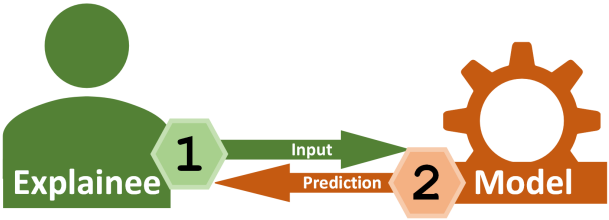
Mediators: Conversational Agents Explaining NLP Model Behavior

Nils Feldhus, Ajay Madhavan Ravichandran, Sebastian Möller

German Research Center for Artificial Intelligence (DFKI), Berlin
Speech and Language Technology group

July 23, 2022
IJCAI-ECAI 2022 Workshop on Explainable Artificial Intelligence (XAI)

Sentiment Analysis



		$p(y\|\mathbf{x};\theta)$	y
x	the year 's best and most unpredictable comedy	0.91	pos
x	we never feel anything for these characters	0.95	neg
x	handsome but unfulfilling suspense drama	0.18	neg

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. 2013.
Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey. 2022.

Feature Attribution

		$p(y\|\mathbf{x};\theta)$	y	c
x	the year 's best and most unpredictable comedy	0.91	pos	pos
x	we never feel anything for these characters	0.95	neg	neg
x	handsome but unfulfilling suspense drama	0.18	neg	pos

Feature Attribution

		$p(y\|\mathbf{x};\theta)$	y	c
x	the year 's best and most unpredictable comedy	0.91	pos	pos
x	we never feel anything for these characters	0.95	neg	neg
x	handsome but unfulfilling suspense drama	0.18	neg	pos

- **static** : One-off explanations do not allow increased interaction (Lakkaraju et al., 2022)

Feature Attribution

		$p(y\|\mathbf{x};\theta)$	y	c
x	the year 's best and most unpredictable comedy	0.91	pos	pos
x	we never feel anything for these characters	0.95	neg	neg
x	handsome but unfulfilling suspense drama	0.18	neg	pos

- **static** : One-off explanations do not allow increased interaction (Lakkaraju et al., 2022)
- **incomplete** : Missing context / narrative (Jacovi et al., 2022)

Feature Attribution

		$p(y\|\mathbf{x};\theta)$	y	c
x	the year 's best and most unpredictable comedy	0.91	pos	pos
x	we never feel anything for these characters	0.95	neg	neg
x	handsome but unfulfilling suspense drama	0.18	neg	pos

- **static** : One-off explanations do not allow increased interaction (Lakkaraju et al., 2022)
- **incomplete** : Missing context / narrative (Jacovi et al., 2022)
- **high cognitive load** : Humans select only most relevant causes for an event (Miller, 2019)

Feature Attribution

		$p(y\|\mathbf{x};\theta)$	y	c
x	the year 's best and most unpredictable comedy	0.91	pos	pos
x	we never feel anything for these characters	0.95	neg	neg
x	handsome but unfulfilling suspense drama	0.18	neg	pos

- **static** : One-off explanations do not allow increased interaction (Lakkaraju et al., 2022)
- **incomplete** : Missing context / narrative (Jacovi et al., 2022)
- **high cognitive load** : Humans select only most relevant causes for an event (Miller, 2019)
- **misaligned** : Laypeople might misinterpret the information (Schuff et al., 2022)

Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey. 2022.
Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective.
Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. Diagnosing AI explanation methods with folk concepts of behavior.
Tim Miller. Explanation in artificial intelligence: Insights from the social sciences.
Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. Human interpretation of saliency-based explanation over text.

Input $x :=$ **the year 's best and most unpredictable comedy**
$\Rightarrow$ **Model prediction** $y =$ **positive**

Method	**Question / User utterance**	**Explanation / Response**
Feature Attribution		

Extended from: Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey.

Input $x :=$ **the year 's best and most unpredictable comedy**
$\Rightarrow$ **Model prediction** $y =$ **positive**

Method	**Question / User utterance**	**Explanation / Response**
Feature Attribution	Which tokens are most important for the prediction?	

Extended from: Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey.

Input $x :=$ **the year 's best and most unpredictable comedy**
$\Rightarrow$ **Model prediction** $y =$ **positive**

Method	**Question / User utterance**	**Explanation / Response**
Feature Attribution	Which tokens are most important for the prediction?	*best* and *unpredictable* are most important.
Adversarial Examples	What would break the model's prediction?	

Extended from: Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey.

Input $x :=$ **the year 's best and most unpredictable comedy**
$\Rightarrow$ **Model prediction** $y =$ **positive**

Method	**Question / User utterance**	**Explanation / Response**
Feature Attribution	Which tokens are most important for the prediction?	*best* and *unpredictable* are most important.
Adversarial Examples	What would break the model's prediction?	Changing *best* to *finest* flips it to negative.
Influential Examples	What training examples influenced the prediction?	

Extended from: Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey.

Input $x :=$ **the year 's best and most unpredictable comedy**
$\Rightarrow$ **Model prediction** $y =$ **positive**

Method	**Question / User utterance**	**Explanation / Response**
Feature Attribution	Which tokens are most important for the prediction?	*best* and *unpredictable* are most important.
Adversarial Examples	What would break the model's prediction?	Changing *best* to *finest* flips it to negative.
Influential Examples	What training examples influenced the prediction?	*a delightfully unpredictable , hilarious comedy* is an influential instance also classified as positive.
Model Rationales	What would a generated natural language explanation be?	

Extended from: Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey.

Input $x :=$ **the year 's best and most unpredictable comedy**
$\Rightarrow$ **Model prediction** $y =$ **positive**

Method	**Question / User utterance**	**Explanation / Response**
Feature Attribution	Which tokens are most important for the prediction?	*best* and *unpredictable* are most important.
Adversarial Examples	What would break the model's prediction?	Changing *best* to *finest* flips it to negative.
Influential Examples	What training examples influenced the prediction?	*a delightfully unpredictable , hilarious comedy* is an influential instance also classified as positive.
Model Rationales	What would a generated natural language explanation be?	*Unpredictable comedies are funny*.

Extended from: Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural NLP: A survey.

Motivation

NLP + XAI

- **Interactivity** and framing the **explanation** process as a **dialogue** = hot topic in HCI!

  Tim Miller. Explanation in artificial intelligence: Insights from the social sciences.
  Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence.
  Q. Vera Liao and Kush R. Varshney. Human-centered explainable AI (XAI): from algorithms to user experiences.
  Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective.
  ...

Motivation

NLP + XAI

- **Interactivity** and framing the **explanation** process as a **dialogue** = hot topic in HCI!
  Tim Miller. Explanation in artificial intelligence: Insights from the social sciences.
  Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence.
  Q. Vera Liao and Kush R. Varshney. Human-centered explainable AI (XAI): from algorithms to user experiences.
  Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective.
  ...
- No prior work has picked this up for **NLP** problems / language models!

Motivation

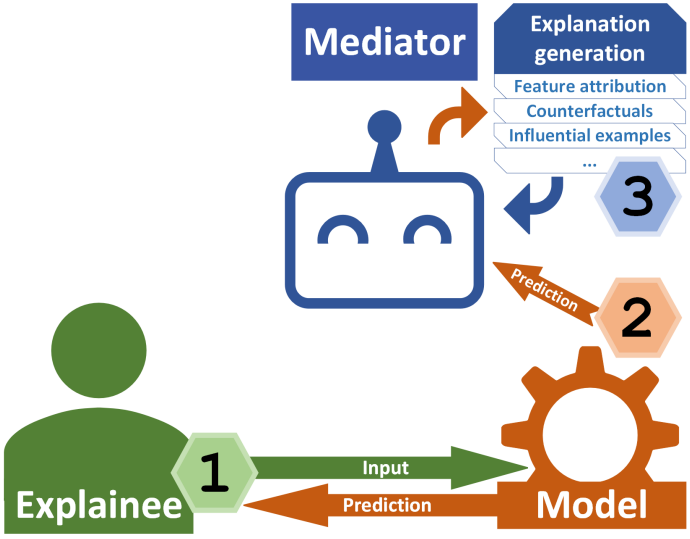
NLP + XAI

- **Interactivity** and framing the **explanation** process as a **dialogue** = hot topic in HCI!
  Tim Miller. Explanation in artificial intelligence: Insights from the social sciences.
  Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence.
  Q. Vera Liao and Kush R. Varshney. Human-centered explainable AI (XAI): from algorithms to user experiences.
  Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective.
  ...
- No prior work has picked this up for **NLP** problems / language models!

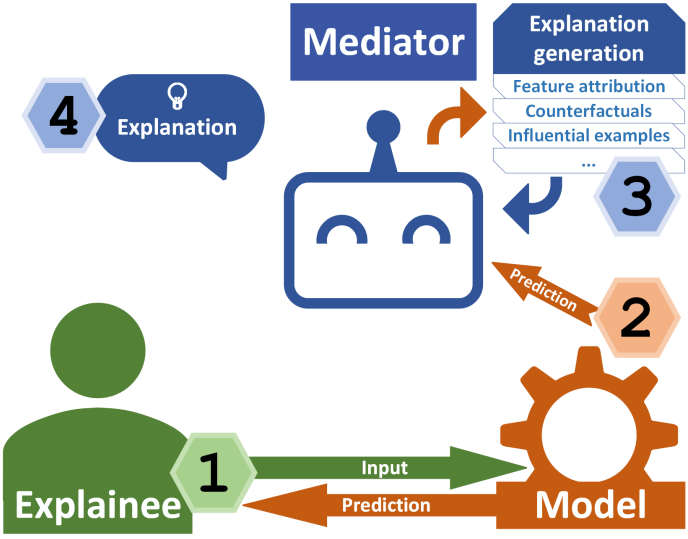
Key factors motivating the need for conversational agents

1. Flexibility of natural language
2. Complementary view of XAI methods
3. Atomic explanations alleviate cognitive load from the explainee

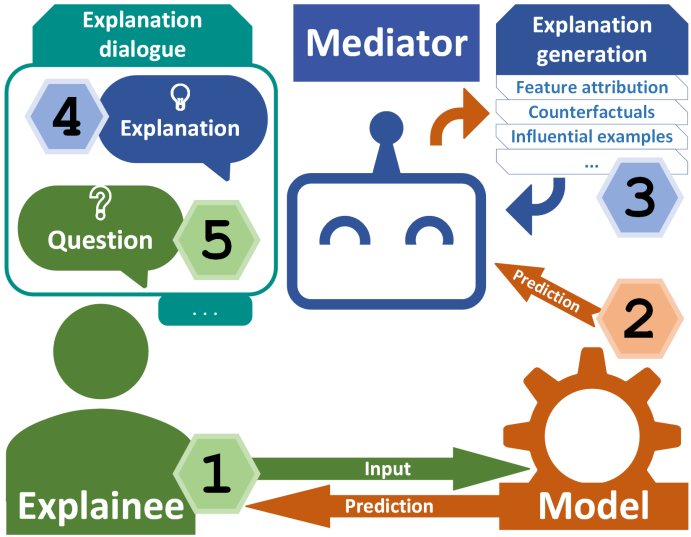
Mediators: Conversational Agents Explaining NLP Models



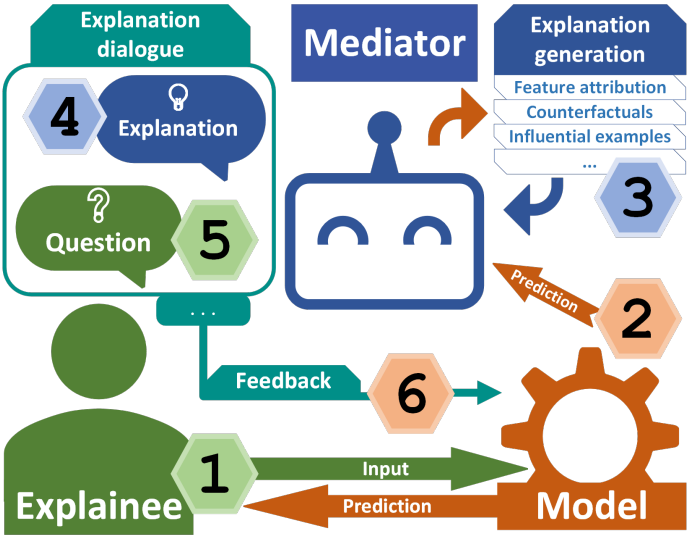
Mediators: Conversational Agents Explaining NLP Models



Mediators: Conversational Agents Explaining NLP Models



Mediators: Conversational Agents Explaining NLP Models



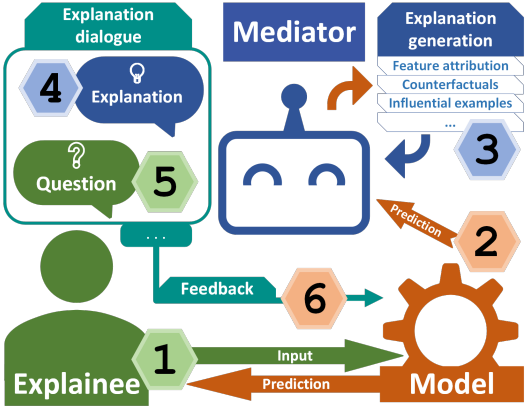
Follow-up questions to initial explanations

Category of follow-up question	Example questions
FQ1 : Input text edits	What if we removed word w from the input? What if the sentence s was in passive voice?
FQ2 : Scope restrictions	What change in the phrase p would flip the prediction? [AE] What is the most salient word in the n-th sentence? [FA]
FQ3 : Foil edits	What are the most salient tokens for class y' instead of y? [FA] What training example from class y' influenced the prediction the most? [IE]

Inspired by Weld & Bansal (2019)

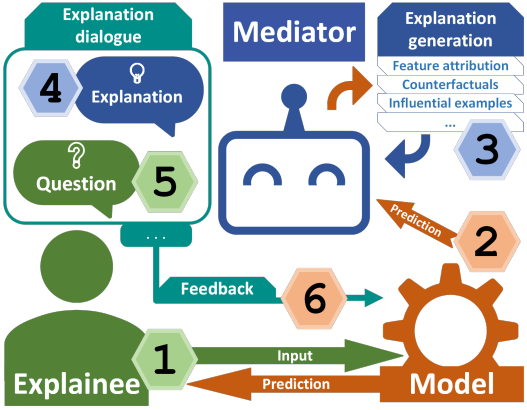
Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence.

Responsibilities of a Mediator



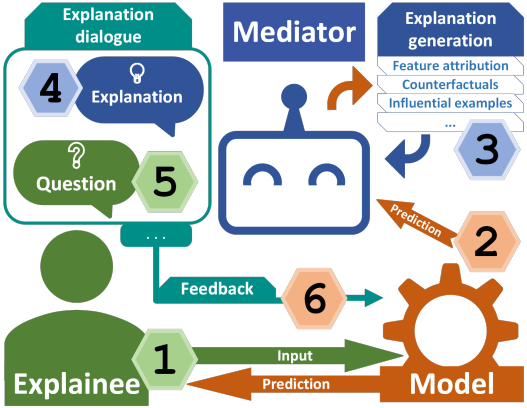
1. Generate bit-sized/atomic explanations

Responsibilities of a Mediator



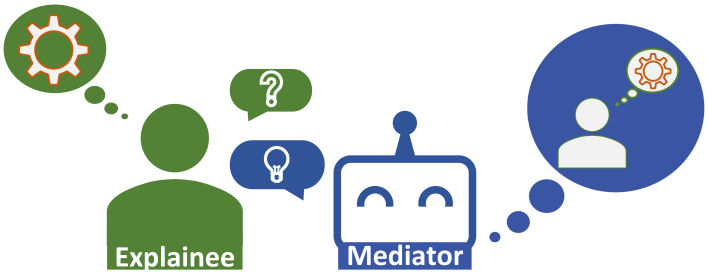
1. Generate bit-sized/atomic explanations
2. Respond to the explainee in NL

Responsibilities of a Mediator



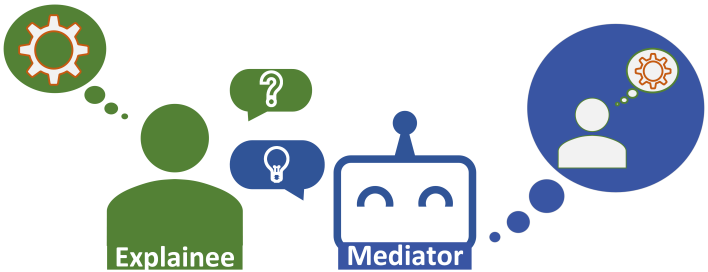
1. Generate bit-sized/atomic explanations
2. Respond to the explainee in NL
3. Understand a explainee's NL input

Mental model and user model



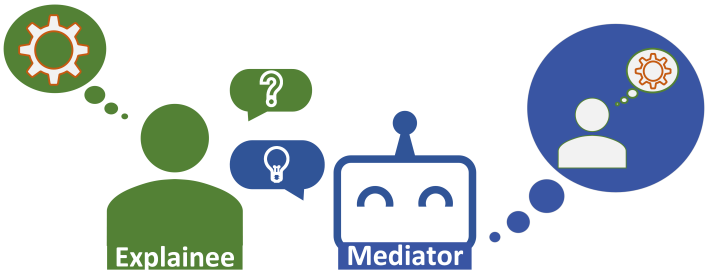
1. Generate bit-sized/atomic explanations
2. Respond to the explainee in NL
3. Understand a explainee's NL input
4. **Keep track of the conversation and user's knowledge**

Mental model and user model



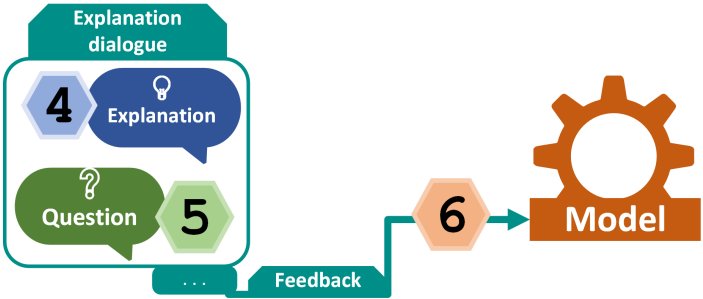
1. Generate bit-sized/atomic explanations
2. Respond to the explainee in NL
3. Understand a explainee's NL input
4. **Keep track of the conversation and user's knowledge**
    - Address misalignments between model and user expectation

Mental model and user model



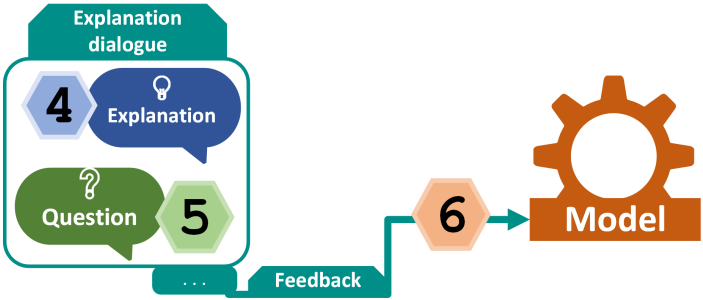
1. Generate bit-sized/atomic explanations
2. Respond to the explainee in NL
3. Understand a explainee's NL input
4. **Keep track of the conversation and user's knowledge**
   - Address misalignments between model and user expectation
   - Consider user expertise (laypeople vs. domain experts vs. model developers)

Beyond Conversations



- Evaluating explanation dialogues is tricky (always domain- and goal-dependent!), but can serve as corrective feedback to the explained model

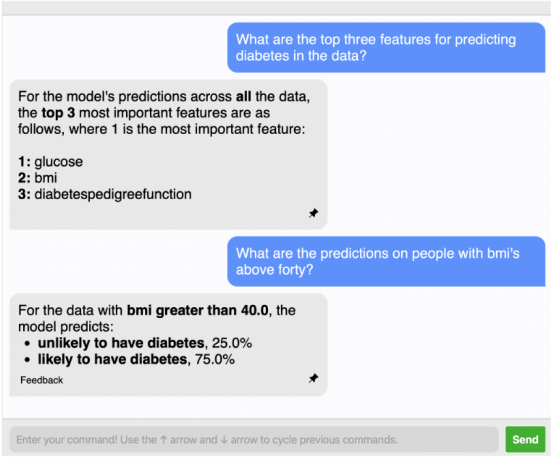
Beyond Conversations



- Evaluating explanation dialogues is tricky (always domain- and goal-dependent!), but can serve as corrective feedback to the explained model
- Training Mediators is an open question,
  because there are no explanation dialogue datasets as of yet
  (Similar: Information-seeking QA & Student-teacher dialogues)

Next steps: Implementation

Example use case in the medical domain from "TalkToModel" by Slack et al. (2022)



Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, Sameer Singh. TalkToModel: Understanding Machine Learning Models With Open Ended Dialogues.

Takeaways

- We engineered a **blueprint** of **Mediators** for the task of sentiment analysis, **conversational agents explaining the behavior of neural models** in an **interactive** fashion.

Takeaways

- We engineered a **blueprint** of **Mediators** for the task of sentiment analysis, **conversational agents explaining the behavior of neural models** in an **interactive** fashion.
- We summarized the desiderata that HCXAI research put forward for dialogue-based explanations and highlighted that the current state of research in NLP has yet to catch up and address the gaps and pitfalls.

Takeaways

- We engineered a **blueprint** of **Mediators** for the task of sentiment analysis, **conversational agents explaining the behavior of neural models** in an **interactive** fashion.
- We summarized the desiderata that HCXAI research put forward for dialogue-based explanations and highlighted that the current state of research in NLP has yet to catch up and address the gaps and pitfalls.
- We recommend employing selection processes in a **complementary view of explanations** and centering the dialogue around **user expectations** by keeping track of their mental models via **rigorous, continuous evaluation**.

Takeaways

- We engineered a **blueprint** of **Mediators** for the task of sentiment analysis, **conversational agents explaining the behavior of neural models** in an **interactive** fashion.
- We summarized the desiderata that HCXAI research put forward for dialogue-based explanations and highlighted that the current state of research in NLP has yet to catch up and address the gaps and pitfalls.
- We recommend employing selection processes in a **complementary view of explanations** and centering the dialogue around **user expectations** by keeping track of their mental models via **rigorous, continuous evaluation**.
- We hope that this inspires **data collection** and **practical applications** of Mediators for NLP model behavior.

Thanks to my co-authors Ajay Madhavan Ravichandran and Sebastian Möller,
the IJCAI-XAI reviewers and organizers
as well as Mareike Hartmann, Aljoscha Burchardt and Jan Nehring for their valuable feedback!

Mediators: Conversational Agents Explaining NLP Models

German Research Center for Artificial Intelligence