“Why is ‘Chicago’ deceptive?”: Towards Building Model-Driven Tutorials for Humans

Vivian Lai, Han Liu, and Chenhao Tan
@vivwylai | @HanLiuAI | @ChenhaoTan
University of Colorado Boulder
machineintheloop.com
AI used in societally critical tasks

Medical diagnosis

Recidivism prediction

Amazon secret AI hiring tool

Autonomous driving

Geiger et al. 2012; European Parliament 2016; Kleinberg et al. 2017; Dastin 2018
Explanations!
Explaining AI is tricky

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!
Why is explaining AI tricky?

Two distinct learning modes

Emulating

Discovering
Why is explaining AI tricky?

Two distinct learning modes

- Emulating
Why is explaining AI tricky?

Two distinct learning modes

Discovering

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!
Why is explaining AI tricky?

Two distinct learning modes

Discovering

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

AI can discover inconspicuous and counterintuitive patterns.
So, how can explaining AI be less tricky?

Model-driven tutorials

- Elucidate counterintuitive patterns
- Enhance humans' ability to understand patterns
Model-driven tutorials: Guidelines

- Deceptive reviews tend to focus on aspects that are external to the hotel being reviewed, e.g., husband, business, vacation.
- Deceptive reviews tend to contain more emotional terms; positive deceptive reviews are generally more positive and negative deceptive reviews are more negative than genuine reviews.
- Genuine reviews tend to include more sensorial and concrete language, in particular, genuine reviews are more specific about spatial configurations, e.g., small, bathroom, on, location.
- Deceptive reviews tend to contain more verbs, e.g., eat, sleep, stay.
- Deceptive reviews tend to contain more superlatives, e.g., cleanest, worst, best.
- Deceptive reviews tend to contain more pre-determiners, which are normally placed before an indefinite article + adjective + noun, e.g., what a lovely day!

State-of-the-art science communication
Model-driven tutorials: Examples

How do we choose examples?

- SP-LIME
- Spaced repetition
Model-driven tutorials: Examples

How do we choose examples?

- SP-LIME
- Spaced repetition

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You are wrong. The review is deceptive.

The AI is right. It predicts this review as deceptive because the red words (ask, back, be, chicago, i, my, will) are associated with deceptive reviews and the green words (on, that, very) are associated with genuine reviews. The darkness shows the degree of the association with the two categories.
Experimental Design & Research Questions

Training

Prediction
Experimental Design & Research Questions

RQ1: Effect of different tutorials

Different tutorials: Training

No assistance: Prediction
Experimental Design & Research Questions

RQ1: Effect of different tutorials
Experimental Design & Research Questions

RQ2: Effect of real-time assistance

Training

Same tutorial (Spaced repetition)

Prediction

Different real-time assistance
Experimental Design & Research Questions

- Training
  - RQ1: Effect of different tutorials

- Prediction
  - RQ2: Effect of real-time assistance
Experimental Design & Research Questions

RQ1 & RQ2

Linear model

RQ3

Deep model
Experimental Design & Research Questions

**Training**
- RQ1: Effect of different tutorials

**Prediction**
- RQ2: Effect of real-time assistance

**RQ3: Effect of model complexity**
Experimental Design & Research Questions

- **Training**
  - RQ1: Effect of different tutorials

- **Prediction**
  - RQ2: Effect of real-time assistance

- **RQ3: Effect of model complexity**

Performed qualitative study to improve interface design.
Research question 1

Can model-driven tutorials improve human performance without any real-time assistance in the prediction phase?
Tutorials are useful to some extent

- Control: 54.6%, p=0.018*
- Guidelines: 60.4%
- Spaced repetition: 57.9%, p=0.1
- Spaced repetition + guidelines: 59.2%

# of stars indicates p-values
***: p < 0.001
**: p < 0.01
*: p < 0.05

Tutorials are useful to some extent.
Tutorials are useful to some extent

- Control: 54.6% (p=0.018*)
- Guidelines: 60.4%
- Spaced repetition: 57.9% (p=0.1)
- Spaced repetition + guidelines: 59.2%

“The tutorial is helpful but it’s just hard not being able to reference it.”
If not, how do varying levels of real-time assistance in prediction phase affect human performance after training?
Prediction: various levels of real-time assistance

Information from AI increases from left to right.
Prediction: various levels of real-time assistance
You made **Chicago** a wonderful stay! The room was gorgeous! I came with **very** little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to **ask**! Thank you so much, I will be back! Very tidy room as well!
Real-time assistance improves performance

- No assistance: 60.4%
- Unsigned: 57.8%
- Signed: 70.7%
- Signed + predicted label + guidelines + accuracy: 74%
- Machine: 86%

# of stars indicates p-values

- ***: p < 0.001
- **: p < 0.01
- *: p < 0.05

Real-time assistance improves performance.
Signed highlights is sufficient

![Bar chart showing accuracy percentages]

- **No assistance**: 60.4%
- **Unsigned**: 57.8%
- **Signed**: 70.7%
- **Signed + predicted label + guidelines + accuracy**: 74%
- **Machine**: 86%

# of stars indicates p-values

- ***: p < 0.001
- **: p < 0.01
- *: p < 0.05

p > 0.05
Gap between human+AI & AI

- No assistance: 60.4%
- Unsigned: 57.8%
- Signed: 70.7%
- Signed + predicted label + guidelines + accuracy: 74%
- Machine: 86%

# of stars indicates p-values

- ***: p < 0.001
- **: p < 0.01
- *: p < 0.05

Poursabzi-Sangdeh et al. 2018; Green & Chen 2019; Lage et al. 2019; Lai & Tan 2019; Carton et al. 2020; Lai et al. 2020
Can our results generalize in other models? How do model complexity and explanation methods affect human performance with/without training?
You made **Chicago** a wonderful stay! The room was gorgeous! I came with **very little on hand** and **my** deluxe room supplied me with everything that I needed, I didn’t even have to ask! Thank you so much, I will be back! Very tidy **room** as well!

You made **Chicago** a wonderful stay! The room was gorgeous! I came with very little on hand and **my** deluxe room supplied **me** with everything that I needed, I didn’t even have to ask! Thank you so much, I will be back! Very tidy **room** as well!
You made **Chicago** a wonderful stay! The room was gorgeous! I came with **very little on hand** and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! **Very tidy room as well!**
Simple model = better human performance

Lai et al. 2019

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Training</th>
<th>No training</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>64.1%</td>
<td>72.8%</td>
<td></td>
</tr>
<tr>
<td>BERT-ATT</td>
<td>58.2%</td>
<td></td>
<td>54.1%</td>
</tr>
<tr>
<td>BERT-LIME</td>
<td>59.2%</td>
<td></td>
<td>64.9%</td>
</tr>
</tbody>
</table>

# of stars indicates p-values

***: p < 0.001
**: p < 0.01
*: p < 0.05
Simple model = better human performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>64.1%</td>
<td>p=0.001***</td>
</tr>
<tr>
<td>BERT-ATT</td>
<td>58.2%</td>
<td>p=0.001***</td>
</tr>
<tr>
<td>BERT-LIME</td>
<td>59.2%</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy (%) vs. Accuracy (%)

Least important: SVM, BERT-ATT
Most important: BERT-LIME

I recently stayed in the **Talbott Hotel** in Chicago with my family of three. I absolutely loved it and will definitely return! The customer service of the staff at the front desk was fantastic. They were very helpful in helping us decide on what activities to do and what restaurants to visit during our stay. For the nights we didn't want to go out, we had room service. The food from the Hotel restaurant was so delicious and the service was very quick. I also loved their commitment to the environment. It is clear to me why they have won so many awards. Thank you **Talbott** for a wonderful stay!
Training leads to better performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>No training</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>72.8%</td>
<td>64.1%</td>
<td>0.001***</td>
</tr>
<tr>
<td>BERT-ATT</td>
<td>58.2%</td>
<td>54.1%</td>
<td>0.001***</td>
</tr>
<tr>
<td>BERT-LIME</td>
<td>64.9%</td>
<td>59.2%</td>
<td>0.001***</td>
</tr>
</tbody>
</table>

# of stars indicates p-values

***: p < 0.001  
**: p < 0.01  
*: p < 0.05
Takeaway

✅ Tutorials somewhat improve human performance

✅ Explanations from simple models are preferred

✅ Future directions for human-centered tutorials and explanations

Vivian Lai, Han Liu, Chenhao Tan
@vivwylai | vivwylai@gmail.com
@HanLiuAI | @ChenhaoTan
University of Colorado Boulder
Website: machineintheloop.com
Paper: https://tinyurl.com/model-driven-tutorials
Workshop: https://tinyurl.com/harness-explanations