

Mediators: Conversational Agents Explaining NLP Model Behavior

Nils Feldhus¹, Ajay Madhavan Ravichandran^{1,2} and Sebastian Möller^{1,2}

¹German Research Center for Artificial Intelligence (DFKI)

²Technische Universität Berlin

{nils.feldhus, ajay_madhavan.ravichandran, sebastian.moeller}@dfki.de

Abstract

The human-centric explainable artificial intelligence (HCXAI) community has raised the need for framing the explanation process as a conversation between human and machine. In this position paper, we establish desiderata for Mediators, text-based conversational agents which are capable of explaining the behavior of neural models interactively using natural language. From the perspective of natural language processing (NLP) research, we engineer a blueprint of such a Mediator for the task of sentiment analysis and assess how far along current research is on the path towards dialogue-based explanations.

1 Introduction

In almost all areas of artificial intelligence, there is continuous evidence that neural models with an ever-growing number of parameters and training data are here to stay, thanks to scaling laws [Kaplan *et al.*, 2020]. Explaining the behavior of these large models has taken the center stage in many areas including NLP research [Madsen *et al.*, 2021]. Interactivity in explainable artificial intelligence (XAI) has been a hot topic for a while [Abdul *et al.*, 2018] and framing the explanation process as a dialogue between the human and the model has solid theoretical foundations in the HCXAI literature [Miller, 2019; Weld and Bansal, 2019; Liao and Varshney, 2021; Lakkaraju *et al.*, 2022; Dazeley *et al.*, 2021; Mariotti *et al.*, 2020], but no prior work has specified how to apply these frameworks to NLP problems and language models. Simultaneously, a big push from the NLP community towards implementing such systems has yet to occur.

In this paper, we highlight three key factors motivating the use of conversational agents to explain the behavior of NLP models (§2): The flexibility of natural language, the need for a complementary view for explainability methods, and the need to alleviate the cognitive load from the explainee.

Following [Sokol and Flach, 2020a] and [Lakkaraju *et al.*, 2022], we envision such conversational agents as complex, modular systems that we coin *Mediators*. Conceptually, a Mediator has to generate appropriate atomic explanations (§3), respond to the user in natural language (§4), understand a user’s natural language input (§5), and keep track of the

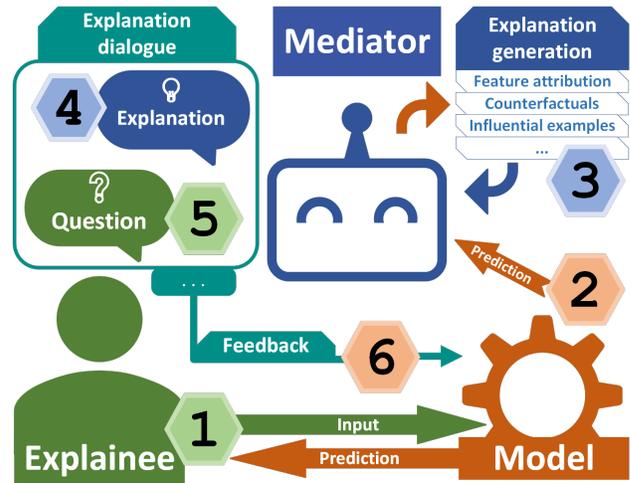


Figure 1: Simplified concept of a Mediator explaining the predictions of a Model to the human Explainee.

- ①: The Explainee provides input to the Model.
- ②: The Model puts out a prediction based on the input.
- ③: The Mediator generates explanations based on the prediction.
- ④: The Mediator holds an explanation dialogue with the Explainee.
- ⑤: The Explainee acts upon the explanation and asks follow-up questions. ④ and ⑤ are repeated until the Explainee is satisfied.
- ⑥: The explanation dialogue serves as corrective feedback which the Model can improve on.

conversation and the user’s knowledge (§6). These tasks are part of a process that we depict in Fig. 1. For the scope of this paper, we focus on purely textual setups and NLP models as the explanandum, but the modular structure of Mediators applies to other types of black boxes and multimodal interactive frameworks [Voigt *et al.*, 2021] as well.

We engineer a blueprint of a Mediator for NLP model behavior by taking its modules apart and devising examples for the downstream task of sentiment analysis. Simultaneously, we extend the seminal work of [Miller, 2019] and the recent works of [Nobani *et al.*, 2021] and [Lakkaraju *et al.*, 2022] by raising awareness of further aspects to consider such as evaluation and customization of the conversation as well as data collection (§7). We hope that this position paper aiming at NLP practitioners helps in closing the gap of conversational explainability.

2 Why we need Mediators

2.1 A complementary view of explainability methods

Natural language is said to be the most accessible and human-centric modality of explanation [Ehsan and Riedl, 2020]. Natural language explanations also exceed other explainability methods in plausibility, i.e. how convincing they are to the human explainee [Lei *et al.*, 2016; Lipton, 2018; Camburu *et al.*, 2018; Jacovi and Goldberg, 2020].

Natural language explanations alone, however, might cause unwarranted trust from the human recipient [Jacovi *et al.*, 2021], because the models which generate them usually are optimized via direct supervision towards a few human-acceptable gold rationales from static datasets [Camburu *et al.*, 2018]. Other explainability methods excel in faithfulness (sometimes called fidelity) [Jacovi and Goldberg, 2020], measuring how truthful an explanation reflects the internal representations of the explained model. In particular, [Wiegrefe *et al.*, 2021] found that faithful explanations are not achievable with free-text rationalization models and truly faithful explanations might be impossible, after all [Jacovi and Goldberg, 2020].

That is why a complementary view of explainability methods (for NLP) [Madsen *et al.*, 2021; Jacovi *et al.*, 2022] is necessary, where faithful and plausible [Jacovi and Goldberg, 2020; Herman, 2017; Gilpin *et al.*, 2018] explanations with a varying scope can both be presented to a user. This has previously been advocated for by works in HCXAI such as [Hohman *et al.*, 2019] and [Yeung *et al.*, 2020] among others. Since people are better at understanding narratives rather than numbers and probabilities, [Reiter, 2019] proposed to present the reasoning done by a numerical non-symbolic model (such as a large neural model) as a narrative including causal and argumentative relations.

2.2 Cognitive load

However, next to explanations being faithful and plausible, we also detect a need for them to be both sufficient and concise: [DeYoung *et al.*, 2020] define sufficiency as a measure evaluating if the amount of information communicated through an explanation is enough to justify some model predictions. While it might be tempting to densely pack as many cognitive chunks [Doshi-Velez and Kim, 2017] into an explanation, it may not be a good explanation in terms of overall satisfaction [Lombrozo, 2007], understandability [Ehsan *et al.*, 2019] and usefulness [Bansal *et al.*, 2021].

The natural process for human explainers is to select only the most relevant causes for an event. Although an event’s causal chain often is much longer, humans have to make conscious decisions about what to say, since the cognitive load on the recipient’s end might become too big otherwise [Hilton, 2017; Miller, 2019; Doshi-Velez and Kim, 2017]. Systems implemented in related work are prone to overwhelm users with too much information at once [Kulesza *et al.*, 2015]. Natural language can help here, because it is flexible, i.e. it can be tuned to custom amounts of cognitive chunks and can be adapted to both different audiences and tasks.

2.3 The case for dialogue-based explanations

The aforementioned cognitive limits motivate a setup where concise explanations are presented to the user at each turn prompting them to ask follow-up questions until their needs for information are satisfied [Leake, 1991], thus engaging in back-and-forth conversations [Weld and Bansal, 2019; Liao and Varshney, 2021; Madumal *et al.*, 2019; Hartmann *et al.*, 2021; Sevastjanova *et al.*, 2018]. These concise explanations should be “atomic” in nature and contain only a few cognitive chunks to not overwhelm the user.

User satisfaction has also been shown to increase under the presence of an option to provide feedback. In particular, [Smith-Renner *et al.*, 2020] concluded that it reduces user frustration and leads them to understand the model in the background better. The findings of [Lakkaraju *et al.*, 2022] indicate that decision-makers would strongly prefer interactions to take the form of natural language dialogues, in order to treat machine learning models as “another colleague” who can be held accountable by asking why they made a particular decision through expressive and accessible natural language interactions.

In summary, we argue that by drawing from many different types of explainability methods and presenting atomic explanations on the basis of cognitive chunks, conversational agents with a modular setup like Mediators can be employed to interact with users in a natural language dialogue.

3 Generating atomic explanations

Explanations should not be static As [Hohman *et al.*, 2019] and [Lakkaraju *et al.*, 2022] pointed out, most of the existing work on explainability focuses on one-off, static explanations, e.g. feature attribution methods producing a single saliency map¹ that conveys a limited amount of information and nothing about the causality involved. Although the explainable NLP community has come up with text- and task-specific methods in the last few years [Madsen *et al.*, 2021], there is still no apparent push towards conversation-based explanations of NLP model behavior.

Explanation generation as a selection process [Miller, 2019] looked into the processes of how people select explanations from available causes, by following common heuristics such as abnormality, intentionality, necessity, sufficiency, and robustness. This is underlined by the notion that there is no single best explanation [Ribera and Lapedriza, 2019]. By framing explanation generation as a search or exploration problem, we can incorporate many factors, including those correlating with human preferences, to find an optimal candidate. This refers to both (1) selecting the best explanation among explanations of the same type, e.g. counterfactuals, and (2) selecting the best explanation type, e.g. a saliency map vs. a counterfactual.

¹In NLP, these explanations are commonly produced with libraries such as AllenNLP Interpret [Wallace *et al.*, 2019] and Captum [Kokhlikyan *et al.*, 2020] which can easily be applied to the wide array of pre-trained language models contributed through, e.g., Hugging Face transformers [Wolf *et al.*, 2020], as demonstrated in Thermostat [Feldhus *et al.*, 2021].

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

Method	Question / User utterance	Explanation / Response from Mediator
Feature Attribution	Which tokens are most important for the prediction?	<u>best</u> and <u>unpredictable</u> are most important.
Adversarial Examples	What would break the model’s prediction?	Changing <u>best</u> to <u>finest</u> flips it to <u>negative</u> .
Influential Examples	What training examples influenced the prediction?	<u>a delightfully unpredictable , hilarious comedy</u> is an influential instance also classified as positive.
Counterfactuals	What does the model consider a valid opposite example?	Changing <u>best</u> to <u>worst</u> and <u>unpredictable</u> to <u>predictable</u> creates a valid <u>negative</u> instance.
Model Rationales	What would a generated natural language explanation be?	<u>Unpredictable comedies are funny.</u>

Table 1: Example instance x from the SST dataset [Socher *et al.*, 2013] that an explainees uses as input to a language model tackling the task of sentiment analysis (distinguishing positive from negative movie reviews). The table depicts the taxonomy of explainability methods and their associated questions as they appear in [Madsen *et al.*, 2021]. (Please consult their paper for more information on available methods.) In the right column, we added explanation candidates that could serve as answers presented to the explainees by the Mediator. The underlined parts are the actual output of the respective method, while the rest is generated through verbalization.

The way explanations are commonly employed may not be sufficient for exploration and continuous discovery from users that have a range of skills and expertise. In this regard, [Sokol and Flach, 2020a] highlight that explanations have properties such as context, scope and breadth that may have to be personalized when generating explanations. Being able to navigate between different kinds of explanations was also advocated for by [Bove *et al.*, 2022] who developed an interface for contextualizing feature attribution explanations. We argue that such an exploration can also be modeled as a conversation. One specific implementation of a search-based explanation generation is the work of [Wiegrefe *et al.*, 2022] whose framework over-generates explanation candidates using a language model and subsequently filters them using a second language model that is trained on human acceptability ratings collected in a crowd study. [Hase *et al.*, 2021] devised methods for searching through the space of possible explanations as an alternative to existing feature attribution methods. [Treviso and Martins, 2020] framed explainability as a communication problem between an explainer and a layperson about a model’s decision and empirically assessed the quality of the explanation.

We argue that the methods mentioned and categorized in [Jacovi *et al.*, 2022] and [Madsen *et al.*, 2021] could be used in a complementary view of explainability for neural models: Feature attribution, counterfactuals, influence functions, and model-generated rationales often occur separately, but a Mediator has to be able to draw upon this pool of explanations generated from different methods dynamically based on user needs.² In Tab. 1, we showcase the example use case of sentiment analysis, a binary text classification task, and devise questions associated to explainability methods from [Madsen *et al.*, 2021].³ Feature attribution explanations and

model rationales can serve as initial explanations letting explainees form their hypotheses about the model’s behavior, while counterfactuals and adversarial examples can be sanity checks that support or counter the hypotheses [Hohman *et al.*, 2019].

Sentiment analysis, like most text classification tasks, is a trivial task in terms of explanation generation. However, since other, more challenging tasks are less explored for most types of explainability methods, we find it fitting for the scope of our work. We identify methods for explaining text generation or, more generally, language modeling [Vafa *et al.*, 2021; Yin and Neubig, 2022] as a very promising avenue for future research. This opens up the pathway towards real-world use cases such as question answering and machine translation.

4 Responding to the user in natural language

After collecting sufficiently many explanations for a Mediator to choose from, the next hurdle is to verbalize and present them in a way that engages the user to start, continue and finish a conversation.⁴ We understand explanations to take the form of information-seeking dialogues [Walton and Krabbe, 1995], where the user seeks the answers to some questions and the Mediator knows and provides them. In the case of explanation dialogues, we identify a mixed initiative setting. The start of the conversation can either be triggered by the Mediator presenting a concise explanation that prompts the user to interact with it, or by the human who already has a clear goal, e.g. an explanation type, in mind. Separating the explanation content planning from the execution of the dialogue has been proposed as early as [Cawsey, 1991]. The Mediator should respond with informative and properly contextualized explanations for why the underlying model made specific decisions [Lakkaraju *et al.*, 2022]. According to the maxims of relation and quantity, it is essential to only relay information that is relevant and necessary at any given point in time. It means that a Mediator has to “know” what the user knows and expects (§7) before determining

²This task is depicted as Step 3 in Fig. 1.

³We identify verbalizing explanations, i.e. translating them into natural language, such as feature attribution and adversarial examples as a missing component. Existing solutions include the works of [Forrest *et al.*, 2018] and [Yao *et al.*, 2021].

⁴This task is depicted as Step 4 in Fig. 1.

Category of follow-up question	Example questions
Input text edits	What if we removed word w from the input? What if we added the phrase p at the end? What if the sentence s was in passive voice?
Scope restrictions	What change in the phrase p would flip the prediction? [AE] How does word w need to be changed in order to flip the prediction? [AE] What is the most salient word in the n -th sentence? [FA]
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]
Explanation source edits	Could you show me the LIME instead of the Shapley Values explanation? [FA] Could you show me the Integrated Gradients explanation with 50 samples? [FA]

Table 2: Categories and examples of follow-up questions beyond the five generic questions in Tab. 1 that can trivially be mapped to an explanation type. Associated explanation type (FA: Feature Attribution, AE: Adversarial Examples, IE: Influential Examples) in square brackets (none means the question is applicable to any type). All except “explanation source edits” were already proposed in [Weld and Bansal, 2019].

the content of the explanation [Sokol and Flach, 2020b; Hartmann *et al.*, 2021]. At the same time, in accordance with the notion of parsimony [Sokol and Flach, 2020b], Mediators should aim to fill in the most gaps [Leake, 1991] with the fewest statements. This ties back in with the issue of cognitive load: Although the frame of a conversation alleviates this issue on a high level, every single response to the user should be selected with user knowledge and expectations in mind.

[Akula *et al.*, 2019] provided answers to user questions using different reasoning paradigms in a visual setting. [Madumal *et al.*, 2019] developed a framework that allows users to follow up on an explanation to reach a comprehensive model understanding.

We urge researchers to investigate different approaches regarding explanation selection and generation: While we argue for a pool of explanations produced by various methods that the Mediator can draw from (§3), one might also train a framework in an end-to-end fashion combining selection, generation and responding.

5 Understanding a user’s natural language input

A Mediator should understand continuous requests for explanations and be able to efficiently map these to appropriate explanation types to generate [Lakkaraju *et al.*, 2022]. This is commonly understood as an intent recognition problem in dialogue and would be realized as one module in a Mediator framework.⁵ [Lakkaraju *et al.*, 2022] identified natural language understanding as a problem, because there is a large set of possible query types, many different ways to phrase ex-

⁵This corresponds to Step 5 in Fig. 1.

plainability questions, and transferring it to different downstream applications is generally very complex. [Weld and Bansal, 2019] outlined the types of follow-up and drill-down actions a user might request upon seeing some initial explanation.

For this module that is tasked with recognizing the user intent, the canvas of [Lim and Dey, 2009] – later expanded upon by [Weld and Bansal, 2019] and [Liao *et al.*, 2021] – can serve as a general-purpose mapping between user utterance and explanation type. We found [Rebanal *et al.*, 2021] to be the only work that concretely built a classifier for user questions. Based on the classification, they generated explanations for the learned representations of the underlying model. We see the taxonomy of [Madsen *et al.*, 2021] (Tab. 1) as a starting point in adapting such a mapping to NLP settings and propose four distinct types of follow-up questions in Tab. 2: Input text edits, scope restrictions and foil edits which were proposed in [Weld and Bansal, 2019] as well as explanation source edits which is more targeted at model developers and ML experts rather than laypeople. The challenge then is to tie the intent recognition module to the search method that identifies potential explanation candidates and optionally combine them to end up with the final answer presented to the explainee. In some cases, e.g. “What if the sentence s was in passive voice?”, the module has to have a thorough understanding of language and perform edits accordingly.

6 Keeping track of the dialogue

When designing dialogue systems, the task of keeping track of the dialogue history is essential to better inform the selection of the next action or response. This is traditionally done by predicting slot-value pairs representing the user’s goals while accounting for the dialogue history at any turn. We point the reader towards recent methods in dialogue state tracking [Balaraman *et al.*, 2021; Hu *et al.*, 2022] and dialogue systems for information acquisition [Cai *et al.*, 2022]. All of them are disconnected from the explainability literature, however.



Figure 2: Depiction of the Explainee’s mental model of the explained Model and the Mediator’s model of the Explainee’s knowledge.

User models and mental models The explainability literature, on the other hand, has explored user models to keep track of the recipient’s knowledge. The idea of a user model has been explored in Cawsey’s EDGE system [1991]: The knowledge that the user has about a phenomenon and their level of expertise should both be updated during the dialogue

[Miller, 2019]. [Stumpf *et al.*, 2009] allowed users to interact with different types of explanations and examined if these enable them to form useful mental models of the system. In a follow-up work, [Kulesza *et al.*, 2015] presented explanatory debugging as a use case to address this question about mental models, allowing users to communicate corrections back to the system. [Weld and Bansal, 2019] pointed out that constructing user models, i.e. tracking explicitly what users know and expect, are typically based on hand-engineered solutions. In the remainder of this section, we will highlight different aspects related to user models and mental models. We illustrate this relation in Fig. 2.

Addressing misalignments between model and user expectation Model predictions, the output of explainability methods and user expectations are often misaligned [Schuff *et al.*, 2022]. This leads us to the analysis of the users’ mental models: Keeping track of the dialogue also means estimating the users’ understanding of the underlying model’s behavior, e.g. by using a formal argumentative dialogue framework [Madumal *et al.*, 2019; Sokol and Flach, 2020a] or by using simulatability [Doshi-Velez and Kim, 2017; Hase and Bansal, 2020] tests prompting them to simulate the model on unseen data. In practice, the latter would usually occur in a separate evaluation stage after users have interacted with the model, but a rigorous evaluation would require a frequent assessment of the user’s understanding.

Another avenue relevant to these misalignments is the one of active learning [Ghai *et al.*, 2020]. For example, the workflow by [Liang *et al.*, 2020] presents most confusing class pairs to human experts and queries them for explanations. Such a system can then utilize this new knowledge to improve the underlying classification model. Hence, an improved underlying model can also mitigate the mismatch between user and model expectations. Two further recent works dealing with learning from natural language explanations have specified empirically which type of explanation data helps models the most in terms of task performance [Hase and Bansal, 2022] and which strategies for training on such data are preferable [Carton *et al.*, 2022].

User expertise [Ehsan *et al.*, 2021] examined and discussed the effect of explanations on groups of users with different AI backgrounds. Elaborate Mediator designs need to take into account that there might not be a one-for-all solution [Sokol and Flach, 2020a] even when a user’s mental model is considered [Chromik *et al.*, 2021]. We think the field will eventually catch up more with targeting laypeople as recipients for explanations, which would make this aspect even more relevant than it already is.

Reacting to user feedback [Miller, 2019] posed the question: “what should an explanatory agent do if the explainee does not accept a selected explanation?” We argue that all user feedback should be considered as training signals for (a) the underlying (explained) model, (b) the explanation-generating model, and (c) the user model that keeps track of user knowledge and the dialogue. Thus, most kind of user feedback to Mediators is *rich feedback* which involves “more expressive forms of corrective feedback which can cause a deeper change in the underlying machine learning algorithm”

[Stumpf *et al.*, 2009]. We attenuate this slightly, because costly model edits or retraining may impede the user experience. It might even degrade the model performance due to misconceptions and biases of the explainee. In addition, incomplete statements such as “The model made a mistake.” or “This explanation is wrong.” should not be taken into account. The advisory dialogues of [Moore and Paris, 1993] explicitly modeled the effect of utterances on the recipient’s mental state allowing for a recovery mechanism from failure and misunderstanding [Miller, 2019].⁶

Finishing an explanation [Miller, 2019] posed the question of “how do we know that an explanation has ‘finished’?” which can involve knowing whether the explainee has correctly understood the explanation. [Alvarez-Melis *et al.*, 2019] proposed a Weight-of-Evidence metric measuring the effect of “explaining away” different outcomes. From a theoretic perspective, it might suffice to exhaust all alternatives, but cannot be guaranteed to fulfill the user’s understanding in practice. Therefore, we highlight this as a potential roadblock.

7 Beyond conversations

Beyond this modular setup we propose for modeling Mediators, we raise awareness of three further aspects to consider: How to evaluate explanation dialogues, when to allow customization of the explanation process, and how to train Mediators.

7.1 Evaluating explanations

In the HCXAI literature, previous works have employed evaluation measures beyond the estimation of the user’s mental model, such as usefulness and satisfaction. This is analogous to the task of natural language generation, in that can show explanations to human subjects and ask them to rate and comment on them in various ways [Reiter, 2019]. [Wiegreffe and Marasović, 2021] identified two paradigms in this regard: Collect-and-Judge and Collect-and-Edit. While the former is about letting crowdworkers assess the quality of the (automatically or human-annotated) collected explanations, the latter necessitates annotators to edit the explanations to reduce annotation artefacts and biases and to improve quality control and linguistic variety. This is echoed by [Arora *et al.*, 2022] who proposed explanation evaluation using iterative editing. [González *et al.*, 2021] evaluated explanations for reading comprehension models and showed that introducing multiple models of various quality and adversarial examples (which can be seen as another complementary type of explanations) can help to account for belief bias effects in human evaluation.

Moreover, we point the reader towards measurements for user trust and reliance on explanations and human-AI task performance [Mohseni *et al.*, 2021] as well as automated metrics for evaluating faithfulness [Hase *et al.*, 2020].

Dialogue evaluation research has raised awareness of measuring flexibility and understanding among many other criteria [Mehri *et al.*, 2022]. There exist automated metrics

⁶Fig. 1 depicts this feedback with Step 6.

based on NLP models for assessing the quality of dialogues, but their correlation with human judgments needs to be improved on. These lines of research are disconnected from each other, which makes the task of evaluating explanation dialogues very challenging. We argue that work on metrics specifically designed for evaluating explanation dialogues is just as important as implementational work.

7.2 Customizing the explanation process

For interactive explanations, [Sokol and Flach, 2020b] raised the need for controllability and customizability to suit a user’s needs, e.g., through adjustable granularity. The complementary view of XAI methods proposed for Mediators allows for enhanced customizations of the dialogue and explanation narrative.

However, this sometimes requires offering additional settings adjustable via user interfaces, because they might not be easily communicated via natural language. Nevertheless, opening up the Pandora’s box of customization options ends up being a question of scalability, i.e. guaranteeing real-time responses [Miller, 2019].

7.3 How to train Mediators

The most apparent open question about Mediators is how to train them. In the following, we will analyze existing datasets, give recommendations about future data collection and how previous frameworks utilized explanations for training models.

There is a distinct lack of natural language explanation datasets [Wiegrefe and Marasović, 2021] covering explanation dialogues, complicating the transfer to domain-specific use cases. [Attari *et al.*, 2019] presented a study on how to collect data from human-human dialogues that can be used to train systems akin to Mediators. [Madumal *et al.*, 2019] analyzed explanation dialogue transcripts and identified key components of such interactions. [Weitz *et al.*, 2021] analyzed how mental models are formed by users playing a collaborative puzzle game including an explanatory dialogue system. We recommend going for this type of data if the Mediator’s main responsibility is to hold dialogues that are as natural as possible and to understand questions that might be outside the scope that we presented in Tab. 1 and Tab. 2.

Another close comparison are information-seeking dialogue or question answering datasets [Choi *et al.*, 2018; Saeidi *et al.*, 2018; Penha *et al.*, 2019; Qu *et al.*, 2018; Feng *et al.*, 2020; Dai *et al.*, 2022] which already are known to be notoriously hard to create [Rogers *et al.*, 2022]. We find, however, that such datasets are more concerned with covering a range of wh-questions instead of explanations (“Why...?”) [Wu *et al.*, 2022].

Based on our findings while reviewing the aforementioned datasets, we propose desiderata for explanation dialogue datasets: At the minimum, they should include user utterances and feedback for post-dialogue user assessment. However, to navigate the narrow path towards such richly annotated data, practitioners might have to invest a lot of effort in terms of implementation, costs and time to generate atomic explanations that humans can give feedback on, e.g. via Likert-scale ratings.

To which degree information-seeking question answering or human-human dialogue datasets can serve as training data for Mediators, depends on the kind of use case and goal. For first test runs, they should be aware of the implementational effort and lack of control over the Mediator’s output and favor Wizard-of-Oz studies [Sokol and Flach, 2020a]. We point the reader towards investigative work on the role of explanation data for training NLP models [Hase and Bansal, 2022; Hartmann and Sonntag, 2022].

8 Related Work

While [Miller, 2019] is the main work advocating for conversational explanations, the recent work of [Lakkaraju *et al.*, 2022] supported these theoretical foundations with a study where they interviewed domain experts about their needs and desires for such explanations. We see [Nobani *et al.*, 2021] as the (proposed) framework that is closest to a Mediator. However, they have yet to present empirical evidence and tie it to an actual use case. Our work is more comprehensive in comparison and connects the dots between the communities of HCXAI and NLP research.

Regarding Mediators in practice, we draw a connection towards applications which allow users to explore NLP models interactively [Tenney *et al.*, 2020; Strobelt *et al.*, 2021; Strobelt *et al.*, 2022; Lee *et al.*, 2022; Perez *et al.*, 2022]. Although not all of their functional features might translate to conversational setups, we expect a dialogue-based interaction on NLP use cases (e.g., next-word prediction, summarization, story generation, question answering) to elicit more useful insights for all parties involved: The user is more engaged in a dialogue with conversational agents and the model can be trained on more elaborate responses.

9 Conclusion

We engineered a blueprint of Mediators, 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 recommended employing search methods in a complementary view of explanations and focussing on user expectations by keeping track of their mental models via rigorous, continuous evaluation. We hope that this position paper inspires data collection and implementational work in Mediators for model behavior.

Acknowledgments

We would like to thank Mareike Hartmann for extensive and fruitful discussions and Jan Nehring and Aljoscha Burchardt as well as the anonymous reviewers at the IJCAI 2022 Workshop on Explainable Artificial Intelligence for their valuable feedback. This work has been supported by the German Federal Ministry of Education and Research as part of the project XAINES (01IW20005).

References

[Abdul *et al.*, 2018] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends

- and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–18, New York, NY, USA, 2018. Association for Computing Machinery.
- [Akula *et al.*, 2019] Arjun R Akula, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. Natural language interaction with explainable AI models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [Alvarez-Melis *et al.*, 2019] David Alvarez-Melis, Hal Daumé III, Jennifer Wortman Vaughan, and Hanna M. Wallach. Weight of evidence as a basis for human-oriented explanations. In *Human-Centric Machine Learning (HCML) Workshop @ NeurIPS 2019*, volume abs/1910.13503, 2019.
- [Arora *et al.*, 2022] Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Thirty-Six AAAI Conference on Artificial Intelligence.*, February 2022.
- [Attari *et al.*, 2019] Nazia Attari, Martin Heckmann, and David Schlangen. From explainability to explanation: Using a dialogue setting to elicit annotations with justifications. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 331–335, Stockholm, Sweden, September 2019. Association for Computational Linguistics.
- [Balaraman *et al.*, 2021] Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 239–251, Singapore and Online, July 2021. Association for Computational Linguistics.
- [Bansal *et al.*, 2021] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.
- [Bove *et al.*, 2022] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 807–819, New York, NY, USA, 2022. Association for Computing Machinery.
- [Cai *et al.*, 2022] Pengshan Cai, Hui Wan, Fei Liu, Mo Yu, Hong Yu, and Sachindra Joshi. Learning as conversation: Dialogue systems reinforced for information acquisition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, United States, April 2022. Association for Computational Linguistics.
- [Camburu *et al.*, 2018] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [Carton *et al.*, 2022] Samuel Carton, Surya Kanoria, and Chenhao Tan. What to learn, and how: Toward effective learning from rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [Cawsey, 1991] Alison Cawsey. Generating interactive explanations. In *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 1*, AAAI'91, page 86–91. AAAI Press, 1991.
- [Choi *et al.*, 2018] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [Chromik *et al.*, 2021] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. I think I get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 307–317, New York, NY, USA, 2021. Association for Computing Machinery.
- [Dai *et al.*, 2022] Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Mike Green, Qazi Rashid, and Kelvin Guu. Dialog inpainting: Turning documents to dialogs. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.
- [Dazeley *et al.*, 2021] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525, 2021.
- [DeYoung *et al.*, 2020] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics.
- [Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017.
- [Ehsan and Riedl, 2020] Upol Ehsan and Mark O. Riedl. Human-centered explainable AI: Towards a reflective

- sociotechnical approach. In Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones, editors, *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, pages 449–466, Cham, 2020. Springer International Publishing.
- [Ehsan *et al.*, 2019] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. Automated rationale generation: A technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, page 263–274, New York, NY, USA, 2019. Association for Computing Machinery.
- [Ehsan *et al.*, 2021] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. The who in explainable AI: how AI background shapes perceptions of AI explanations. *CoRR*, abs/2107.13509, 2021.
- [Feldhus *et al.*, 2021] Nils Feldhus, Robert Schwarzenberg, and Sebastian Möller. Thermostat: A large collection of NLP model explanations and analysis tools. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 87–95, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [Feng *et al.*, 2020] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online, November 2020. Association for Computational Linguistics.
- [Forrest *et al.*, 2018] James Forrest, Somayajulu Sripada, Wei Pang, and George Coghill. Towards making NLG a voice for interpretable machine learning. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 177–182, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics.
- [Ghai *et al.*, 2020] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Klaus Mueller. Explainable active learning (XAL): an empirical study of how local explanations impact annotator experience. In *CSCW 2020: ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2020.
- [Gilpin *et al.*, 2018] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.
- [González *et al.*, 2021] Ana Valeria González, Anna Rogers, and Anders Søgaard. On the interaction of belief bias and explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942, Online, August 2021. Association for Computational Linguistics.
- [Hartmann and Sonntag, 2022] Mareike Hartmann and Daniel Sonntag. A survey on improving NLP models with human explanations. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 40–47, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [Hartmann *et al.*, 2021] Mareike Hartmann, Ivana Kruijff-Korbayová, and Daniel Sonntag. Interaction with explanations in the xaines project. In *Trustworthy AI in the Wild Workshop 2021*, 9 2021.
- [Hase and Bansal, 2020] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics.
- [Hase and Bansal, 2022] Peter Hase and Mohit Bansal. When can models learn from explanations? A formal framework for understanding the roles of explanation data. In *ACL 2022 Workshop on Natural Language Supervision*, 2022.
- [Hase *et al.*, 2020] Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online, November 2020. Association for Computational Linguistics.
- [Hase *et al.*, 2021] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [Herman, 2017] Bernease Herman. The promise and peril of human evaluation for model interpretability. *Interpretable ML Symposium at the 31st Conference on Neural Information Processing Systems*, 2017.
- [Hilton, 2017] Denis Hilton. Social attribution and explanation. *The Oxford Handbook of Causal Reasoning*, 2017.
- [Hohman *et al.*, 2019] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.
- [Hu *et al.*, 2022] Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. *CoRR*, 2022.
- [Jacovi and Goldberg, 2020] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

- [Jacovi *et al.*, 2021] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA, 2021. Association for Computing Machinery.
- [Jacovi *et al.*, 2022] Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. Diagnosing AI explanation methods with folk concepts of behavior. *CoRR*, abs/2201.11239, 2022.
- [Kaplan *et al.*, 2020] Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.
- [Kokhlikyan *et al.*, 2020] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.
- [Kulesza *et al.*, 2015] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, page 126–137, New York, NY, USA, 2015. Association for Computing Machinery.
- [Lakkaraju *et al.*, 2022] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner’s perspective. *CoRR*, abs/2202.01875, 2022.
- [Leake, 1991] David B. Leake. Goal-based explanation evaluation. *Cognitive Science*, 15(4):509–545, 1991.
- [Lee *et al.*, 2022] Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.
- [Lei *et al.*, 2016] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics.
- [Liang *et al.*, 2020] Weixin Liang, James Zou, and Zhou Yu. ALICE: Active learning with contrastive natural language explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4380–4391, Online, November 2020. Association for Computational Linguistics.
- [Liao and Varshney, 2021] Q. Vera Liao and Kush R. Varshney. Human-centered explainable AI (XAI): from algorithms to user experiences. *CoRR*, abs/2110.10790, 2021.
- [Liao *et al.*, 2021] Q. Vera Liao, Milena Pribic, Jaesik Han, Sarah Miller, and Daby Sow. Question-driven design process for explainable AI user experiences. *CoRR*, abs/2104.03483, 2021.
- [Lim and Dey, 2009] Brian Y. Lim and Anind K. Dey. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, page 195–204, New York, NY, USA, 2009. Association for Computing Machinery.
- [Lipton, 2018] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, jun 2018.
- [Lombrozo, 2007] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257, 2007.
- [Madsen *et al.*, 2021] Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ArXiv*, abs/2108.04840, 2021.
- [Madumal *et al.*, 2019] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1033–1041, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.
- [Mariotti *et al.*, 2020] Ettore Mariotti, Jose M. Alonso, and Albert Gatt. Towards harnessing natural language generation to explain black-box models. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 22–27, Dublin, Ireland, November 2020. Association for Computational Linguistics.
- [Mehri *et al.*, 2022] Shikib Mehri, Jinho Choi, L. F. D’Haro, Jan Deriu, Maxine Eskénazi, Milica Gasic, Kallirroi Georgila, Dilek Z. Hakkani-Tür, Zekang Li, Verena Rieser, Samira Shaikh, David R. Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *ArXiv*, abs/2203.10012, 2022.
- [Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
- [Mohseni *et al.*, 2021] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), aug 2021.
- [Moore and Paris, 1993] Johanna D. Moore and Cecile L. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694, 1993.

- [Nobani *et al.*, 2021] Navid Nobani, Fabio Mercorio, and Mario Mezzanzanica. Towards an explainer-agnostic conversational xai. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4909–4910. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Doctoral Consortium.
- [Penha *et al.*, 2019] Gustavo Penha, Alexandru Balan, and Claudia Hauff. Introducing mantis: a novel multi-domain information seeking dialogues dataset. *CoRR*, abs/1912.04639, 2019.
- [Perez *et al.*, 2022] Ethan Perez, Saffron Huang, H. Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *CoRR*, abs/2202.03286, 2022.
- [Qu *et al.*, 2018] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, page 989–992, New York, NY, USA, 2018. Association for Computing Machinery.
- [Rebanal *et al.*, 2021] Juan Rebanal, Jordan Combitsis, Yuqi Tang, and Xiang 'Anthony' Chen. Xalgo: A design probe of explaining algorithms' internal states via question-answering. In *26th International Conference on Intelligent User Interfaces, IUI '21*, page 329–339, New York, NY, USA, 2021. Association for Computing Machinery.
- [Reiter, 2019] Ehud Reiter. Natural language generation challenges for explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NLXAI 2019)*, pages 3–7. Association for Computational Linguistics, 2019.
- [Ribera and Lapedriza, 2019] Mireia Ribera and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *Joint Proceedings of the ACM IUI 2019 Workshops*, page 7, Los Angeles, USA, mar 2019.
- [Rogers *et al.*, 2022] Anna Rogers, Matt Gardner, and Isabelle Augenstein. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 2022.
- [Saeidi *et al.*, 2018] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [Schuff *et al.*, 2022] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. Human interpretation of saliency-based explanation over text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul, South Korea, 2022. Association for Computing Machinery.
- [Sevastjanova *et al.*, 2018] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel A. Keim, and Mennatallah El-Assady. Going beyond visualization : Verbalization as complementary medium to explain machine learning models. In *Workshop on Visualization for AI Explainability at IEEE (VIS)*, 2018.
- [Smith-Renner *et al.*, 2020] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. *No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2020.
- [Socher *et al.*, 2013] 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. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
- [Sokol and Flach, 2020a] Kacper Sokol and Peter Flach. One explanation does not fit all. *KI - Künstliche Intelligenz*, 34(2):235–250, Jun 2020.
- [Sokol and Flach, 2020b] Kacper Sokol and Peter A. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.
- [Strobel *et al.*, 2021] Hendrik Strobel, Benjamin Hoover, Arvind Satyanaryan, and Sebastian Gehrmann. LMDiff: A visual diff tool to compare language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 96–105, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [Strobel *et al.*, 2022] Hendrik Strobel, Jambay Kinley, Robert Krueger, Johanna Beyer, Hanspeter Pfister, and Alexander M. Rush. Genni: Human-ai collaboration for data-backed text generation. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1106–1116, 2022.
- [Stumpf *et al.*, 2009] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dieterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, 2009.
- [Tenney *et al.*, 2020] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. The language interpretability tool: Extensible, interactive visualizations and

- analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online, October 2020. Association for Computational Linguistics.
- [Treviso and Martins, 2020] Marcos Treviso and André F. T. Martins. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118, Online, November 2020. Association for Computational Linguistics.
- [Vafa *et al.*, 2021] Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10314–10332, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [Voigt *et al.*, 2021] Henrik Voigt, Monique Meuschke, Kai Lawonn, and Sina Zarriß. Challenges in designing natural language interfaces for complex visual models. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 66–73, Online, April 2021. Association for Computational Linguistics.
- [Wallace *et al.*, 2019] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. AllenNLP interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [Walton and Krabbe, 1995] Douglas N. Walton and Erik C.W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, 1995.
- [Weitz *et al.*, 2021] Katharina Weitz, Lindsey Vanderlyn, Ngoc Thang Vu, and Elisabeth André. “it’s our fault!”: Insights into users’ understanding and interaction with an explanatory collaborative dialog system. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 1–16, Online, November 2021. Association for Computational Linguistics.
- [Weld and Bansal, 2019] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79, may 2019.
- [Wiegrefe and Marasović, 2021] Sarah Wiegrefe and Ana Marasović. Teach me to explain: A review of datasets for explainable nlp. In *Proceedings of NeurIPS*, 2021.
- [Wiegrefe *et al.*, 2021] Sarah Wiegrefe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [Wiegrefe *et al.*, 2022] Sarah Wiegrefe, Jack Hessel, Swabha Swayamdipta, Mark O. Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, United States, April 2022. Association for Computational Linguistics.
- [Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
- [Wu *et al.*, 2022] Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. QAConv: Question answering on informative conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5389–5411, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [Yao *et al.*, 2021] Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. Refining language models with compositional explanations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [Yeung *et al.*, 2020] Arnold Y. S. Yeung, Shalmali Joshi, Joseph Jay Williams, and Frank Rudzicz. Sequential explanations with mental model-based policies. In *ICML 2020 Workshop on Human Interpretability in Machine Learning*, 2020.
- [Yin and Neubig, 2022] Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. *CoRR*, abs/2202.10419, February 2022.