application:

Basic Information about the Loan	
1. Loan Amount:	< $5,000
2. Issued Month:	Sep
Basic Information about the Applicant	
3. Annual Income:	$80,000 - $100,000
4. State of Address:	Texas
5. Credit Score:	Good
6. Month of Earliest Credit Account:	Sep

Make Your Prediction:

Do you think this applicant will default on the loan?

- Yes, I think this applicant **will** default on the loan.
- No, I think this applicant **will not** default on the loan.

Machine Learning Prediction:

Our machine learning model predicts that this applicant **will not** default on the loan.
The two features that contributes the most to the model's prediction is **Credit Score (Good) and Loan Amount (< $5,000)**.

Make your final prediction:

Now, do you think this applicant will default on the loan?

- Yes, I think this applicant **will** default on the loan.
- No, I think this applicant **will not** default on the loan.

Next

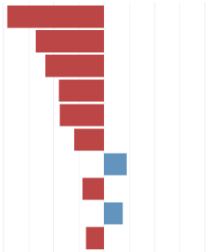
Experimental Procedure

Phase 1: Before AI Update

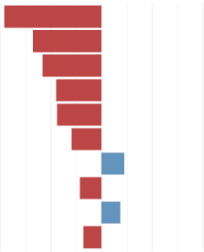
Phase 2: After AI Update

Experimental Treatments:

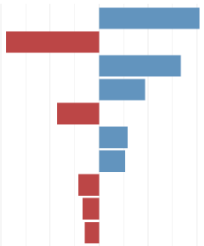
- High Similarity Update
- Medium Similarity Update
- Low Similarity Update



high-similarity update

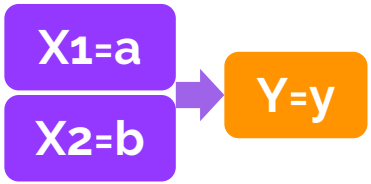


low-similarity update



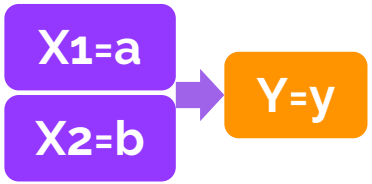
Experimental Treatments

Explanation Before AI Update:



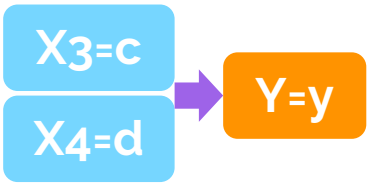
Experimental Treatments

Explanation Before AI Update:



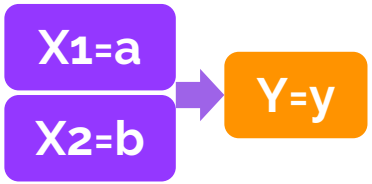
Explanation After AI Update:

low-similarity



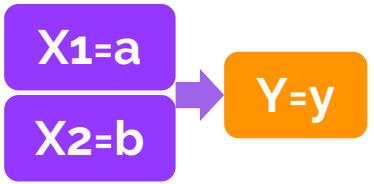
Experimental Treatments

Explanation Before AI Update:

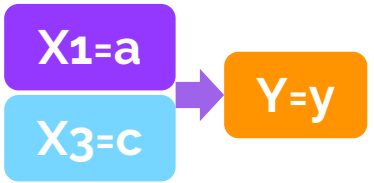


Explanation After AI Update:

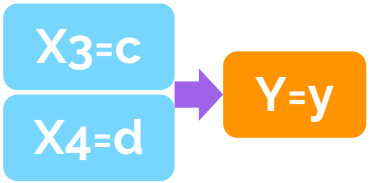
high-similarity



medium-similarity



low-similarity



Experimental Design

Experiment 2.2: Phase 1 relevant

We obtained users' general common knowledge through a separate pilot study.

Basic Information about the Loan	
1. Loan Amount:	< $5,000
2. Issued Month:	Sep
Basic Information about the Applicant	
3. Annual Income:	$80,000 - $100,000
4. State of Address:	Texas
5. Credit Score:	Good
6. Month of Earliest Credit Account:	Sep

low-similarity update

Basic Information about the Loan	
1. Loan Amount:	< $5,000
2. Issued Month:	Mar
Basic Information about the Applicant	
3. Annual Income:	$40,000 - $60,000
4. State of Address:	California
5. Credit Score:	Good
6. Month of Earliest Credit Account:	Aug

Measures & Results

1. Can end-users perceive the changes ?

Perceived Explanation Change: self-report after Phase 2

- Yes.

2. Will the changes affect end-users' trust and satisfaction with the AI?

Objective Trust Gain: % human final prediction = AI prediction → Phase 2 – Phase 1

Subjective Trust Gain: self-report → Phase 2 – Phase 1

Subjective Satisfaction Gain : self-report → Phase 2 – Phase 1

2. Will the changes affect end-users' trust and satisfaction with the AI?

Objective Trust Gain

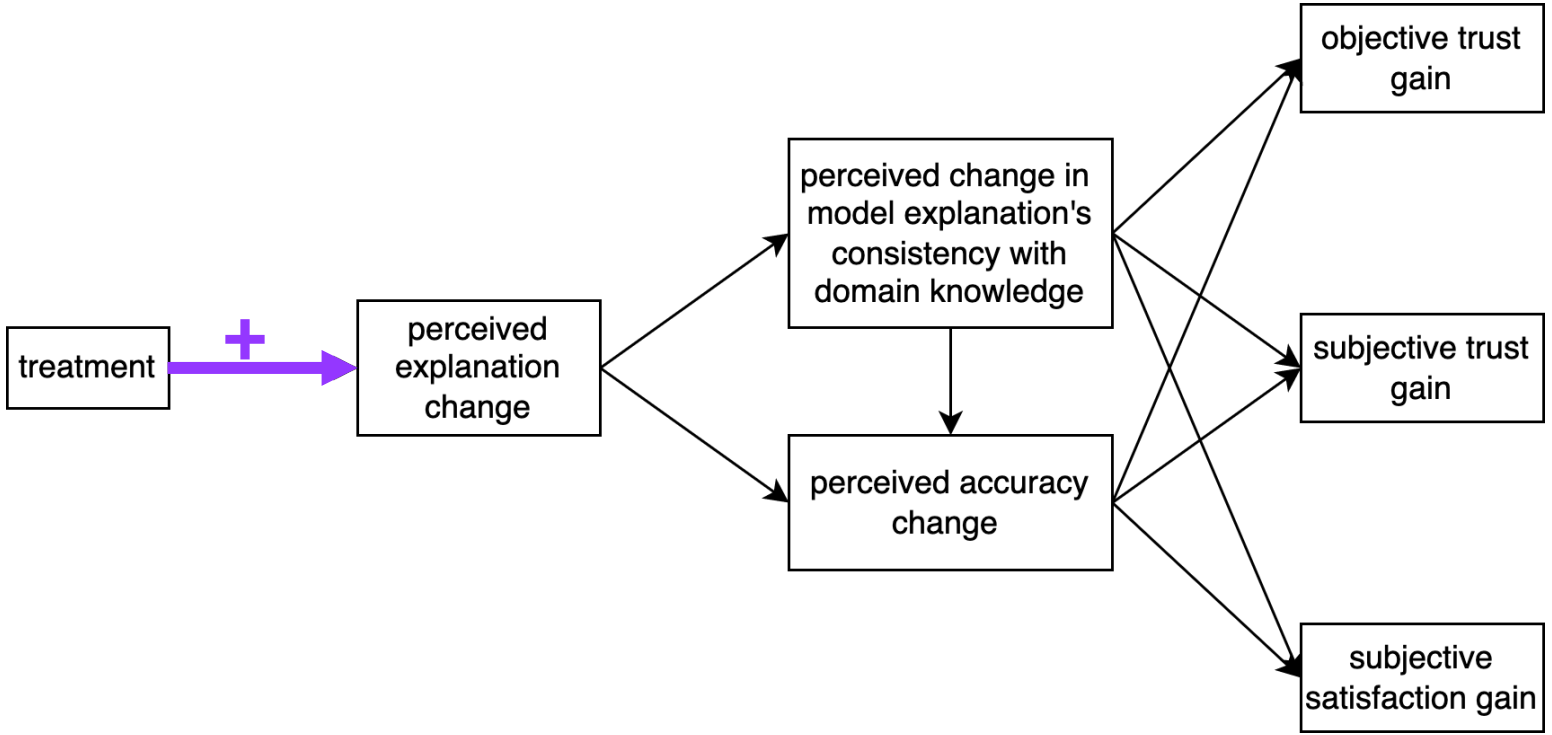
- No

Subjective Trust Gain

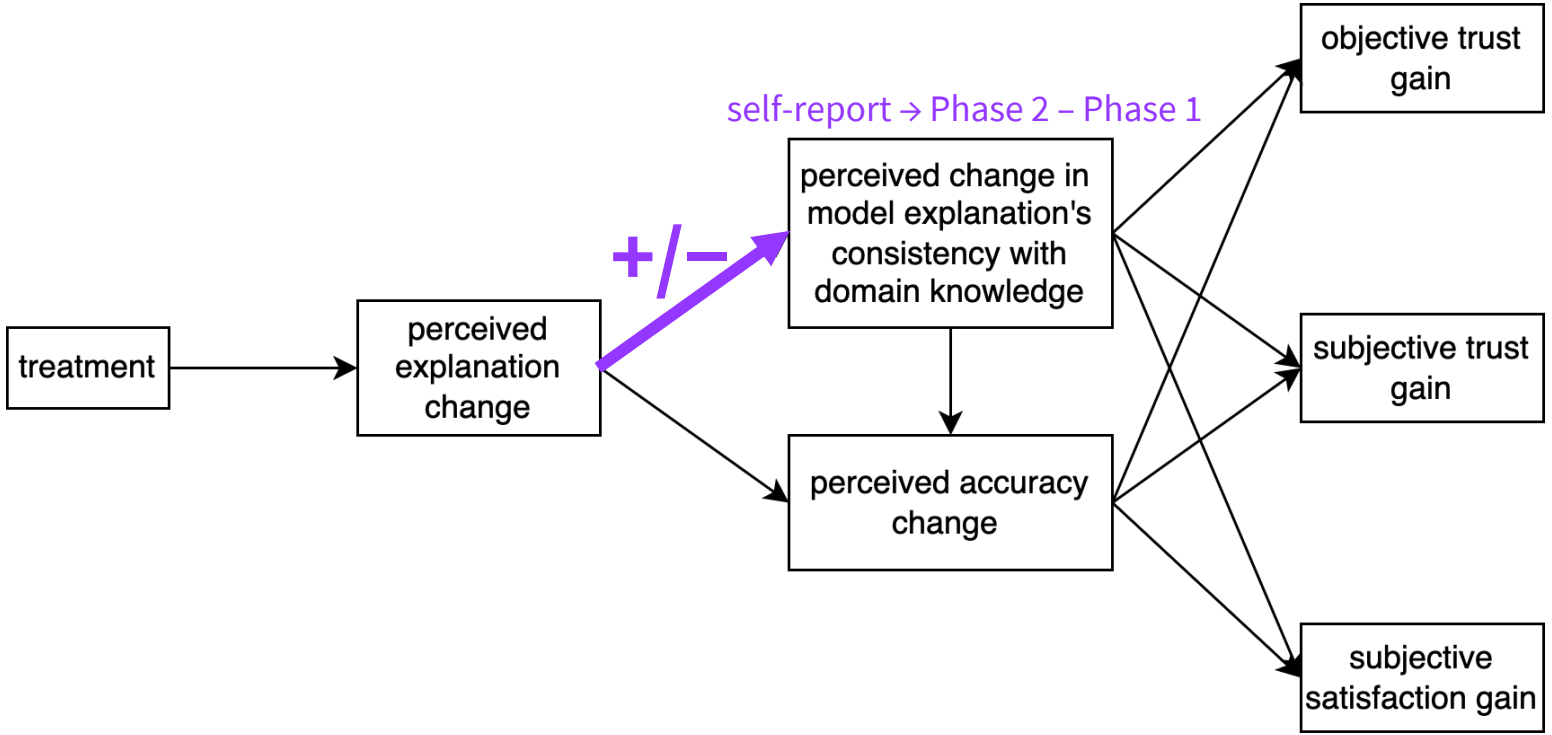
Subjective Satisfaction Gain

- Yes, when users have some prior knowledge
  - **increase / decrease** when the new AI explanation is **consistent / inconsistent** with the human rationale.

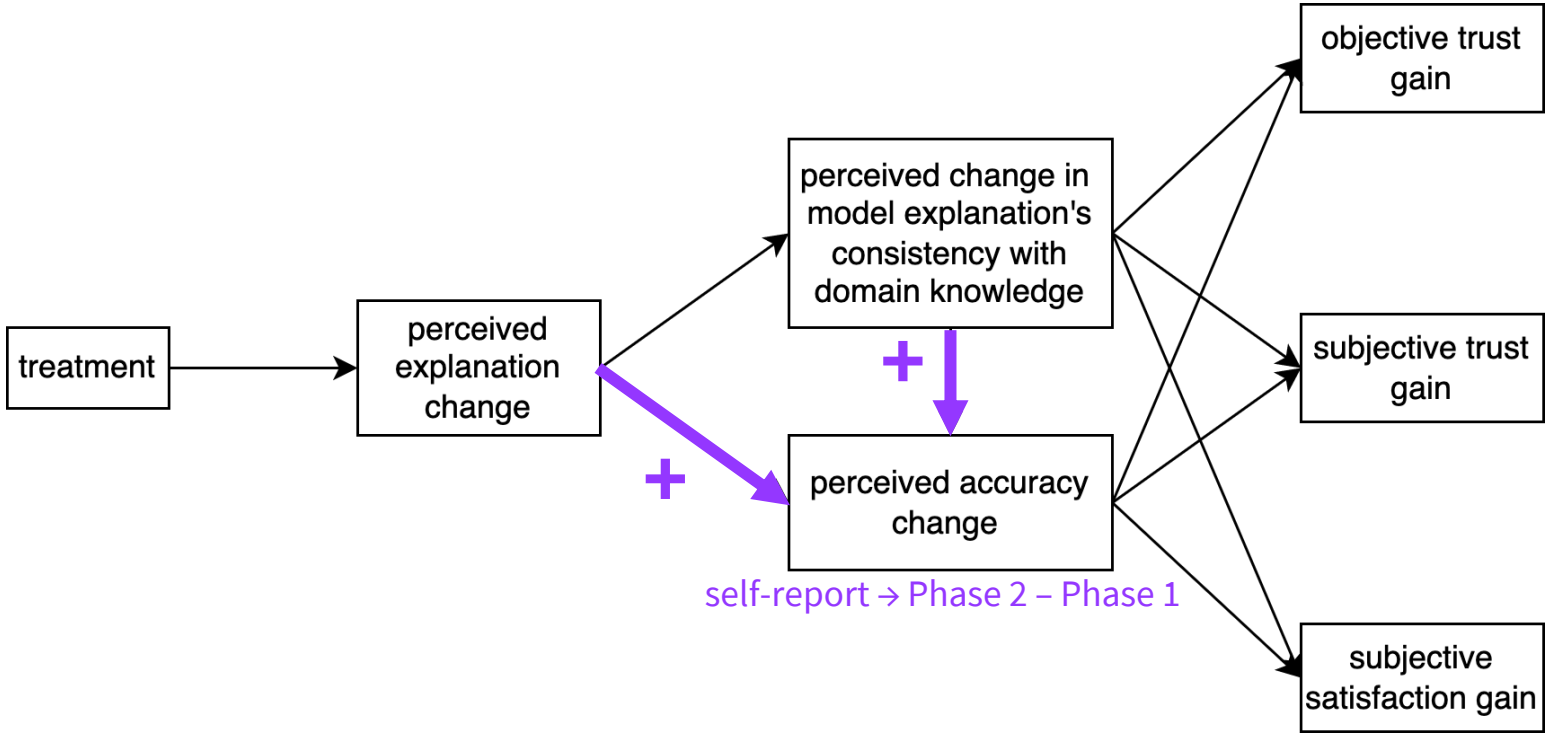
3. What are the underlying mechanisms of these effects?



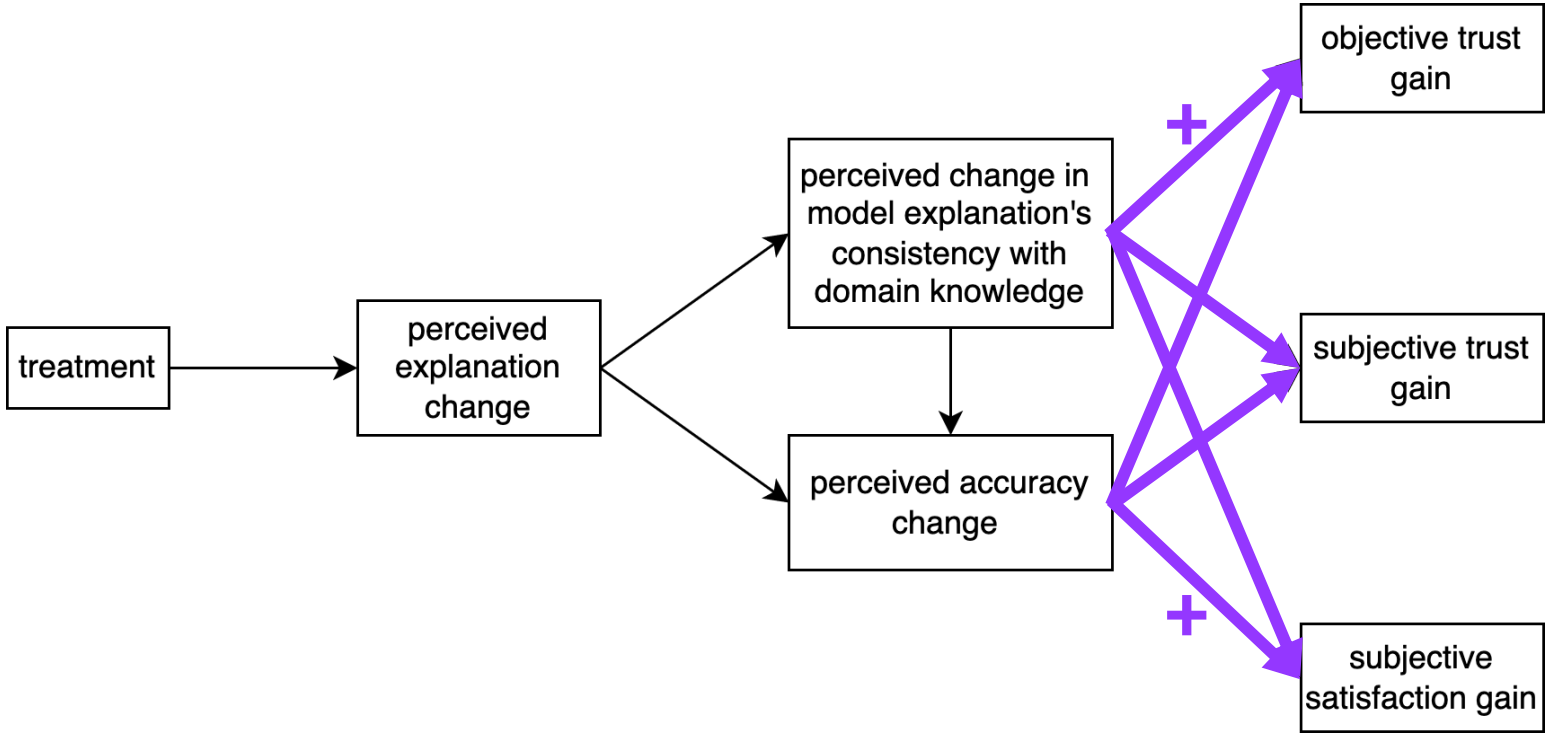
3. What are the underlying mechanisms of these effects?



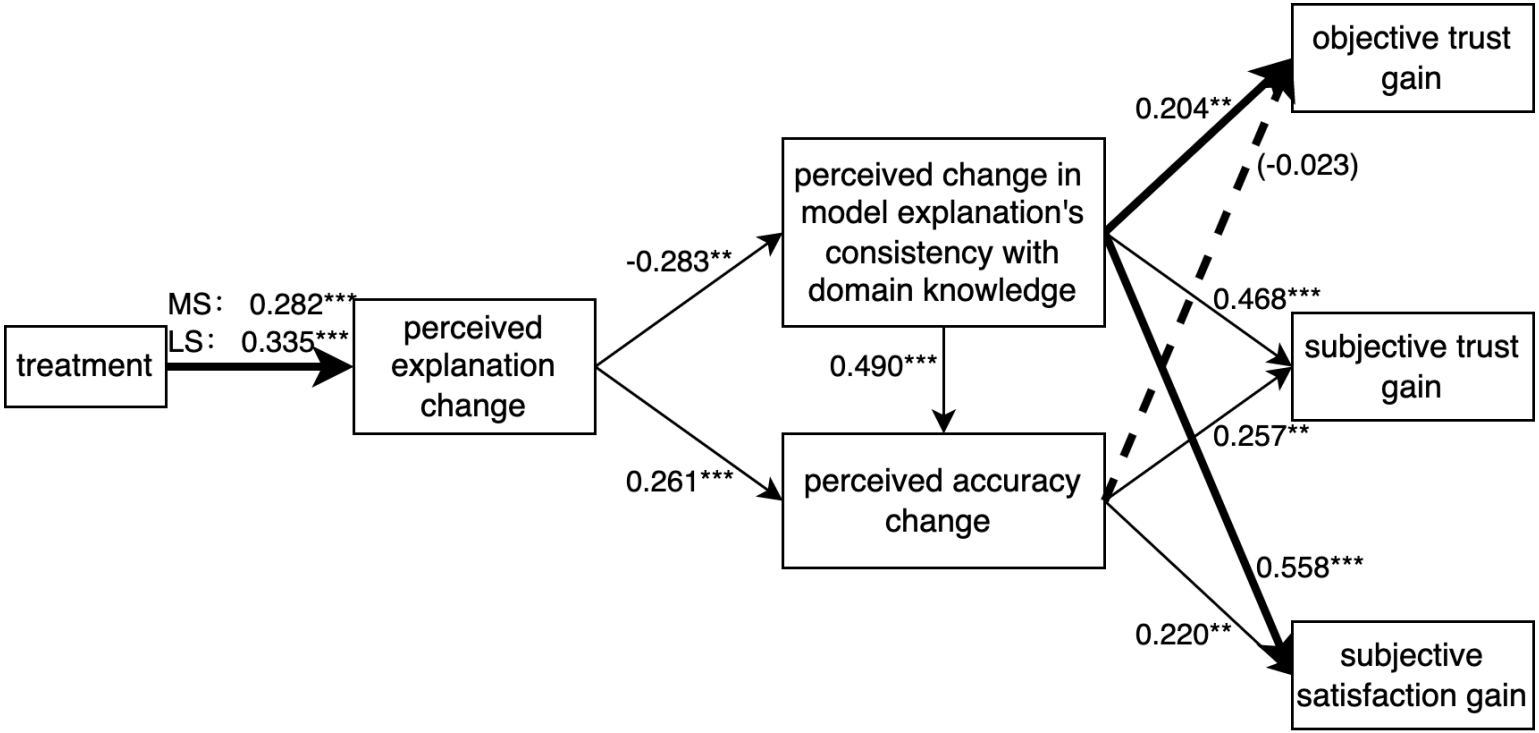
3. What are the underlying mechanisms of these effects?



3. What are the underlying mechanisms of these effects?



3. What are the underlying mechanisms of these effects?



Take Away

As the AI model gets updated,

- End-users can perceive changes in AI explanations

- Changes in AI explanation may change users' subjective perception in the AI model

- Design implications for XAI methods in a fast-evolving AI lifecycle
  - Integrating human expertise into the AI explanation updating processes
  - Highlighting the changes in the AI explanation

Thank You!