

# Thomson Reuters Labs

## The Role of Explanations of AI Systems: Beyond Trust and Helping to Form Mental Models

Milda Norkute

September 5th, 2021

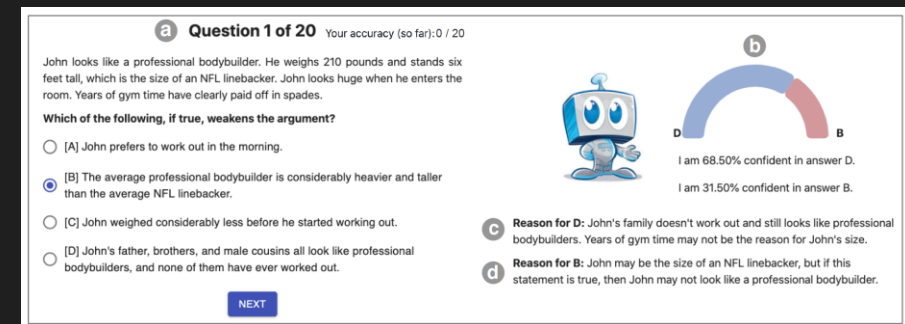
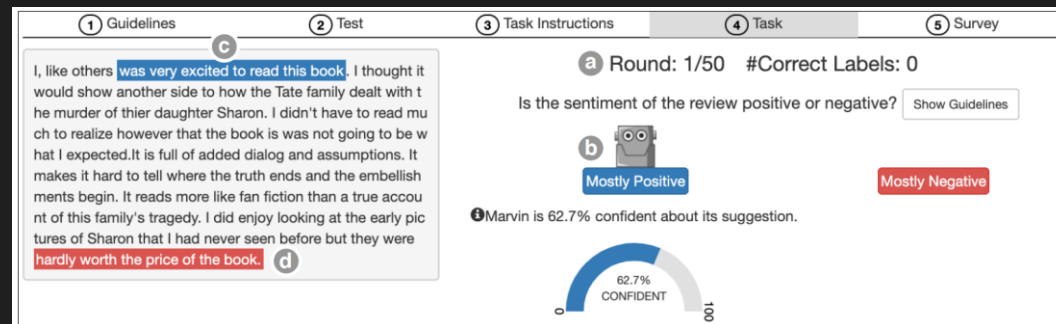
# Roles of explanations

- Mediate trust
- Help form mental model

# Trust and mental models

Bansal et al. 2021 observed that explanations increased the chance that humans will accept the AI's recommendation, regardless of its correctness.

The authors suggest that explanations should be informative, instead of convincing to follow the recommendation.



Some participants developed mental models of the AI's confidence score to determine when to trust the AI, but they built different mental models.

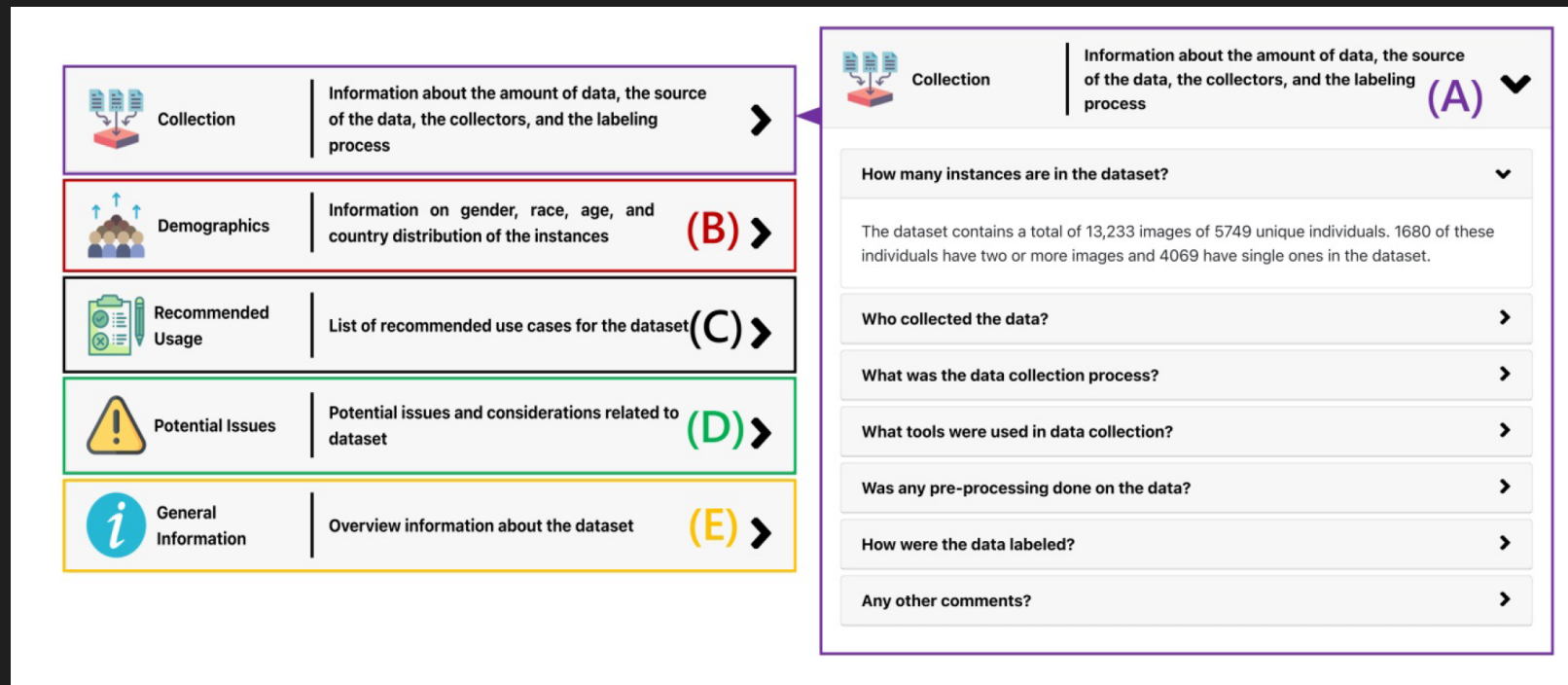
Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3411764.3445717>

# Roles of explanations

- Mediate trust
- Help form mental model
- Inform, not convince

# Trust and mental models

Anik and Bunt (2021) explored the concept of data-centric explanations where the explanations describe the training data to end-users. They found that participants trust in AI system was impacted positively when the training data seemed balanced and negatively when the explanations revealed problems.




Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3411764.3445736>

# Roles of explanations

- Mediate trust
- Help form mental model
- Inform, not convince
- Assess fairness?

# Highlighting for Summary verification



Make the edits to the Allegations Summary for the document and submit when ready.

Defendant failed to maintain its premises in a reasonably safe and suitable condition, causing Plaintiff to trip, fall and sustain injuries.

[Link to PDF](#) [Submit](#)

## Attention Highlights

corporation incorporated in Pennsylvania with a registered office address of 1357 Samantha Way, North Huntingdon, Westmoreland County, Pennsylvania 15642.

Page 6

5. The premises in question is the residential home of Mr. and Mrs. <PERSON\_1> as well as the business premises of Defendant <PERSON\_1> & Company located at 1357 Samantha Way, North Huntingdon, Westmoreland County, Pennsylvania 15642. 6. <PERSON\_10> times relevant and material hereto, the Defendants leased, owned, operated, possessed, controlled, managed and/or maintained the premises and had a duty to inspect, maintain, repair, control, supervise and oversee the at-issue premises and to warn of and correct the dangerous conditions. 7. <PERSON\_10> times relevant and material hereto, the Defendants acted by and through their agents, servants, employees, representatives, assignees, subsidiaries, predecessors and successors in interest. 8. On or about February 25, 2014, the Plaintiff was lawfully on the aforementioned premises for a business purpose. 9. At all times relevant and material hereto, there existed a dangerous, defective, hazardous and unsafe condition on the premises of the Defendants, characterized by ice which was allowed to accumulate on the driveway of the premises. 10. Plaintiff was caused to trip, slip and/or otherwise lose her balance as a result of coming into contact with the aforementioned defective condition. 11. <PERSON\_11> and proximate result of the aforementioned accident, Plaintiff sustained the following injuries, some or all of which are or may be permanent:

a. Comminuted fracture of the left distal radius;

b. Bruises, contusions and other injuries in or about nerves, muscles, bones, tendons, ligaments, tissues and vessels of the body; and

c. Nervousness, emotional tension, anxiety and depression.

Page 7

12. <PERSON\_11> and proximate result of the aforementioned accident, Plaintiff sustained

**Influence of text on the summary**

weak strong

**Pick a page to jump to**

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9



Make the edits to the Allegations Summary for the document and submit when ready.

Defendant failed to maintain its premises in a reasonably safe and suitable condition, causing Plaintiff to trip, fall and sustain injuries.

[Link to PDF](#) [Submit](#)

## Source Highlights

corporation incorporated in Pennsylvania with a registered office address of 1357 Samantha Way, North Huntingdon, Westmoreland County, Pennsylvania 15642.

Page 6

5. The premises in question is the residential home of Mr. and Mrs. <PERSON\_1> as well as the business premises of Defendant <PERSON\_1> & Company located at 1357 Samantha Way, North Huntingdon, Westmoreland County, Pennsylvania 15642. 6. <PERSON\_10> times relevant and material hereto, the Defendants leased, owned, operated, possessed, controlled, managed and/or maintained the premises and had a duty to inspect, maintain, repair, control, supervise and oversee the at-issue premises and to warn of and correct the dangerous conditions. 7. <PERSON\_10> times relevant and material hereto, the Defendants acted by and through their agents, servants, employees, representatives, assignees, subsidiaries, predecessors and successors in interest. 8. On or about February 25, 2014, the Plaintiff was lawfully on the aforementioned premises for a business purpose. 9. At all times relevant and material hereto, there existed a dangerous, defective, hazardous and unsafe condition on the premises of the Defendants, characterized by ice which was allowed to accumulate on the driveway of the premises. 10. Plaintiff was caused to trip, slip and/or otherwise lose her balance as a result of coming into contact with the aforementioned defective condition. 11. <PERSON\_11> and proximate result of the aforementioned accident, Plaintiff sustained the following injuries, some or all of which are or may be permanent:

a. Comminuted fracture of the left distal radius;

b. Bruises, contusions and other injuries in or about nerves, muscles, bones, tendons, ligaments, tissues and vessels of the body; and

c. Nervousness, emotional tension, anxiety and depression.

Page 7

12. <PERSON\_11> and proximate result of the aforementioned accident, Plaintiff sustained

**Predicted influence of text on the summary**

☒ Text that had influence

☐ Text that had no influence

**Pick a page to jump to**

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

Participants were faster with attention but not source highlights when reviewing the summaries

## Attention highlights:

- gave them a sense the whole complaint was looked at
- took them to key details included in the summary
- helped spot additional details
- helped to correct the summary
- task more enjoyable

## Source Highlights:

- helped navigate to the section of the document where relevant details might be
- Taks more enjoyable

Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, Sally Gao. 2021. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 10 pages. DOI: <https://doi.org/10.1145/3411763.3443441>

# Roles of explanations

- Mediate trust
- Help form mental model
- Inform, not convince
- Increase efficiency
- Increase satisfaction
- Enrich AI suggestions
- Correct AI suggestions



# Roles of explanations

## Global

- Mediate trust
- Help form mental model
- Inform, not convince
- Help assess fairness
- Increase efficiency
- Increase satisfaction
- Enrich AI suggestions
- Correct AI suggestions
- ?

## Task specific

- Navigation of document
- Highlighting relevant details

# Can explanations have standalone value?

	Explanation	No Explanation
Summary	Summary <i>and</i> Text Highlighting	Summary Only
No Summary	Text Highlighting Only	No ML Outputs

What value, if any, could the highlighting have alone?

# Can explanations have standalone value?

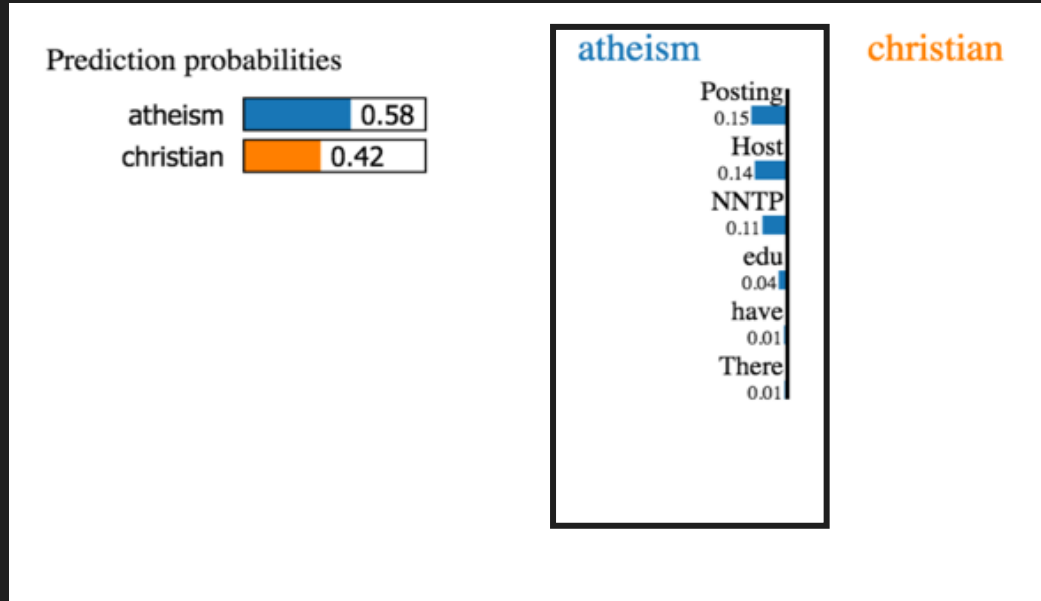


Image taken from:  
<https://github.com/marcotcr/lime>

Could seeing only a list of words that influenced the decision as guidance on what to look for what deciding? What task specific value could that have?

# Conclusions

Explanations can play many roles

Global vs task specific value (context specific)

Standalone value of explanations?

# Thank you



milda.norkute@tr.com



@milda\_nor