CS 4501/6501 Interpretable Machine Learning

Interpretation and Human Understanding

Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
{hc9mx, yangfeng}@virginia.edu

Properties

- Faithfulness to model

  How accurately an interpretation reflects the true reasoning process of the model

- Plausibility to humans

  How convincing the interpretation is to humans

[Jacovi and Yoav, 2020]

Properties

- Faithfulness to model

  How accurately an interpretation reflects the true reasoning process of the model

- Plausibility to humans

  How convincing the interpretation is to humans

[Jacovi and Yoav, 2020]

- Generally, it is not easy to satisfy both criteria because of the gap between model reasoning and human understanding
- Faithfulness is the primary criterion

Simulatability

A model is simulatable when a person can predict its behavior on new inputs

[Doshi-Velez and Kim, 2017]

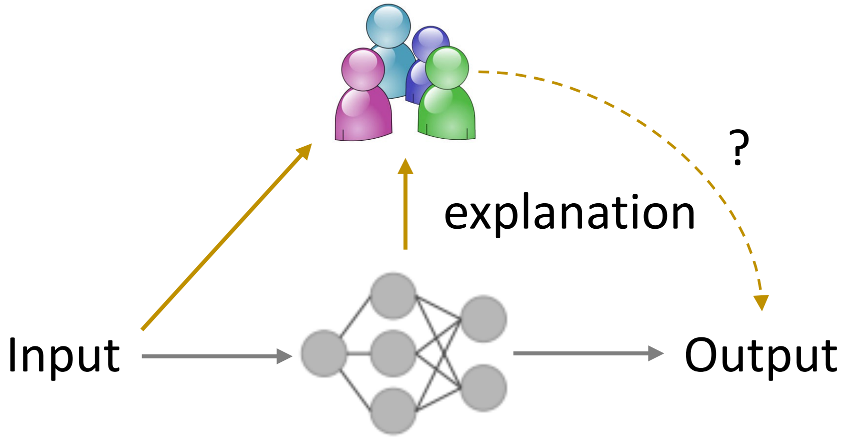
Simulatability

A model is simulatable when a person can predict its behavior on new inputs

[Doshi-Velez and Kim, 2017]

Human-subject tasks

- Forward simulation: given an input and an explanation, users must predict what a model would output for the given input



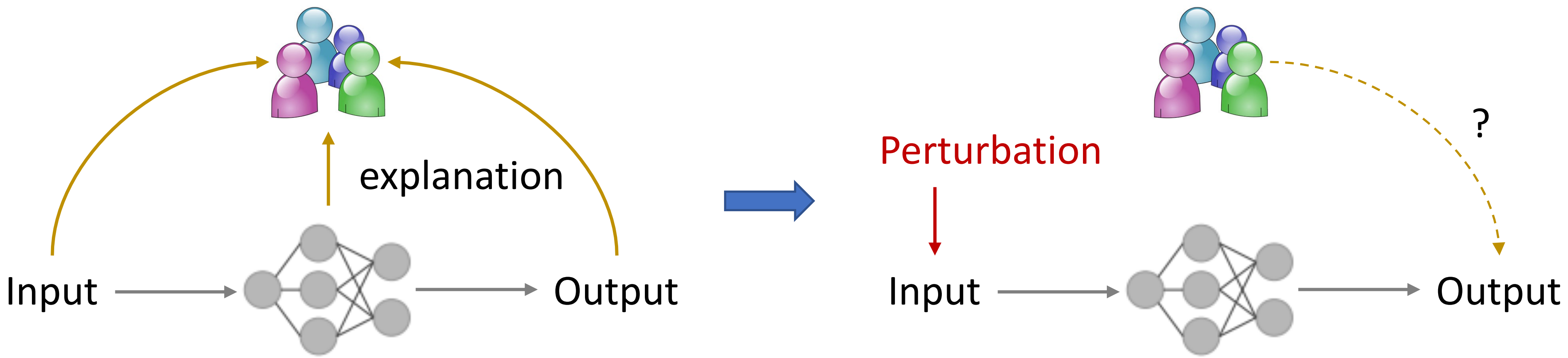
Simulatability

A model is simulatable when a person can predict its behavior on new inputs

[Doshi-Velez and Kim, 2017]

Human-subject tasks

- Counterfactual simulation: users are given an input, a model's output for that input, and an explanation of that output, and then they must predict what the model will output when given a perturbation of the original input



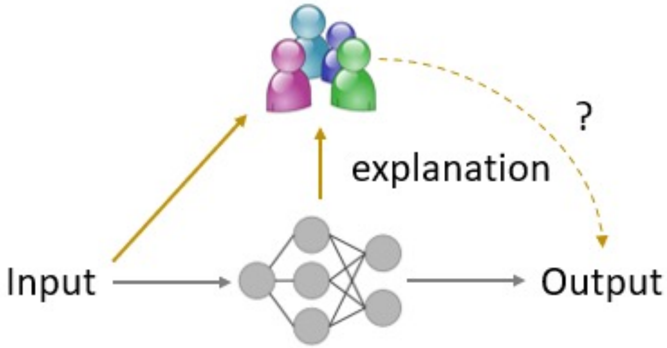
Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Peter Hase and Mohit Bansal

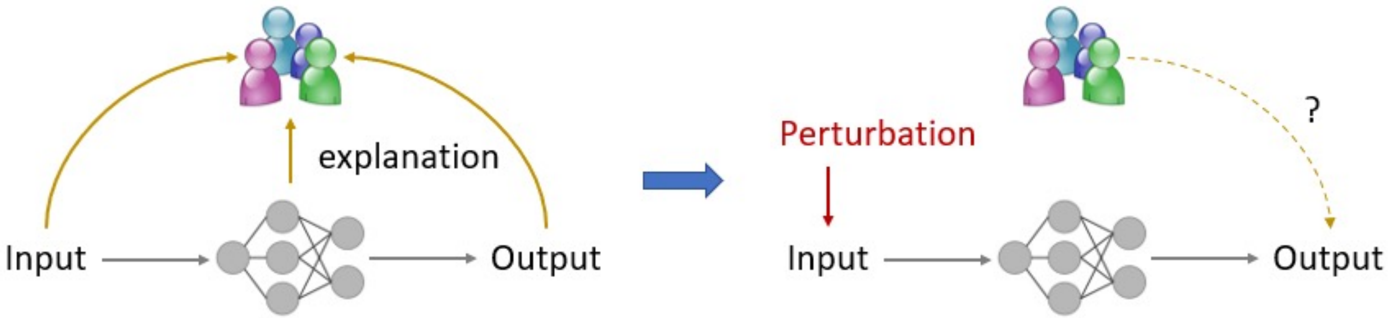
(ACL, 2020)

Problem

- Forward simulation



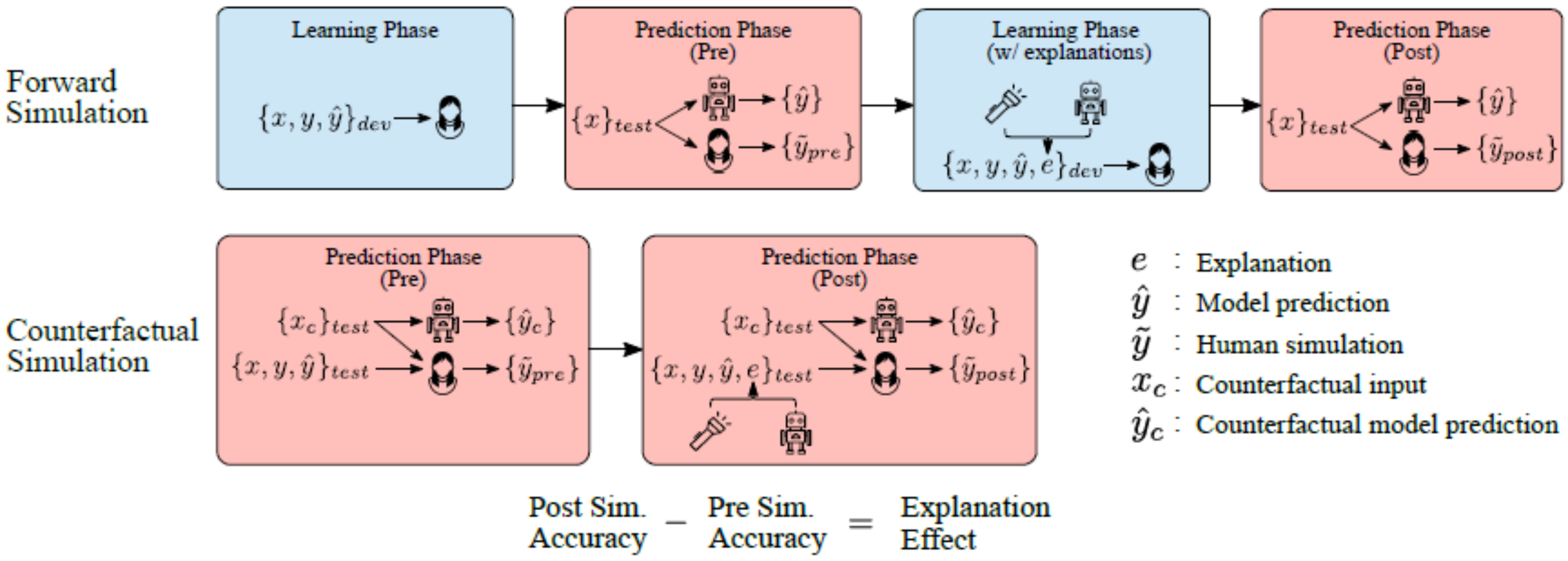
- Counterfactual simulation



- Humans really understand model prediction behavior?
- Explanations give away the answers?

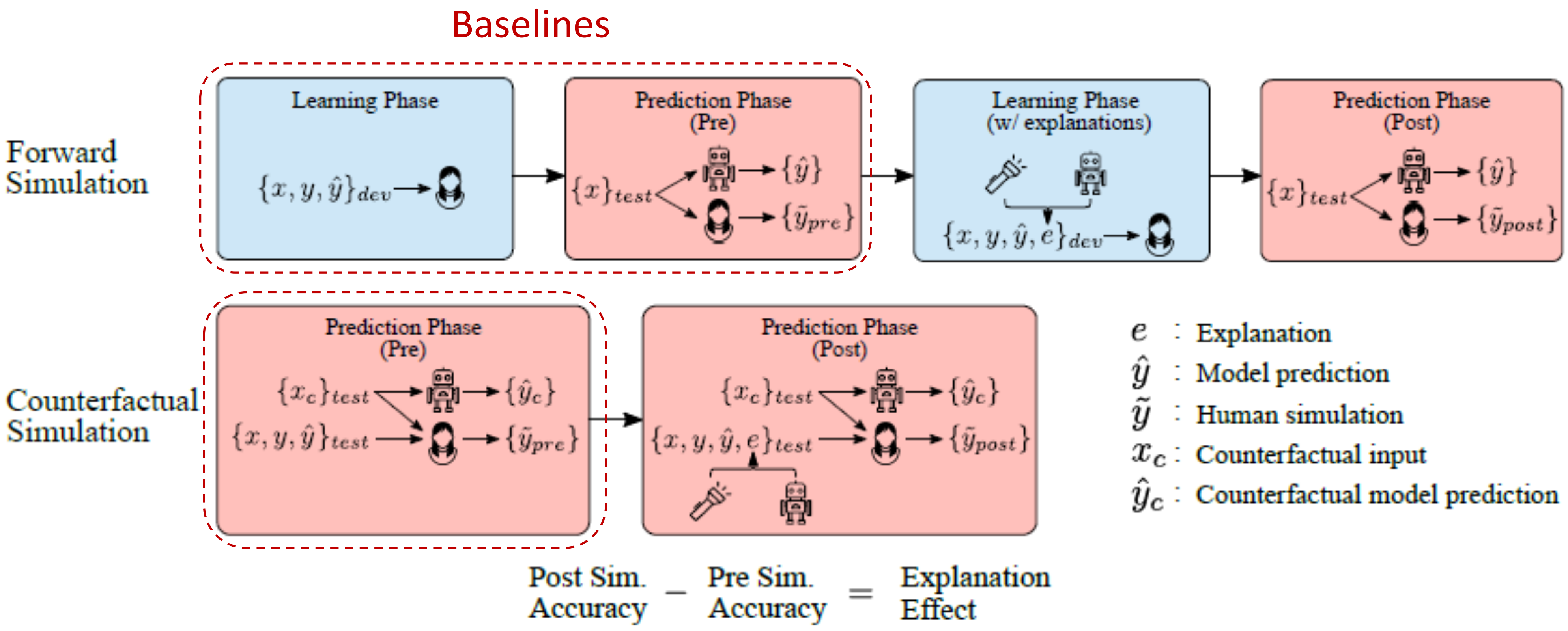
Method

- Separate the explained instances from the test instances to prevent explanations from giving away the answers



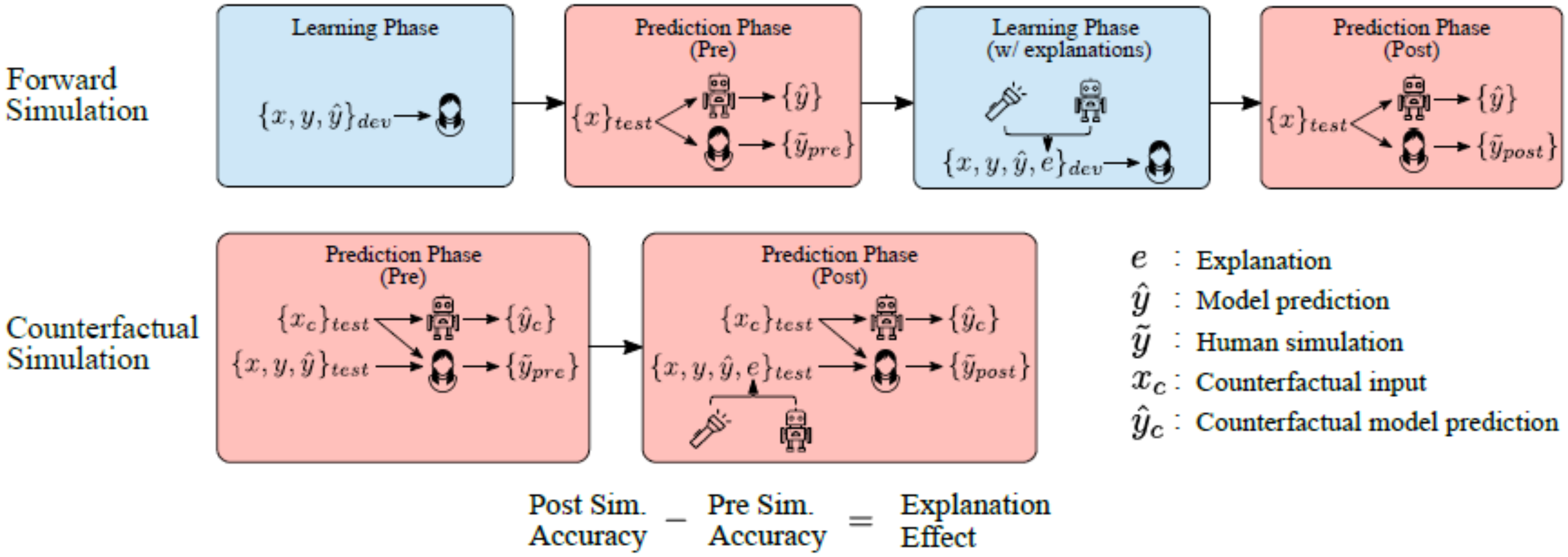
Method

- Separate the explained instances from the test instances to prevent explanations from giving away the answers
- Evaluate the effect of explanations against a baseline where users see the same example data points without explanations



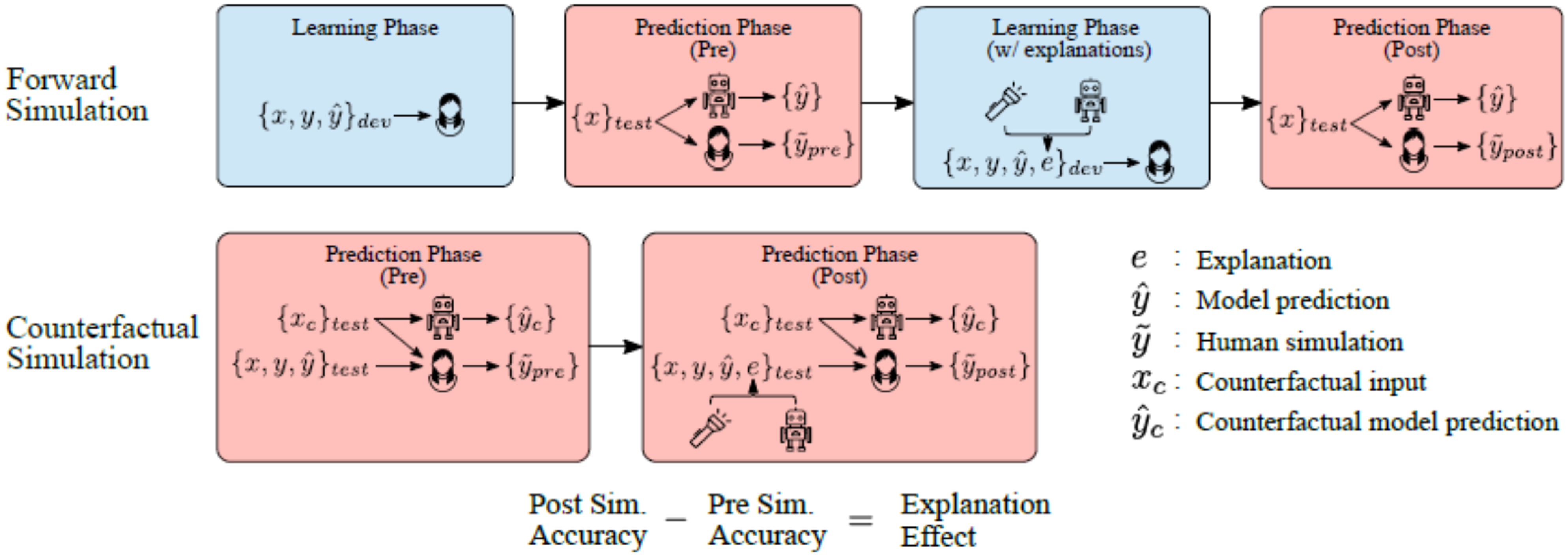
Method

- Separate the explained instances from the test instances to prevent explanations from giving away the answers
- Evaluate the effect of explanations against a baseline where users see the same example data points without explanations
- Balance data by model correctness (users cannot succeed simply by guessing the true label)



Method

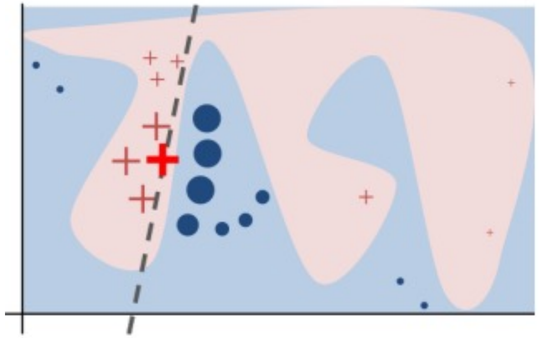
- Separate the explained instances from the test instances to prevent explanations from giving away the answers
- Evaluate the effect of explanations against a baseline where users see the same example data points without explanations
- Balance data by model correctness (users cannot succeed simply by guessing the true label)
- Force user predictions on every input (not favor some explanations)



Question?

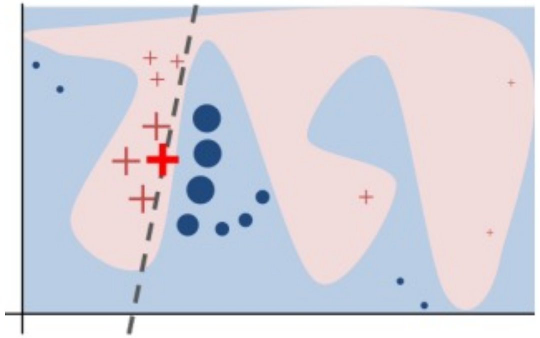
Explanations

LIME [Ribeiro, 2016]



Explanations

LIME [Ribeiro, 2016]



Anchors [Ribeiro, 2018]

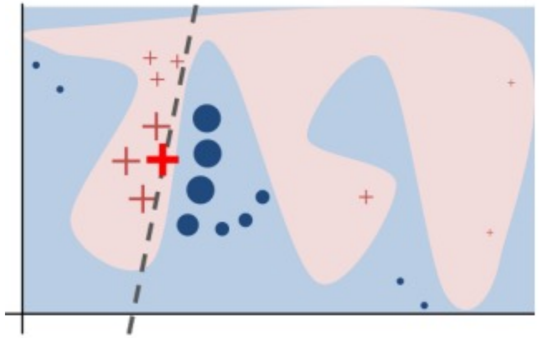
Local rule lists

+ This movie is not bad.

{"not", "bad"} → Positive

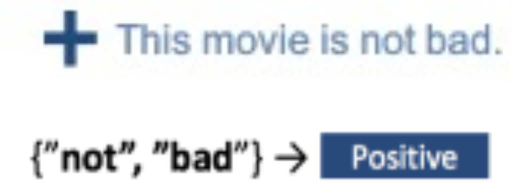
Explanations

LIME [Ribeiro, 2016]



Anchors [Ribeiro, 2018]

Local rule lists



Prototype

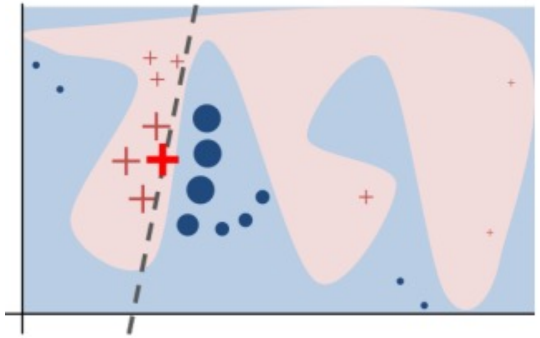
$$f(x)_c = \max_{p_k \in P_c} a(g(x), p_k)$$

$$Attr(x_i) = f(x)_c - f\left(x_{\backslash x_i}\right)_c$$

c: class
P_c: a set of prototype vectors
a: similarity function
g: a neural network

Explanations

LIME [Ribeiro, 2016]



Anchors [Ribeiro, 2018]

Local rule lists

+ This movie is not bad.

{"not", "bad"} → Positive

Prototype

$$f(x)_c = \max_{p_k \in P_c} a(g(x), p_k)$$

$$Attr(x_i) = f(x)_c - f\left(x_{\setminus x_i}\right)_c$$

c: class
P_c: a set of prototype vectors
a: similarity function
g: a neural network

Decision Boundary

Input, Label, and Model Output

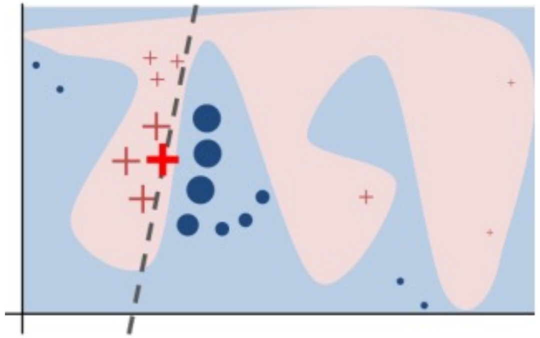
x = Despite modest aspirations its occasional charms are not to be dismissed.
y = Positive $\hat{y}$ = Negative

Decision Boundary

Step 0	Evidence Margin: -5.21
Step 1	occasional → rare Evidence Margin: -3.00
Step 2	modest → impressive Evidence Margin: +0.32
$x^{(c)}$	Despite ***impressive*** aspirations its ***rare*** charms are not to be dismissed.

Explanations

LIME [Ribeiro, 2016]



Anchors [Ribeiro, 2018]

Local rule lists

+ This movie is not bad.

{"not", "bad"} → Positive

Prototype

$$f(x)_c = \max_{p_k \in P_c} a(g(x), p_k)$$

$$Attr(x_i) = f(x)_c - f(x_{\setminus x_i})_c$$

c: class

P_c: a set of prototype vectors

a: similarity function

g: a neural network

Decision Boundary

Input, Label, and Model Output

x = Despite modest aspirations its occasional charms are not to be dismissed.

y = Positive $\hat{y}$ = Negative

Decision Boundary

Step 0	Evidence Margin: -5.21
Step 1	occasional → rare Evidence Margin: -3.00
Step 2	modest → impressive Evidence Margin: +0.32
$x^{(c)}$	Despite ***impressive*** aspirations its ***rare*** charms are not to be dismissed.

Composite Approach

Combine LIME/Anchors/Prototype/Decision Boundary

Question?

Experiments

- Data and task models

  - Movie Reviews [Pang et al., 2002]

    Task: binary sentiment classification

    Model: hierarchical attention network [Yang et al., 2016]

  - Tabular Adult Data [Dua and Graff, 2017]

    Task: predict whether the annual income is more than $50,000

    Model: a neural network with two hidden layers [Ribeiro, 2018]

Experiments

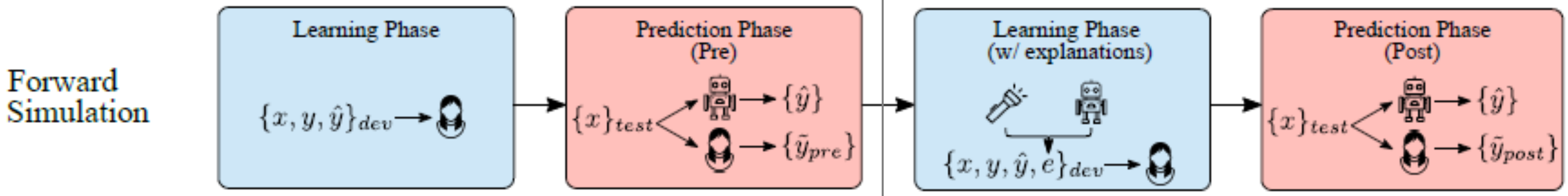
- User pool
  - 32 trained undergraduates who had taken at least one course in computer science or statistics
  - gather over 2100 responses via in-person tests
  - screen out invalid responses (low scores in screening test, task completion time is extremely low)

Experiments

- Forward simulation

First round (baseline)

Second round



16 examples: labels, model predictions, no explanations

Predict the model output for 16/32 new examples

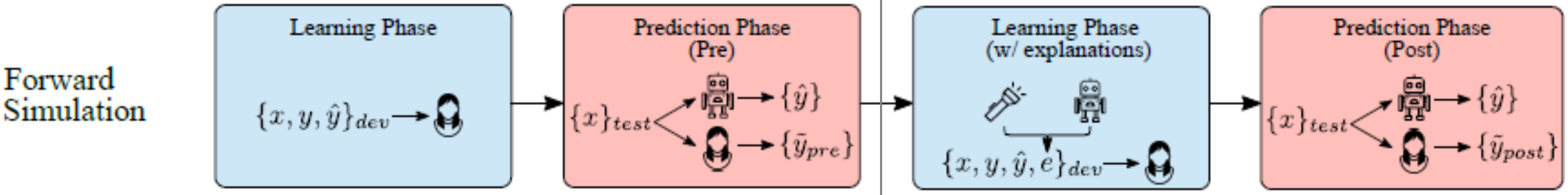
The same examples: labels, model predictions, with explanations

Predict the model behavior again

Experiments

- Forward simulation

First round (baseline) | Second round



16 examples: labels, model predictions, no explanations

Predict the model output for 16/32 new examples

The same examples: labels, model predictions, with explanations

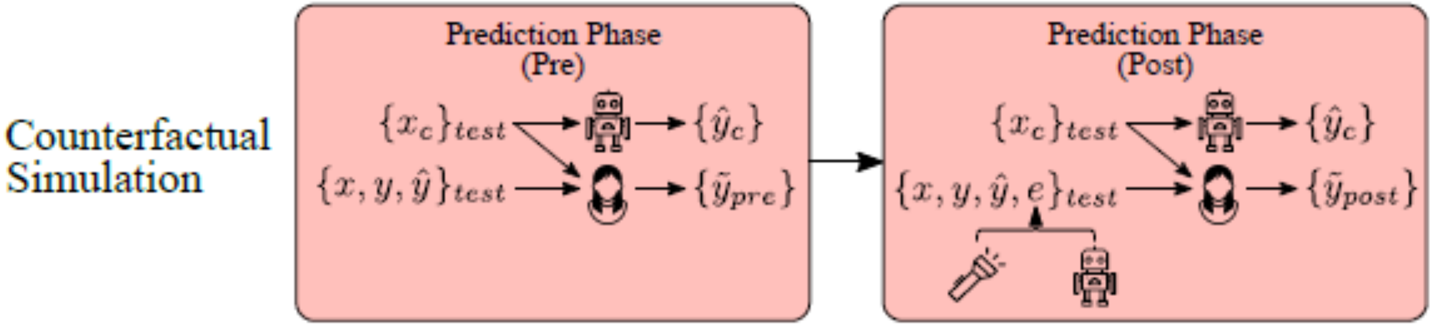
Predict the model behavior again

Post Sim. Accuracy − Pre Sim. Accuracy = Explanation Effect

Experiments

- Counterfactual simulation

Ask users to predict how a model will behave on a perturbation of a given data point



Examples: labels,
model predictions,
no explanations,
perturbations

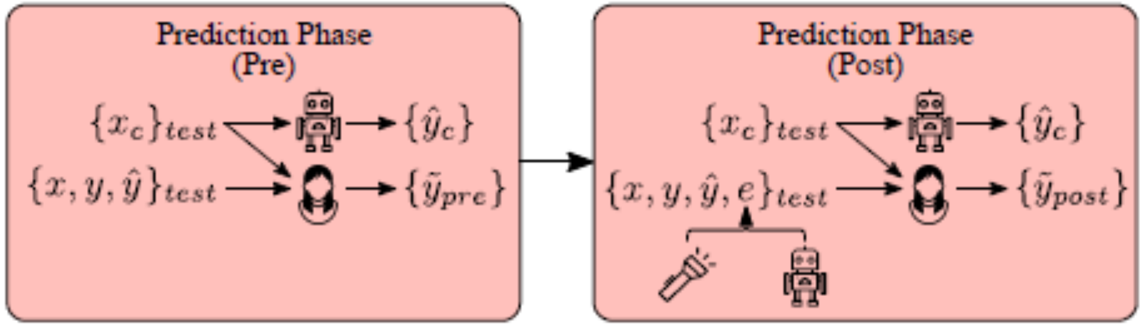
(e.g., randomly substitute words with their neighbors)

Experiments

- Counterfactual simulation

  Ask users to predict how a model will behave on a perturbation of a given data point

Counterfactual Simulation



Examples: labels, model predictions, no explanations, perturbations

(e.g., randomly substitute words with their neighbors)

The same examples with explanations

Post Sim. Accuracy − Pre Sim. Accuracy = Explanation Effect

Experiments

- Data Balancing

  - Goal : prevent users from succeeding on the tests simply by guessing the true label
  - True positives, false positives, true negatives, and false negatives are equally represented
  - For the counterfactual test, there is a 50% chance that the perturbation receives the same prediction as the original input

Results

Do explanations help users?

Explanation effectiveness: the difference in user accuracy across prediction phases in simulation tests

confidence interval

	Text					Tabular				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1144	62.67	-	7.07	-	1022	70.74	-	6.96	-
LIME	190	-	0.99	9.58	.834	179	-	**11.25**	8.83	.014
Anchor	181	-	1.71	9.43	.704	215	-	5.01	8.58	.234
Prototype	223	-	3.68	9.67	.421	192	-	1.68	10.07	.711
DB	230	-	−1.93	13.25	.756	182	-	5.27	10.08	.271
Composite	320	-	3.80	11.09	.486	254	-	0.33	10.30	.952

Results

Do explanations help users?

- LIME improves simulatability with tabular data, while other methods do not definitively improve simulatability in either domain

	Text					Tabular				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1144	62.67	-	7.07	-	1022	70.74	-	6.96	-
LIME	190	-	0.99	9.58	.834	179	-	**11.25**	8.83	.014
Anchor	181	-	1.71	9.43	.704	215	-	5.01	8.58	.234
Prototype	223	-	3.68	9.67	.421	192	-	1.68	10.07	.711
DB	230	-	−1.93	13.25	.756	182	-	5.27	10.08	.271
Composite	320	-	3.80	11.09	.486	254	-	0.33	10.30	.952

Results

Do explanations help users?

- LIME improves simulatability with tabular data, while other methods do not definitively improve simulatability in either domain
- Even with combined explanations in the Composite method, no definitive effects on model simulatability

	Text					Tabular				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1144	62.67	-	7.07	-	1022	70.74	-	6.96	-
LIME	190	-	0.99	9.58	.834	179	-	**11.25**	8.83	.014
Anchor	181	-	1.71	9.43	.704	215	-	5.01	8.58	.234
Prototype	223	-	3.68	9.67	.421	192	-	1.68	10.07	.711
DB	230	-	−1.93	13.25	.756	182	-	5.27	10.08	.271
Composite	320	-	3.80	11.09	.486	254	-	0.33	10.30	.952

Results

Do explanations help users?

- LIME improves simulatability with tabular data, while other methods do not definitively improve simulatability in either domain
- Even with combined explanations in the Composite method, no definitive effects on model simulatability

	Text					Tabular				
Method	n	Pre	Change	CI	p	n	Pre	Change	CI	p
User Avg.	1144	62.67	-	7.07	-	1022	70.74	-	6.96	-
LIME	190	-	0.99	9.58	.834	179	-	**11.25**	8.83	.014
Anchor	181	-	1.71	9.43	.704	215	-	5.01	8.58	.234
Prototype	223	-	3.68	9.67	.421	192	-	1.68	10.07	.711
DB	230	-	−1.93	13.25	.756	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Composite	320	-	3.80	11.09	.486	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Explanation methods may not help users understand how models will behave

Results

How do users rate explanations?

Users rate each method on a 7-point scale, in response to the question, “Does this explanation show me why the system thought what it did?”

	Text Ratings				Tabular Ratings			
Method	n	μ	CI	σ	n	μ	CI	σ
LIME	144	4.78	1.47	1.76	130	5.36	0.63	1.70
Anchor	133	3.86	0.59	1.79	175	4.99	0.71	1.38
Prototype	191	4.45	1.02	2.08	144	4.20	0.82	1.88
DB	224	3.85	0.60	1.81	144	4.61	1.14	1.86
Composite	240	4.47	0.58	1.70	192	5.10	1.04	1.42

Results

How do users rate explanations?

Users rate each method on a 7-point scale, in response to the question,
"Does this explanation show me why the system thought what it did?"

- Users rated explanations based on quality rather than model correctness
- Ratings are generally higher for tabular data, relative to text data
- The Composite and LIME methods receive the highest ratings

	Text Ratings				Tabular Ratings			
Method	n	μ	CI	σ	n	μ	CI	σ
LIME	144	4.78	1.47	1.76	130	5.36	0.63	1.70
Anchor	133	3.86	0.59	1.79	175	4.99	0.71	1.38
Prototype	191	4.45	1.02	2.08	144	4.20	0.82	1.88
DB	224	3.85	0.60	1.81	144	4.61	1.14	1.86
Composite	240	4.47	0.58	1.70	192	5.10	1.04	1.42

Results

Can users predict explanation effectiveness?

Measure how explanation ratings relate to user correctness in the Post phase of the counterfactual simulation test

Results

Can users predict explanation effectiveness?

Measure how explanation ratings relate to user correctness in the Post phase of the counterfactual simulation test

There is no evidence that explanation ratings are predictive of user correctness

Example:

Rating: 4 -> 5
Correctness: -2.9 ~ 5.2 percentage point change

Qualitative Analysis

- Explanation failure example

Only 7 of 13 responses were correct after seeing explanations (with no method improving correctness)

Original $(\hat{y} = pos)$: "A bittersweet film, simple in form but rich with human events."

Counterfactual $(\hat{y}_c = neg)$: "A teary film, simple in form but vibrant with devoid events."

Discussion

- Forward tests stretch user memory

  Some users reported that it was difficult to retain insights from the learning phase during later prediction rounds

- Counterfactual examples are out of the data distribution

Question?

Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations

Siddhant Arora, Danish Pruthi, Norman Sadeh,
WilliamW. Cohen, Zachary C. Lipton, Graham Neubig

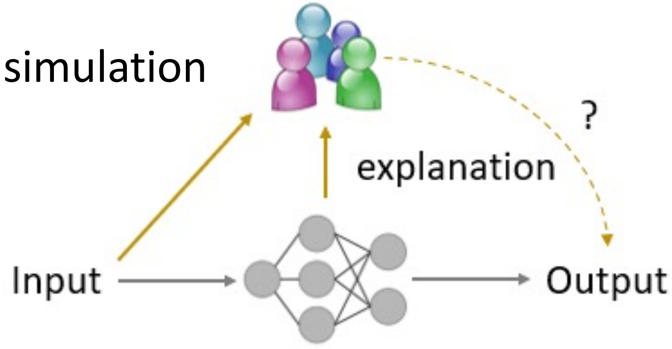
(AAAI, 2022)

Problem

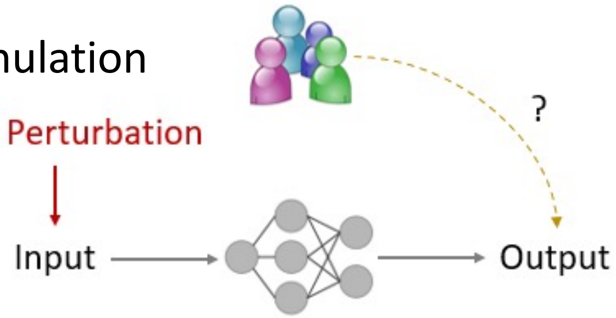
Doshi-Velez and Kim, 2017

Explanations give away the answers

Forward simulation

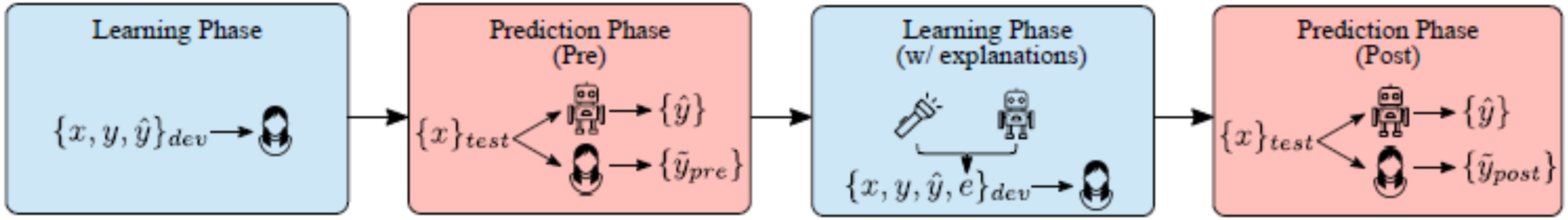


Counterfactual simulation



Hase and Bansal, 2020

Separate the explained instances from the test instances, compare with a baseline

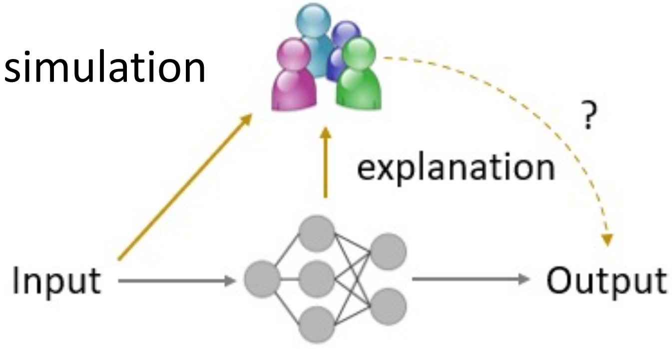


Problem

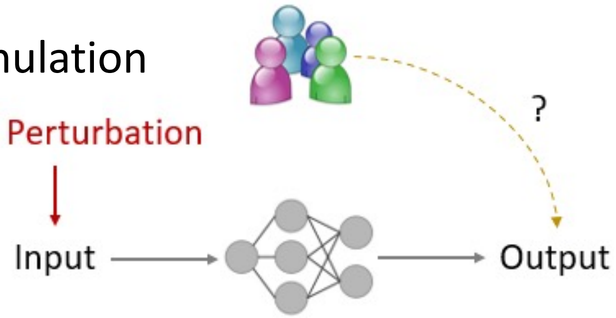
Doshi-Velez and Kim, 2017

Explanations give away the answers

Forward simulation



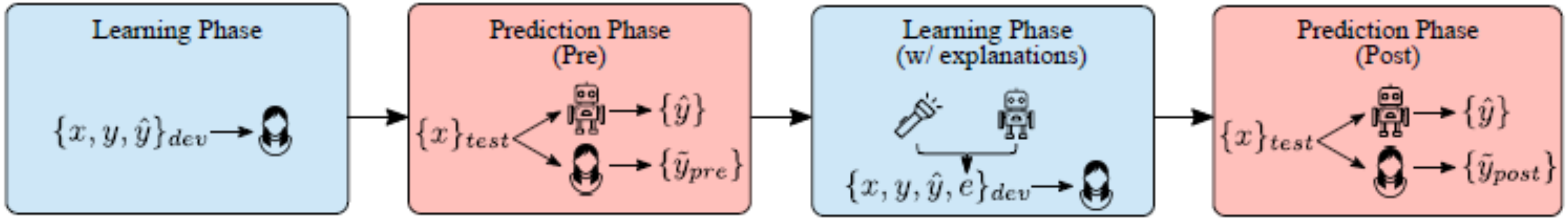
Counterfactual simulation



Hase and Bansal, 2020

Separate the explained instances from the test instances, compare with a baseline

Stretch user memory, no interaction between users and models



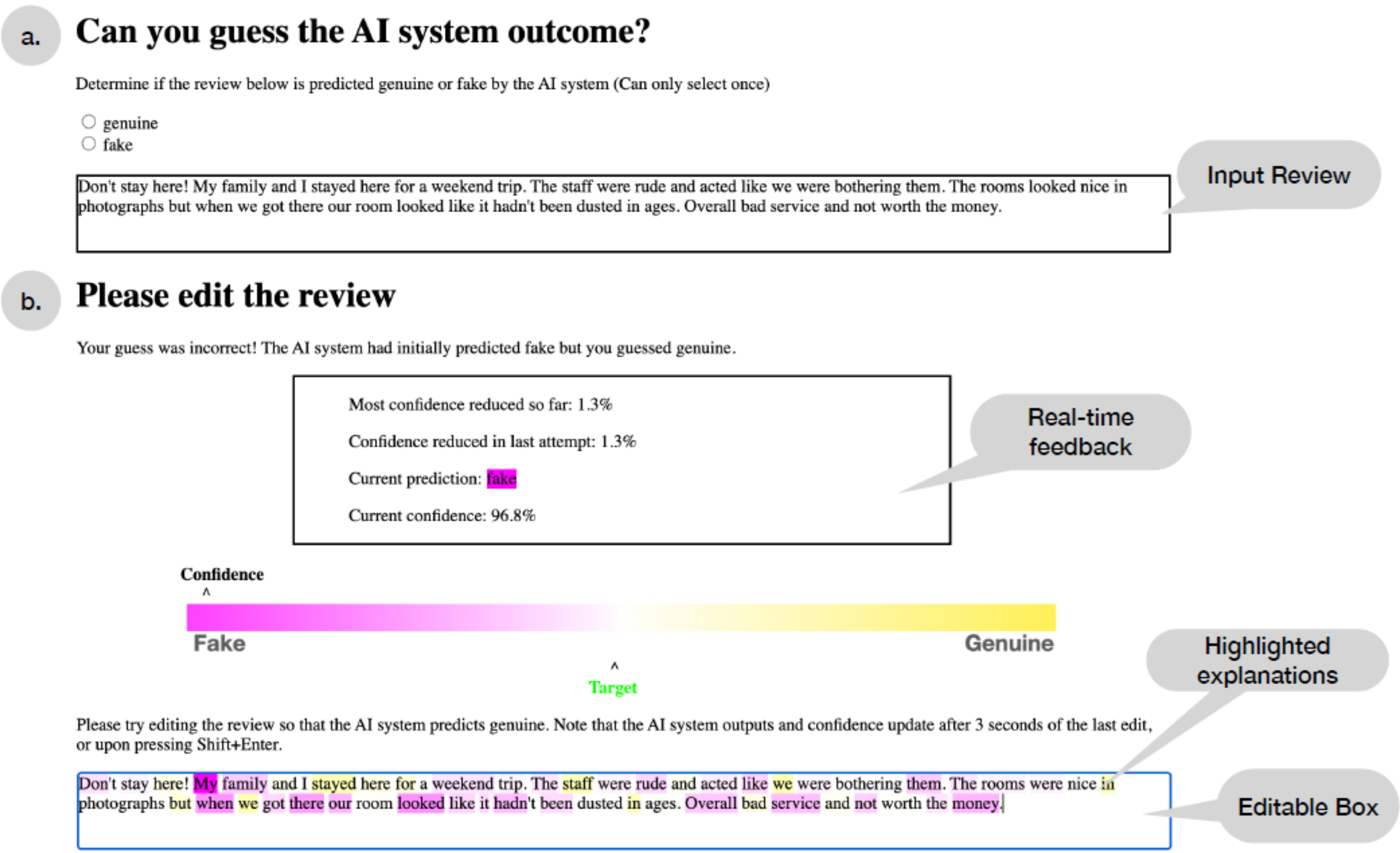
Method

- Provide participants with query access to the model

  Users can alter input documents to observe how model predictions and explanations change in real time

- Prompt participants to edit examples to reduce the model confidence towards the predicted class

Interface


a.
Can you guess the AI system outcome?
Determine if the review below is predicted genuine or fake by the AI system (Can only select once)
genuine
fake
Don't stay here! My family and I stayed here for a weekend trip. The staff were rude and acted like we were bothering them. The rooms looked nice in photographs but when we got there our room looked like it hadn't been dusted in ages. Overall bad service and not worth the money.
Input Review
b.
Please edit the review
Your guess was incorrect! The AI system had initially predicted fake but you guessed genuine.
Most confidence reduced so far: 1.3%
Confidence reduced in last attempt: 1.3%
Current prediction: fake
Current confidence: 96.8%
Real-time feedback
Confidence
Fake
Genuine
Target
Highlighted explanations
Please try editing the review so that the AI system predicts genuine. Note that the AI system outputs and confidence update after 3 seconds of the last edit, or upon pressing Shift+Enter.
Don't stay here! My family and I stayed here for a weekend trip. The staff were rude and acted like we were bothering them. The rooms were nice in photographs but when we got there our room looked like it hadn't been dusted in ages. Overall bad service and not worth the money.
Editable Box

Research Questions

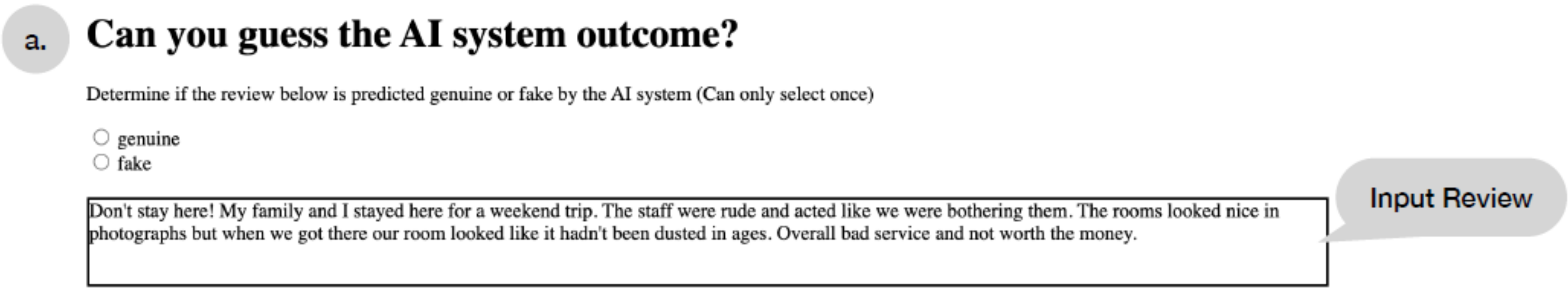
- Which attribution techniques improve humans' ability to guess the model output, or edit the input examples to lower the model confidence?

- Whether the interactive environment with query access to the models makes it possible to distinguish the relative value of different attributions?

Experiments

Training phase → Test phase

Participants first read the input example, and are challenged to guess the model prediction



Experiments

Training phase → Test phase

Then participants see the model output, model confidence and an explanation

Your guess was incorrect! The AI system had initially predicted fake but you guessed genuine.

Most confidence reduced so far: 1.3%

Confidence reduced in last attempt: 1.3%

Current prediction: fake

Current confidence: 96.8%

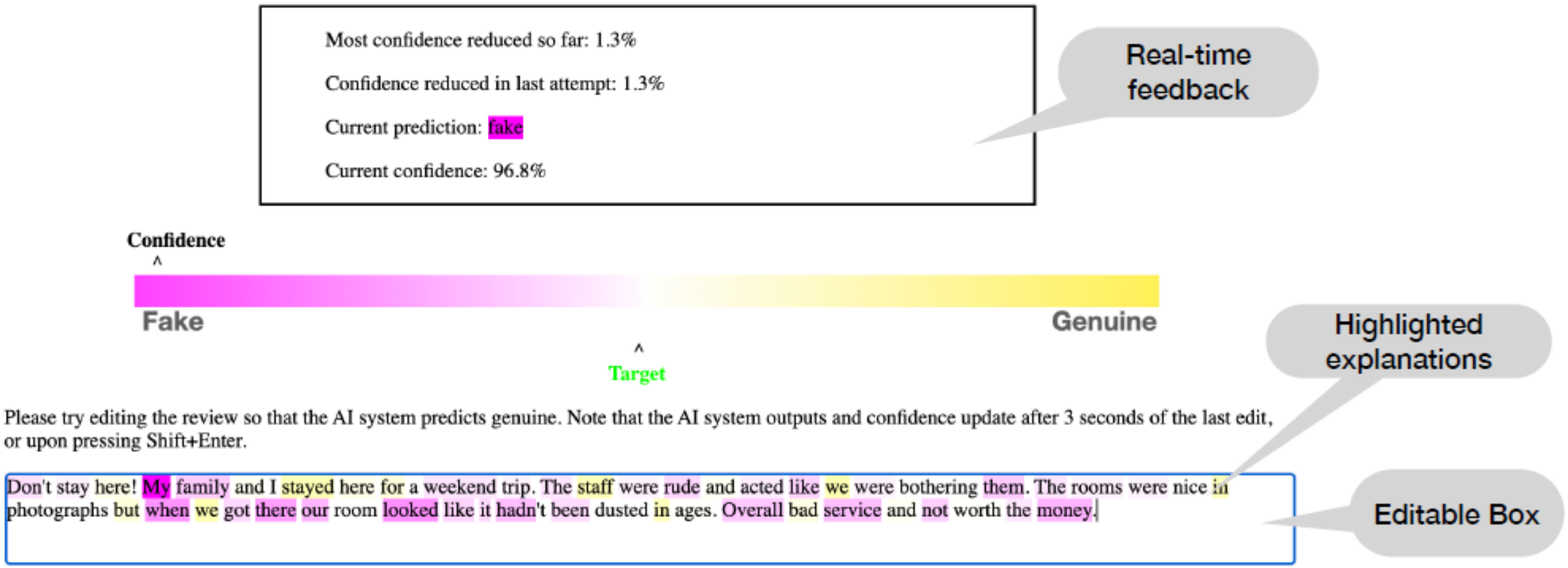
Real-time feedback

Don't stay here! My family and I stayed here for a weekend trip. The staff were rude and acted like we were bothering them. The rooms were nice in photographs but when we got there our room looked like it hadn't been dusted in ages. Overall bad service and not worth the money.

Experiments

Training phase → Test phase

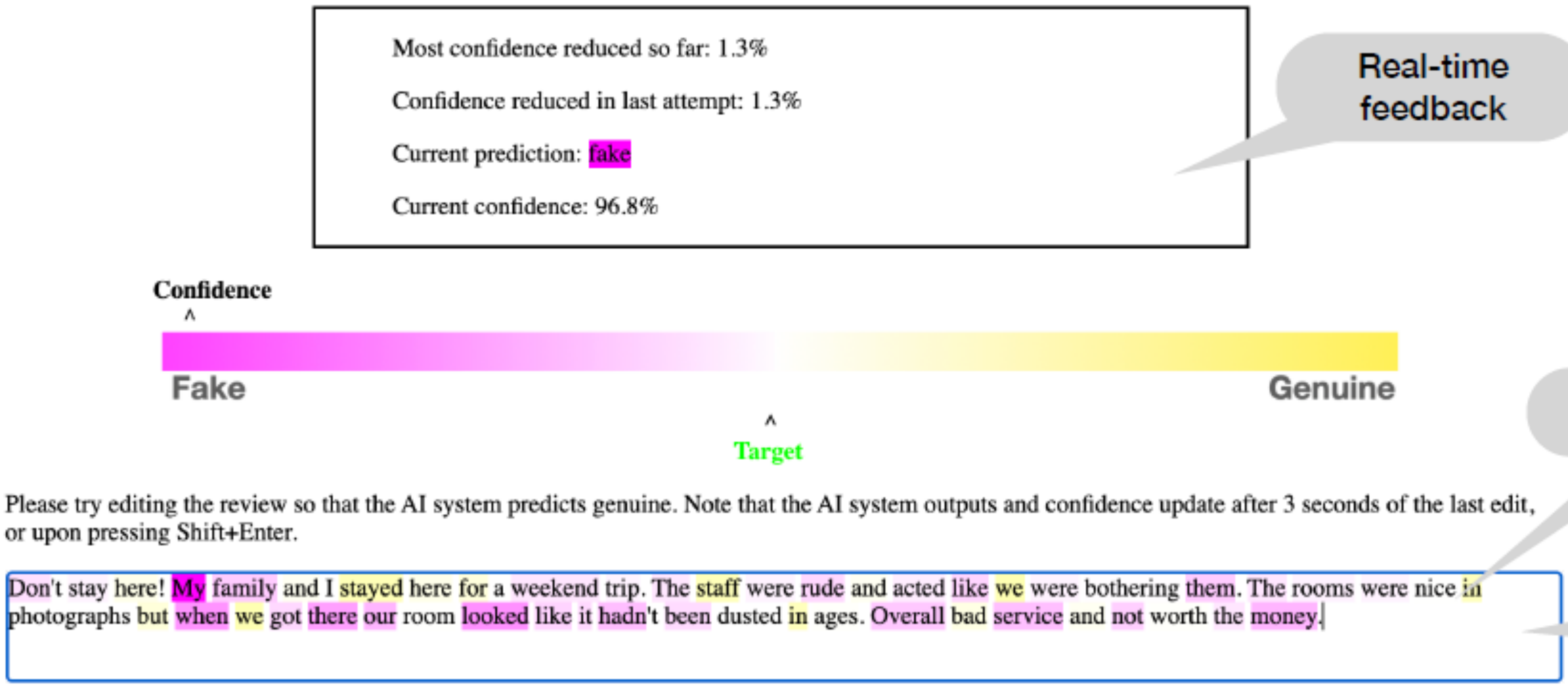
Prompt participants to edit the input text with a goal to lower the confidence of the model prediction



Experiments

Training phase → Test phase

Prompt participants to edit the input text with a goal to lower the confidence of the model prediction



Real-time feedback

Users can validate any hypothesis about the input-output associations

Highlighted explanations

Editable Box

Experiments

Training phase → Test phase

- Similar to the training phase
- Explanations are not available during testing

  Eliminate concerns that the explanations might trivially leak the output

- Iterative training and test

  two training examples + one test example

Question?

A Case Study of Deception Detection

- Task: distinguishing between fake and real hotel reviews [Ott et al., 2011]

- Machine learning models perform much better than humans
- Models may exploit subtle, unknown and possibly counter-intuitive associations to drive prediction

Model	Accuracy
Human Accuracy (Ott et al. 2011)	$\approx 60\%$
Logistic Regression	87.8%
BERT	89.8%

A Case Study of Deception Detection

- Task: distinguishing between fake and real hotel reviews [Ott et al., 2011]

- Machine learning models perform much better than humans
- Models may exploit subtle, unknown and possibly counter-intuitive associations to drive prediction

Model	Accuracy
Human Accuracy (Ott et al. 2011)	$\approx$ 60%
Logistic Regression	87.8%
BERT	89.8%

Explanations help humans in understanding the input-output associations that models exploit?

A Case Study of Deception Detection

- What are permissible edits?

  - Participants cannot alter the staying experience conveyed through the hotel review
  - If the review is positive, negative or mixed, then the edited version should maintain that stance
  - Participants are allowed to paraphrase and can remove or change information not relevant to the experience about the hotel

"My husband and I" -> "We" ✓

Add "The staff was unfriendly" ✗

A Case Study of Deception Detection

- Model and explanations
  - Logistic regression

    Explanations: feature coefficients of unigram features

  - BERT

    Local explanations: LIME, IG

    Global explanations:

    Linear *student* model ≈ BERT

    ⇓

    feature coefficients

Results

Do explanations help humans simulate models?

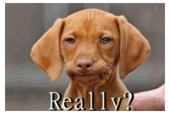
Investigate if the query access to the model's predictions and explanations during the training phase enables participants to understand the models sufficiently to simulate its output on unseen test examples

Results

Do explanations help humans simulate models?

Investigate if the query access to the model's predictions and explanations during the training phase enables participants to understand the models sufficiently to simulate its output on unseen test examples

No evidence of improved simulatability



Model	Treatments	Simulation Accuracy	Phase	Examples flipped (Percentage)	Avg. Confidence Reduced
Logistic Regression	Control	54.5 [51.0, 58.0]	Train	8.2 [5.4, 11.6]	8.0 [7.0, 9.0]
			Test	15.0 [10.8, 19.4]	5.9 [4.3, 7.8]
	Feature coefficients	53.1 [50.0, 57.0]	Train	**36.7 [24.8, 49.3]**	**21.3 [19.5, 23.1]**
			Test	16.0 [10.8, 21.6]	**8.9 [7.2, 10.6]**
BERT	Control	57.1 [54.0, 61.0]	Train	15.0 [11.6, 18.8]	10.7 [8.6, 12.8]
			Test	12.4 [7.6, 18.1]	9.2 [6.6, 11.9]
	LIME	56.4 [53.0, 60.0]	Train	14.4 [10.5, 19.5]	10.2 [8.2, 12.3]
			Test	7.7 [4.4, 11.3]	6.1 [4.1, 8.2]
	Integrated gradients	56.6 [54.0, 60.0]	Train	**23.6 [19.4, 28.0]**	**16.5 [14.0, 19.2]**
			Test	13.6 [8.2, 19.3]	10.4 [7.7, 13.3]
	Feature coefficients (from a linear student)	60.5 [57.0, 64.0]	Train	**32.2 [27.1, 37.3]**	**22.6 [19.7, 25.6]**
			Test	**21.3 [15.7, 27.4]**	**14.9 [11.6, 18.4]**
	+ global cues (from a linear student)	55.7 [51.0, 60.0]	Train	**40.6 [32.0, 49.6]**	**29.9 [26.8, 33.0]**
			Test	**31.6 [23.2, 40.8]**	**23.6 [19.7, 27.6]**

No explanations

None of the explanations help improve simulation accuracy

Results

Do explanations help humans perform edits that reduce the model confidence?

Examine if participants gain sufficient understanding during the training phase to perform edits that cause the models to lower the confidence towards the originally predicted class

Results

Do explanations help humans perform edits that reduce the model confidence?

Examine if participants gain sufficient understanding during the training phase to perform edits that cause the models to lower the confidence towards the originally predicted class

Model	Treatments	Simulation Accuracy	Phase	Examples flipped (Percentage)	Avg. Confidence Reduced
Logistic Regression	Control	54.5 [51.0, 58.0]	Train	8.2 [5.4, 11.6]	8.0 [7.0, 9.0]
			Test	15.0 [10.8, 19.4]	5.9 [4.3, 7.8]
	Feature coefficients	53.1 [50.0, 57.0]	Train	**36.7 [24.8, 49.3]**	**21.3 [19.5, 23.1]**
			Test	16.0 [10.8, 21.6]	**8.9 [7.2, 10.6]**
BERT	Control	57.1 [54.0, 61.0]	Train	15.0 [11.6, 18.8]	10.7 [8.6, 12.8]
			Test	12.4 [7.6, 18.1]	9.2 [6.6, 11.9]
	LIME	56.4 [53.0, 60.0]	Train	14.4 [10.5, 19.5]	10.2 [8.2, 12.3]
			Test	7.7 [4.4, 11.3]	6.1 [4.1, 8.2]
	Integrated gradients	56.6 [54.0, 60.0]	Train	**23.6 [19.4, 28.0]**	**16.5 [14.0, 19.2]**
			Test	13.6 [8.2, 19.3]	10.4 [7.7, 13.3]
	Feature coefficients (from a linear student)	60.5 [57.0, 64.0]	Train	**32.2 [27.1, 37.3]**	**22.6 [19.7, 25.6]**
			Test	**21.3 [15.7, 27.4]**	**14.9 [11.6, 18.4]**
	+ global cues (from a linear student)	55.7 [51.0, 60.0]	Train	**40.6 [32.0, 49.6]**	**29.9 [26.8, 33.0]**
			Test	**31.6 [23.2, 40.8]**	**23.6 [19.7, 27.6]**

Logistic regression coefficient weights help participants reduce the model confidence

Results

Do explanations help humans perform edits that reduce the model confidence?

Examine if participants gain sufficient understanding during the training phase to perform edits that cause the models to lower the confidence towards the originally predicted class

Model	Treatments	Simulation Accuracy	Phase	Examples flipped (Percentage)	Avg. Confidence Reduced
Logistic Regression	Control	54.5 [51.0, 58.0]	Train	8.2 [5.4, 11.6]	8.0 [7.0, 9.0]
			Test	15.0 [10.8, 19.4]	5.9 [4.3, 7.8]
	Feature coefficients	53.1 [50.0, 57.0]	Train	**36.7** **[24.8, 49.3]**	**21.3** **[19.5, 23.1]**
			Test	16.0 [10.8, 21.6]	**8.9** **[7.2, 10.6]**
BERT	Control	57.1 [54.0, 61.0]	Train	15.0 [11.6, 18.8]	10.7 [8.6, 12.8]
			Test	12.4 [7.6, 18.1]	9.2 [6.6, 11.9]
	LIME	56.4 [53.0, 60.0]	Train	14.4 [10.5, 19.5]	10.2 [8.2, 12.3]
			Test	7.7 [4.4, 11.3]	6.1 [4.1, 8.2]
	Integrated gradients	56.6 [54.0, 60.0]	Train	**23.6** **[19.4, 28.0]**	**16.5** **[14.0, 19.2]**
			Test	13.6 [8.2, 19.3]	10.4 [7.7, 13.3]
	Feature coefficients (from a linear student)	60.5 [57.0, 64.0]	Train	**32.2** **[27.1, 37.3]**	**22.6** **[19.7, 25.6]**
			Test	**21.3** **[15.7, 27.4]**	**14.9** **[11.6, 18.4]**
	+ global cues (from a linear student)	55.7 [51.0, 60.0]	Train	**40.6** **[32.0, 49.6]**	**29.9** **[26.8, 33.0]**
			Test	**31.6** **[23.2, 40.8]**	**23.6** **[19.7, 27.6]**

During the training phase, users are able to flip more predictions, however, this ability does not transfer to the test phase

Results

Do explanations help humans perform edits that reduce the model confidence?

Examine if participants gain sufficient understanding during the training phase to perform edits that cause the models to lower the confidence towards the originally predicted class

Model	Treatments	Simulation Accuracy	Phase	Examples flipped (Percentage)	Avg. Confidence Reduced
Logistic Regression	Control	54.5 [51.0, 58.0]	Train	8.2 [5.4, 11.6]	8.0 [7.0, 9.0]
			Test	15.0 [10.8, 19.4]	5.9 [4.3, 7.8]
	Feature coefficients	53.1 [50.0, 57.0]	Train	**36.7** **[24.8, 49.3]**	**21.3** **[19.5, 23.1]**
			Test	16.0 [10.8, 21.6]	**8.9** **[7.2, 10.6]**
BERT	Control	57.1 [54.0, 61.0]	Train	15.0 [11.6, 18.8]	10.7 [8.6, 12.8]
			Test	12.4 [7.6, 18.1]	9.2 [6.6, 11.9]
	LIME	56.4 [53.0, 60.0]	Train	14.4 [10.5, 19.5]	10.2 [8.2, 12.3]
			Test	7.7 [4.4, 11.3]	6.1 [4.1, 8.2]
	Integrated gradients	56.6 [54.0, 60.0]	Train	**23.6** **[19.4, 28.0]**	**16.5** **[14.0, 19.2]**
			Test	13.6 [8.2, 19.3]	10.4 [7.7, 13.3]
	Feature coefficients (from a linear student)	60.5 [57.0, 64.0]	Train	**32.2** **[27.1, 37.3]**	**22.6** **[19.7, 25.6]**
			Test	**21.3** **[15.7, 27.4]**	**14.9** **[11.6, 18.4]**
	+ global cues (from a linear student)	55.7 [51.0, 60.0]	Train	**40.6** **[32.0, 49.6]**	**29.9** **[26.8, 33.0]**
			Test	**31.6** **[23.2, 40.8]**	**23.6** **[19.7, 27.6]**

For the BERT model, neither LIME nor IG help participants flip more predictions or reduce confidence at the test phase

Results

Do explanations help humans perform edits that reduce the model confidence?

Examine if participants gain sufficient understanding during the training phase to perform edits that cause the models to lower the confidence towards the originally predicted class

Model	Treatments	Simulation Accuracy	Phase	Examples flipped (Percentage)	Avg. Confidence Reduced
Logistic Regression	Control	54.5 [51.0, 58.0]	Train	8.2 [5.4, 11.6]	8.0 [7.0, 9.0]
			Test	15.0 [10.8, 19.4]	5.9 [4.3, 7.8]
	Feature coefficients	53.1 [50.0, 57.0]	Train	**36.7** **[24.8, 49.3]**	**21.3** **[19.5, 23.1]**
			Test	16.0 [10.8, 21.6]	**8.9** **[7.2, 10.6]**
BERT	Control	57.1 [54.0, 61.0]	Train	15.0 [11.6, 18.8]	10.7 [8.6, 12.8]
			Test	12.4 [7.6, 18.1]	9.2 [6.6, 11.9]
	LIME	56.4 [53.0, 60.0]	Train	14.4 [10.5, 19.5]	10.2 [8.2, 12.3]
			Test	7.7 [4.4, 11.3]	6.1 [4.1, 8.2]
	Integrated gradients	56.6 [54.0, 60.0]	Train	**23.6** **[19.4, 28.0]**	**16.5** **[14.0, 19.2]**
			Test	13.6 [8.2, 19.3]	10.4 [7.7, 13.3]
	Feature coefficients (from a linear student)	60.5 [57.0, 64.0]	Train	**32.2** **[27.1, 37.3]**	**22.6** **[19.7, 25.6]**
			Test	**21.3** **[15.7, 27.4]**	**14.9** **[11.6, 18.4]**
	+ global cues (from a linear student)	55.7 [51.0, 60.0]	Train	**40.6** **[32.0, 49.6]**	**29.9** **[26.8, 33.0]**
			Test	**31.6** **[23.2, 40.8]**	**23.6** **[19.7, 27.6]**

Global interpretations from the linear student model help participants flip more predictions and reduce confidence

Results

Do participants edit tokens highlighted as explanations? Are their edits effective?

- Monitor whether participants are paying attention to the explanations, specifically by measuring how they respond to highlighted words
- Record the fraction of times edits are performed on a word that is among the top-20% of highlighted words in a given input text

Results

Do participants edit tokens highlighted as explanations? Are their edits effective?

- Monitor whether participants are paying attention to the explanations, specifically by measuring how they respond to highlighted words
- Record the fraction of times edits are performed on a word that is among the top-20% of highlighted words in a given input text

Yes, participants edit the highlighted words significantly more often

Results

Do participants edit tokens highlighted as explanations? Are their edits effective?

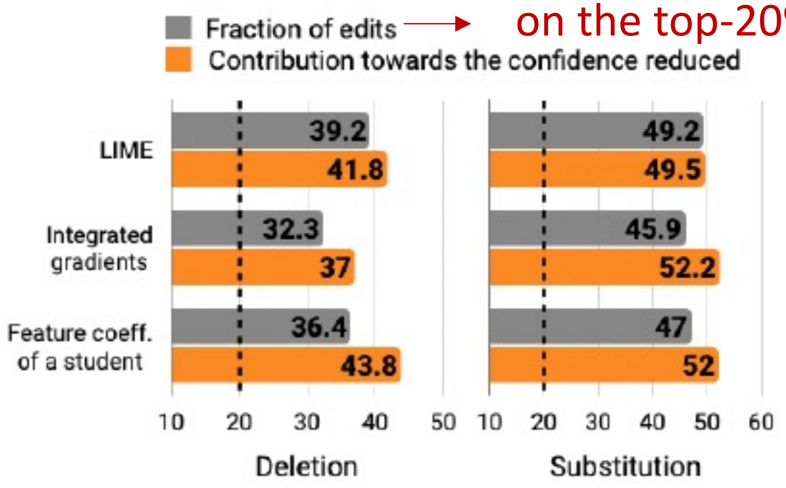
- Monitor whether participants are paying attention to the explanations, specifically by measuring how they respond to highlighted words
- Record the fraction of times edits are performed on a word that is among the top-20% of highlighted words in a given input text

Yes, participants edit the highlighted words significantly more often

The edits on the top-20% highlighted words are effective in reducing model confidence?

- Yes, the edits on highlighted words are more effective
- IG and global interpretations are more effective than LIME

on the top-20% highlighted words



Discussion

- Separating learning and predicting phase is too challenging for humans to understand model prediction behavior

- The number of examples for learning is limited

Question?

Reference

- Jacovi, Alon, and Yoav Goldberg. "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?." *arXiv preprint arXiv:2004.03685* (2020).
- Hase, Peter, and Mohit Bansal. "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?." *arXiv preprint arXiv:2005.01831* (2020).
- Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." *arXiv preprint arXiv:1702.08608* (2017).
- Arora, Siddhant, et al. "Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations." *arXiv preprint arXiv:2112.09669* (2021).
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model agnostic explanations. In AAAI Conference on Artificial Intelligence.