$<<$ **Beware the Rationalization Trap!** $>>$

When Language Model Explainability Diverges from our Mental Models of Language

Rita Sevastjanova[1] , **Mennatallah El-Assady**[2]

[1]University of Konstanz
[2]ETH Zürich
rita.sevastjanova@uni-konstanz.de, melassady@ai.ethz.ch

Abstract

Language models learn and represent language differently than humans; they learn the *form* and not the *meaning*. Thus, to assess the success of language model explainability, we need to consider the impact of its divergence from a user's mental model of language. In this position paper, we argue that in order to avoid harmful rationalization and achieve truthful understanding of language models, explanation processes must satisfy three main conditions: (1) explanations have to truthfully represent the model behavior, i.e., have a high fidelity; (2) explanations must be complete, as missing information distorts the truth; and (3) explanations have to take the user's mental model into account, progressively verifying a person's knowledge and adapting their understanding. We introduce a decision tree model to showcase potential reasons why current explanations fail to reach their objectives. We further emphasize the need for human-centered design to explain the model from multiple perspectives, progressively adapting explanations to changing user expectations.

1 Introduction

Machine learning models are widely used for diverse applications and use cases, such as machine translation, image recognition, patient diagnostics, etc. These models are typically black boxes as they do not provide any information considering how they evolve nor how they make their predictions. Thus, the Explainable Artificial Intelligence (XAI) research field [Arrieta *et al.*, 2020] has grown rapidly in recent years.

One specific area of interest is modeling and understanding natural language. A wide range of deep-learning-based language models (LMs) (e.g., BERT [Devlin *et al.*, 2019]) have been developed in recent years, reaching high performances in various natural language understanding and generation tasks. These models are pre-trained on a large number of text documents (e.g., *Wikipedia* articles), and are capable of learning various language structures in an unsupervised manner. The popularity of deep-learning-based language models and their widespread usage in diverse natural language processing (NLP) applications for prediction and decision-making [Vig

and Belinkov, 2019] drove the necessity for their explainability. In this paper, we want to emphasize the **challenges related to language model explainability**.

Language (text data) is ***deceivably* intelligible**. We all have an inherent understanding of language, as this is one of the most effective communication mediums for humans. Most people are often under the illusion they understand all aspects of language. However, the reality is that we usually understand the text *semantics* (intelligible) but not as much the text *structure* and linguistic *function* (non-intelligible). Unless we specifically analyze text data, we are usually interpreting language based on our intuition and experience. Similarly to other habitual and intuition-based tasks (think: car driving, for example), understanding language is something that humans do not consciously or effortfully *think* about. However, this is in stark contrast to what we need for understanding language models. These models process text data in a sequential and logical manner, scouting patterns to learn decision boundaries. These often enough do not align with human intuition, making them ideal candidates for *harmful rationalizations*.

Hence, the first major challenge in explaining language models is the misalignment between how language models learn and represent language and the way humans perceive it (through their intuition). In particular, language models currently learn only parts of the semantic and functionality spectrum [Dasgupta *et al.*, 2018]. Thus, also explanation methods might not be faithful to the model's decision-making and are likely to be incomplete. Humans typically tend to "fill the missing gaps" of the explanations by relying on their world understanding. This is dangerous, since such a rationalization can lead to wrong conclusions about the model's behavior. Therefore, it is especially important to think about the **quality** and **completeness** when designing new explanation methods.

The second major challenge is effectively tailoring explanations to specific stakeholders. It has been shown that current language model explanation methods have limited success when it comes to increasing user understanding of the model's behavior [Arora *et al.*, 2021; Hase and Bansal, 2020]. Since there are several user groups involved in the explainability process (e.g., NLP researchers with theoretical or computational background, decision-makers, etc.), they require different types of explanations matching their expectations. Each group of users has its specific mental model of lan-

guage, based on their background, education, or world experience. To reach a positive explainability outcome, it is thus important to design human-centered explanations that align with the different **human mental models of language**.

To summarize, in this position paper we argue that addressing user mental models (i.e., their understanding and use of language) is especially important for the NLP explainability, since humans have a strong mental model of language based on their experience. And when we design or consume explanations, we sometimes forget that language models learn only the form and not meaning [Bender and Koller, 2020]. In this paper, we show that the success of explanations, i.e., *truthful understanding of model behavior*, depends on three main conditions – **explanation quality**, **completeness**, and their **fit to user's mental models**. Only with complete, high fidelity explanations that match or manage to calibrate user mental models can we guarantee to not fall into a *harmful rationalization trap*.

2 Background: Human-Centered XAI

In the following, we describe the types of current explanation methods and the human role in the explanation process.

2.1 Types of Explanations

Most commonly, explanations in XAI are categorized, as follows – (1) whether the explanation is for an individual prediction (local) or the model's prediction process as a whole (global), and (2) whether the explanation is emerging directly from the prediction process (self-explaining) or it requires post-processing (post-hoc) [Guidotti *et al.*, 2018]. Additionally, we can distinguish between (3) explanations on the data-level (model-agnostic) or explanation taking the model behavior into account (model-aware) [Spinner *et al.*, 2020]. Lastly, (4) explanations can be top-down (deductive), bottom-up (inductive), or contrastive [El-Assady *et al.*, 2019]. Regardless of their type, the objectives of these explanations are clear – they should faithfully characterize the models' behavior (i.e., be faithful), and increase user understanding and trust in black-box models [Jacovi and Goldberg, 2020]. They should be interpretable, accurate, and with a high fidelity, i.e., the model should be able to accurately imitate a black-box predictor [Guidotti *et al.*, 2018].

2.2 Human Role in the Explanation Process

The users to whom the explanations are provided play a crucial role in the explanation process and in its success. Yet, it has been criticized that this role is often underestimated. [Ehsan and Riedl, 2020] recently wrote that "*explainability in AI is as much of a Human-Computer Interaction (HCI) problem as it is an AI problem, if not more. Yet, the human side of the equation is often lost in the technical discourse of XAI.*" Also, [Liao and Varshney, 2021] emphasize that the choice of an explanation method should be based on target users' explainability needs, whereby "*their explainability needs can vary significantly depending on their goals, backgrounds, usage contexts, and more.*" So, who are the users for whom the explanations are designed? In the literature, one can find five

main user groups for model explainability: (1) **model developers**, (2) **business owners or administrators**, (3) **decision-makers**, (4) **impacted groups**, (5) **regulatory bodies** [Arrieta *et al.*, 2020]. Based on their needs, each group might prefer a different type of explanation. At the same time, there exist from user groups independent, fundamental properties of explanations. [Miller, 2019] in his survey summarized their four major properties: (1) **explanations are contrastive**, i.e., people do not ask why an event happened, but rather why *it* happened instead of *something else*; (2) **explanations are selected** by humans, and they rarely expect explanations to consist of a complete cause of an event; (3) **explanations are social**, and they are part of a conversation; and (4) **probabilities matter less** than causal relationships.

One of the objectives of good explanations is their faithful characterization of a models' behavior. When these explanations are fragmented, i.e., they partially explain what the model learns or when they don't fully match one's expectations, people start to rationalize about the explanations as well as the explained model. *Rationalization* is an attempt to find reasons for behavior, decisions, etc. [Vanhoucke, 2018] If explanations miss relevant aspects for a complete understanding, users can make false conclusions about the model's behavior and explainability process fails.

3 Language Model Explainability

With the emergence of Transformer architecture [Vaswani *et al.*, 2017], there was a paradigm shift away from building NLP applications on, e.g., Recurrent-Neural-Networks (RNNs) [Bengio *et al.*, 1994] or Long Short-Term Memory networks (LSTMs) [Hochreiter and Schmidhuber, 1997] and the focus was put on (self-) attention mechanisms. The self-attention mechanism processes an input word, accounting for its context, in a bidirectional manner [Vaswani *et al.*, 2017]. Similar to other deep learning models, language models are often fine-tuned with additional output layers for classification tasks (see [Zhuang *et al.*, 2020]). The main idea is to first pre-train the model on large-scale corpora and then fine-tune it on an additional, task-specific labeled corpus [Howard and Ruder, 2018]. And, hence, these models are widely applied in different domains for prediction- and decision-making (e.g., [Sun *et al.*, 2019; Han and Eisenstein, 2019]). Researchers in the explainable NLP field actively analyze which characteristics these models capture both during the pre-training, and after they are fine-tuned for specific tasks (see, e.g., [Rogers *et al.*, 2020; Danilevsky *et al.*, 2021]).

Language models are trained on plain natural language; they have no additional input about word classes, their semantic meaning, or functionality. During the training process, the model represents a word through a contextualized word embedding – word vector that is sensitive to the context in which it appears [Naseem *et al.*, 2021]. [Ethayarajh, 2019] and [Sevastjanova *et al.*, 2021] have explored the embedding contextualization, i.e., how they change within the model's architecture, and showed that the contextualization increases with language model's layers. It has also been shown that in BERT's upper layers, words within the context (i.e., sentence) become more similar to each other [Ethayarajh, 2019].

Limitations of Current Explanation Methods – In a recent survey, [Shen and Huang, 2021] present an overview of methods for NLP explainability, summarizing and tagging over 200 papers that cover related topics. Their work shows that most popular explanation methods in NLP are *feature attribution* (sometimes called *saliency* or *importance*) methods, followed by *tuple*, *rule*, or *concept* formats to demonstrate the model's reasoning process. Less than 1% use *counterfactuals* – methods that according to the theory are preferred by humans [Miller, 2019]. However, the feature attribution methods that are used in around 44% of related papers, are statistical explanations and, theoretically, less sufficient for humans than, for instance, causal explanations [Miller, 2019].

The popular feature attribution methods are local posthoc explanations that show the importance of a token with respect to the model's prediction. A higher score denotes greater importance. Example methods include integrated gradients [Sundararajan *et al.*, 2017] and LIME [Ribeiro *et al.*, 2016]. These scores are often visualized in saliency maps in the form of heatmaps (e.g., [Tenney *et al.*, 2020; Sinha *et al.*, 2021]). It is assumed that users can easily understand why models make their decisions [Sun *et al.*, 2021] or when models pay attention to the wrong words [Fleisher, 2021] by examining the most salient part(s) in the text.

According to the interpretable NLP systems objectives [Jacovi and Goldberg, 2020], saliency visualizations should faithfully characterize the important features and increase the user's understanding of the language model's behavior. However, current studies show that saliency visualizations **do not** help users to understand or simulate the model or flip its output [Arora *et al.*, 2021; Hase and Bansal, 2020], but when they help – the improvement in human performance is only minor [Lai and Tan, 2019]. Although they are commonly used as explanation methods, they do not satisfy the interpretable NLP system objectives. What could be the reason?

4 A Thought Experiment on Explainability

The goal of explainability is to make a model's behavior understandable to different types of users through XAI methods. This explainability process in NLP involves three main aspects, as depicted in Figure 1. First, we have the language models themselves with their architectures, behavior, and learning techniques. Second, we have the various types of explanations by the XAI methods. Third, we have the users and their mental models of language, described in detail in subsection 4.1. Although it would be wishful that these three aspects operate in sync, i.e., the explanations correctly depict model behavior, and the models represent language the way people use it, in reality these aspects are often disconnected.

The two main issues disrupting this process are, on the one hand, that the fidelity of some explanation methods is still disputed; on the other hand, the lack of human-like concept understanding in language models, causing their learning and results to be widely criticized. For instance, [Bender and Koller, 2020] have criticized the hype of language models and scrutinized exaggerated statements, such as, their ability to *understand* or *comprehend* natural language. They argue that a language model that is trained purely on form will not learn meaning – it simply lacks the signal for it.

However, when people learn, they learn from the world around them and from the interaction with other people in that world [Tomasello and Farrar, 1986], i.e., language learning for humans does not happen in isolation, nor based on a single signal. Since this topic is discussed in detail in context of language model learning (please, see [Bender and Koller, 2020]), we will not go in more detail here. But, when designing explanation, we need to consider that language models and humans learn language in vastly different ways.

4.1 Mental Models in NLP

Human mental models of language are critical factors influencing the XAI process outcome.

Mental Models are Human and Subjective – According to [Jones *et al.*, 2011], "*Mental models are personal, internal representations of external reality that people use to interact with the world around them.*" Mental models are crucial in our lives, we use them to reason and make decisions. They also provide the mechanism through which we filter and store new information. Everyone who uses language has a mental model for it. Mental models are subjective, and they depend on a person's background, unique life experiences, perceptions, and understandings of the world [Jones *et al.*, 2011], as well as different circumstances in which the models are used. For example, a person's mental model of a car depends on their association to cars in different roles. Being a car designer or mechanic conjures into the mind a complete different conceptualization than being a driver or passenger.

People Adapt Their Mental Models – In the scope of HCI research field, it has been discussed that mental models evolve as users interact with a system [Norman, 2014]. It is similar with our mental model of language. During our childhood or education phase, we learn things such as word semantic meaning, syntactic structures, word functionality in a sentence, and with each new learned concept we update our mental model of language. If we acquire a second language as adults, we extend our mental model of language, but instead of starting with a blank canvas, we learn in a contrastive manner to our existing understanding.
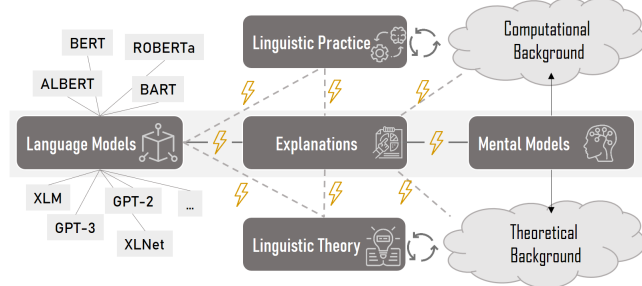

Figure 1: Explainability process in NLP involves three main aspects: language models, explanations, and user mental models. Mental models of language, however, depend on user backgrounds and could be influenced by, e.g., linguistic practice or theory. Since language models learn language differently from humans, there can be many mismatches in the explanation process.
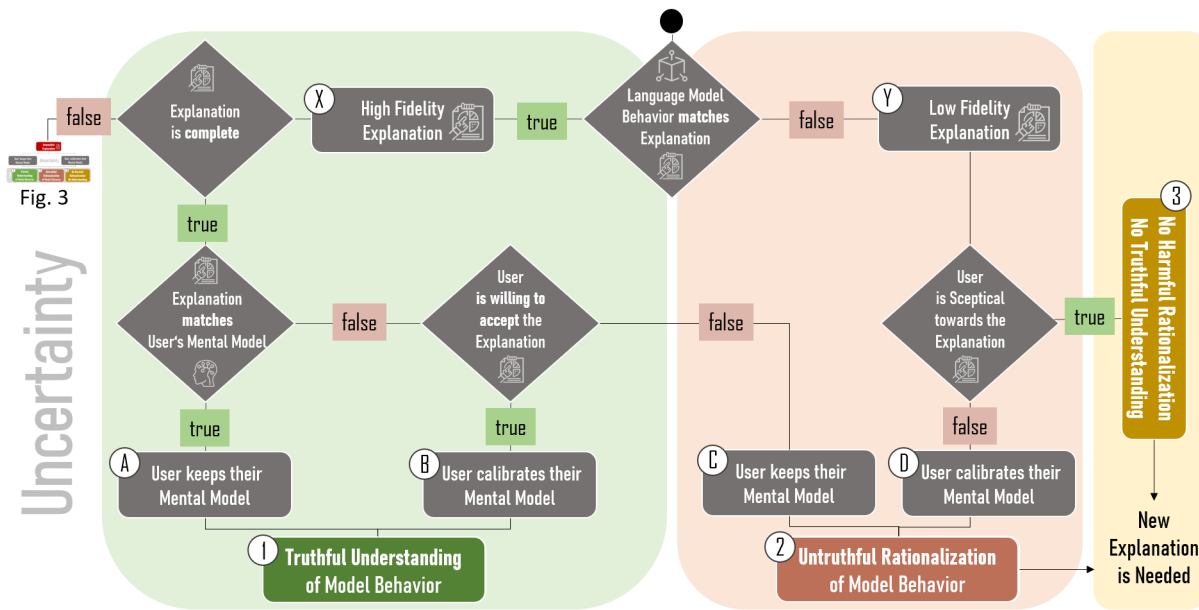
Figure 2: Decision tree showing the explainability process in NLP. (1) **Truthful Understanding** of model behavior can be reached when high fidelity (X), complete explanations match (A) or adapt (B) user mental models. (2) **Untruthful Rationalization** begins when explanations have low fidelity (Y) that change user mental models (D), or users are not willing to change their mental models in favor of high-fidelity explanations (C). Only when low-fidelity explanations raise skepticism in users, (3) **No Truthful Understanding** will be generated.

Types of Mental Models in NLP – As mentioned earlier, mental models are subjective, and they depend on a person's background and tasks at hand. Thus, also in NLP as a field, researchers and practitioners, depending on their educational backgrounds and experiences, can have different mental models of language. On the broadest level, we can differentiate between computational and theoretical linguistic backgrounds. As depicted in Figure 1, mental models of researchers and practitioners with computational background are impacted by their linguistic practice, e.g., which algorithms, techniques they use to (pre-) process text data in their research. In contrast, mental models of researchers and practitioners with theoretical background are framed according to the linguistic theory they study. Although mental models from these two groups of people share a common core, there might also be vast differences between them, for example when judging the importance of specific word categories (e.g., stopwords) for a text analysis task.

4.2 The Process of Explainability and Mental Model Adaptation

As mentioned earlier, the explainability process in NLP consists of three aspects: the language model, the explanations, and the user's mental model. The outcome of the explanation process depends on how these aspects align, i.e., whether the explanations correctly explain the model's behavior (high fidelity), whether they match the user's mental models, and if not, whether the user is willing to adapt their mental model in favor of the explanations.

In the following, we describe the explainability process with its components in a form of a decision tree (shown in Figure 2). We show that the user mental model adaptations

based on truthful understanding or untruthful rationalization both generate knowledge [Pirolli and Card, 2005] and thus need to be carefully audited.

Towards Truthful Understanding – The primary goal of the explanation process is for the user to reach a **truthful understanding** of the model behavior (see Figure 2–1). In an ideal explainability scenario, a model explanation would have a *high fidelity* (see Figure 2–X), i.e., it would match the language model behavior. Moreover, the explanation would be *complete*, covering all the different model's properties. Even if we would have this ideal situation (which is obviously not a simple task to achieve), there would be another important factor that influences the success of the explainability process, i.e., the user's mental model. If the explanation matches the user's mental model (e.g., the explanation confirms that what they know about language models and/or language), then they most likely would keep their mental model (see Figure 2–A) and gain some understanding of the model's behavior. If the explanation differs from the user's expectations, the truthful understanding could be reached only if the user is willing to adapt their mental model to match the high fidelity explanation (see Figure 2–B). The explanation should thus use an appropriate language that is legible for the target user, and it should clearly illustrate the differences between the way the language model captures natural language properties and the particular target user with their specific knowledge and expertise. This case also covers users who are learning about language models for the first time and with no prior conception (and thus no expectation of a mental model match) willing to accept the explanation, fostering a better understanding of the language model behavior.

People Tend to Rationalize – If the high fidelity explanation does not match the user's expectations, and they are not willing to adapt their mental model (see Figure 2–C), there is another possible outcome – the user begins to *rationalize* (see Figure 2–2). One of the reasons might be a lack of trust in the underlying model or the explanation process as such. The mismatch between the explanation outcomes and one's expectations may raise doubts about the model's learning capabilities, as well as the explanation correctness/effectiveness. The reason for users tending to rationalize when interpreting language model outcomes lies in the nature of the input data, i.e., the natural language. In particular, natural language, i.e., text has a meaning (also known as *semantics*). Semantics in some cases are difficult to represent through computational methods. For instance, [Sheth *et al.*, 2005] define three types of semantics, i.e., implicit, formal, and powerful semantics, and write that not all of them are machine processable. For instance, the powerful semantics allow uncertainty permitting not only inductive but also abductive reasoning, i.e., making conclusions from what one knows rather than what one observes. Such semantics are difficult or even impossible to learn and represent by a language model. Thus, there can be a potential mismatch between what the user knows about the language and what they observe in the language model explanations. The user can also misinterpret explanations due to their abductive reasoning. The rationalization can also occur in situations, in which the XAI provides *low fidelity explanations* (see Figure 2–Y). If the user is not skeptical about such explanations and decides to falsely calibrate their mental model (see Figure 2–D), the outcome is dangerous. Previous studies have shown that explanations of language models, even when wrong, can increase human trust in machine predictions; e.g., in experiments showing low-accuracy statements or random heatmaps [Lai and Tan, 2019].

A less severe outcome in the low fidelity case occurs when the users are *skeptical* and don't change their mental model. Here, their vigilance will keep users from accepting a harmful rationalization, but no model behavior understanding will be achieved (see Figure 2–3). Although this outcome does not directly cause any harm, it hinders users from understanding the true potential of the models and might influence their trust in future explanations.

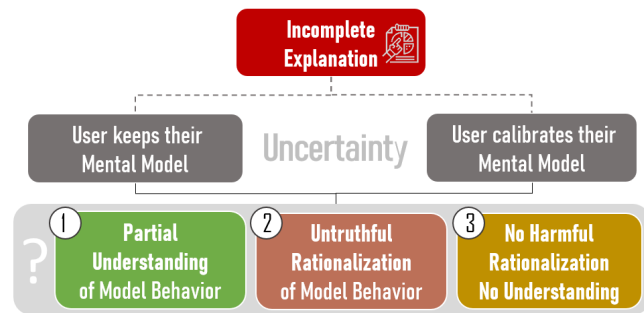Lastly, sometimes the explanation fidelity is *unknown*; in

Figure 3: In reality, many NLP explanations are incomplete, i.e., they partially explain one phenomenon in isolation. Since partial explanations leave empty space for rationalization, there is a high uncertainty in the outcome of the explainability process.

such cases the effect of the explanation is *uncertain*. This is similar to the case of *incomplete explanations*, as described in detail below.

Incomplete Explanations Cause Uncertainty – Commonly, model explanations are tailored toward explaining one or a few phenomena in the black-box model at a time. At the same time, explanations are commonly designed to show a result of one (e.g., attribution method) or a few methods in isolation. As depicted in Figure 3, such incomplete explanations introduce uncertainty. The outcome of incomplete explanations can be threefold. Although two outcomes are less harmful, i.e., either these explanations generate partial understanding of the model's behavior (see Figure 3–1) or no harm at all (see Figure 3–3), there is also a possibility that the *missing information* will prompt users to untruthful rationalization (see Figure 3–2). Since the generation of complete explanations is difficult (if not impossible in some cases), most of the current explanation processes contain a high uncertainty in their outcome. Thus, we emphasize the need for a more human-centered explanation process where the explanations clearly highlight the differences between the users' understanding of language and the language model behavior and where they are communicated in a way that allows the users to keep or calibrate their mental models with high certainty.

5 Supporting Mental Model Calibration

In the following, we describe actionable steps that can be undertaken to support users in gaining truthful understanding of language model behavior.

Avoiding False Rationalization – In order to make the explanation process successful, we need to avoid several factors: (1) low fidelity explanations, (2) incomplete explanations, and (3) detachment from users' mental models. To create high fidelity explanations, we need to create means and metrics to appropriately measure their quality [Guidotti *et al.*, 2018]. The design of the explanations must be motivated by language model inner-workings; the user should get the chance to get insights into relevant model properties. The explanations must use language that is familiar to the users. We need to first understand the user's mental models and expectations about a phenomenon before we try to explain it, ideally relying on metaphors familiar to the user [Jentner *et al.*, 2018].

Catch and Counteract Pitfall Scenarios – To ensure that users don't fall into a rationalization deadlock, we need to implement means to detect when the explanations have brought users to the wrong *path*. Verification steps of that what users understand from the explanations could be integrated in the explainability process. In particular, from time to time, we could test what the user has understood about the model's working mechanisms, e.g., using verification questions [El-Assady *et al.*, 2019]. If the generated insights are wrong, we would need to correct and educate the user, possibly adapting the used explanation method.

Explanation as an Adaptive Process – Explaining a phenomenon is as much about choosing the right methods as

it is about enabling the person receiving the explanation to gradually adapt their mental model. Building XAI pipelines that consist of smaller explainer units has been suggested by [Spinner *et al.*, 2020] to allow for the progressive development of explanations based on the user's understanding and tasks. Further, we can enhance this progressive process by deploying co-adaptive guidance and analysis strategies [Sperrle *et al.*, 2021] to capture and react to varying users' mental models – if the explanation system detects that the user misunderstands the presented explanations, the system could present a different type of explanation or enhance the explanation with more details.

Contrastive Explanations – Top-down explanations is a common practice in XAI research. Researchers define hypotheses and test, e.g., through probing classifiers [Belinkov, 2018; Conneau *et al.*, 2018], whether the model captures specific phenomena. Probing classifiers are typically black-box models and, thus, such explanations can lead to harmful rationalizations. Contrasting such explanations with bottom-up or counterfactual reasoning can lead to more nuanced perspectives, e.g., by unfolding new specificities about a model's behavior. These explanations are especially important when users have strong mental models about related phenomena, such as on language model learning mechanisms, requiring them to undergo a necessary calibration of their mental models. Hence, we might want to consider designing systems that support open-ended exploration, as well as diverse contrastive and bottom-up explanations.

Human-like Explanations – [Ehsan *et al.*, 2018] has introduced a technique called *AI Rationalization* for explanation generation that translates internal state-actions of an autonomous agent into natural language. They argue that such a human-like communication is more accessible and intuitive to users. We believe that such explanations are particularly relevant for user groups whose mental models about a specific phenomenon differs from the way language models depict it. In these cases, pure statistical explanations, e.g., saliency scores, are incomplete and insufficient to provide truthful understanding. Thus, we might consider extending statistical explanations with explanations in natural language that translates which features are considered as important for prediction making and why in a simple and accessible way [Sevastjanova *et al.*, 2018].

Collaborative Learning – To ease the learning process, we could benefit from collaborative environments where multiple users work together. Collaborative learning motivates people to share their mental models of language and, thus, enable them to learn from both, the system and their peers in the explanation environment.

6 Conclusion

This paper discusses the need for proper human-centered design solutions for NLP explainability. Our thought experiment illustrated the explainability process and its possible outcomes under the consideration of three factors; explanation completeness, explanation quality, and user mental models. We argue that only complete, high fidelity explanations

that match or calibrate user mental models will create user truthful understanding in language model behavior. To support users to gain truthful understanding of language model behavior, we introduce several human-centered solutions.

In the future, we will work on quantifying the effects of rationalizations through different user studies using explainability interfaces. Additionally, we aim to study the differences of explaining intelligible versus non-intelligible data, specifically the effects of intelligibility on rationalization.

References

[Arora *et al.*, 2021] 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. *arXiv preprint arXiv:2112.09669*, 2021.

[Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[Belinkov, 2018] Yonatan Belinkov. *On internal language representations in deep learning: An analysis of machine translation and speech recognition*. PhD thesis, Massachusetts Insitute of Technology, 2018.

[Bender and Koller, 2020] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics.

[Bengio *et al.*, 1994] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[Conneau *et al.*, 2018] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Danilevsky *et al.*, 2021] Marina Danilevsky, Shipi Dhanorkar, Yunyao Li, Lucian Popa, Kun Qian, and Anbang Xu. *Explainability for Natural Language Processing*, page 4033–4034. Association for Computing Machinery, New York, NY, USA, 2021.

[Dasgupta *et al.*, 2018] Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. Evaluating Compositionality in Sentence Embeddings. *CoRR*, abs/1802.04302, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Ehsan and Riedl, 2020] Upol Ehsan and Mark O Riedl. Human-centered explainable AI: Towards a reflective sociotechnical approach. In *Int. Conf. on Human-Computer Interaction*, pages 449–466. Springer, 2020.

[Ehsan *et al.*, 2018] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proc. of the Conf. on AI, Ethics, and Society*, AIES '18, page 81–87, New York, NY, USA, 2018. Association for Computing Machinery.

[El-Assady *et al.*, 2019] Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, and Daniel Keim. Towards Explainable Artificial Intelligence: Structuring the Processes of Explanations. In *ACM CHI Workshop: Human–Centered Machine Learning Perspectives*, 2019.

[Ethayarajh, 2019] Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proc. of the Conf. on Empirical Methods in Natural Language Proc. and the Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Fleisher, 2021] Will Fleisher. Understanding, Idealization, and Explainable AI. *Idealization, and Explainable AI (October 19, 2021)*, 2021.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[Han and Eisenstein, 2019] Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *EMNLP*, 2019.

[Hase and Bansal, 2020] Peter Hase and Mohit Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *ACL*, 2020.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Jacovi and Goldberg, 2020] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

[Jentner *et al.*, 2018] Wolfgang Jentner, Rita Sevastjanova, Florian Stoffel, Daniel A Keim, Jürgen Bernard, and Mennatallah El-Assady. Minions, sheep, and fruits: Metaphorical narratives to explain artificial intelligence and build trust. In *Workshop on Visualization for AI Explainability at IEEE*, 2018.

[Jones *et al.*, 2011] Natalie A Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. Mental models: An interdisciplinary synthesis of theory and methods. *Ecology and Society*, 16(1), 2011.

[Lai and Tan, 2019] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proc. of the Conf. on Fairness, Accountability, and Transparency*, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery.

[Liao and Varshney, 2021] Q Vera Liao and Kush R Varshney. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790*, 2021.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[Naseem *et al.*, 2021] Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35, 2021.

[Norman, 2014] Donald A Norman. *Some observations on mental models*. Psychology Press, 2014.

[Pirolli and Card, 2005] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. of Int. Conf. on Intelligence Analysis*, volume 5, pages 2–4. McLean, VA, USA, 2005.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[Rogers *et al.*, 2020] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Trans. of the Association for Computational Linguistics*, 8:842–866, 2020.

[Sevastjanova *et al.*, 2018] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel 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*. oct 2018.

[Sevastjanova *et al.*, 2021] Rita Sevastjanova, Aikaterini-Lida Kalouli, Christin Beck, Hannah Hauptmann, and Mennatal-

lah El-Assady. Explaining Contextualization in Language Models using Visual Analytics. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics*, ACL 2021. Association for Computational Linguistics, 2021.

[Shen and Huang, 2021] Hua Shen and Ting-Hao (Kenneth) Huang. Explaining the road not taken. *ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI*, 2021.

[Sheth *et al.*, 2005] Amit Sheth, Cartic Ramakrishnan, and Christopher Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 1(1):1–18, 2005.

[Sinha *et al.*, 2021] Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing inputs for fragile interpretations in deep natural language processing. *arXiv preprint arXiv:2108.04990*, 2021.

[Sperrle *et al.*, 2021] Fabian Sperrle, Astrik Jeitler, Jürgen Bernard, Daniel Keim, and Mennatallah El-Assady. Co-adaptive visual data analysis and guidance processes. *Computers & Graphics*, 100:93–105, 2021.

[Spinner *et al.*, 2020] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. 26(1):1064–1074, 2020.

[Sun *et al.*, 2019] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *CCL*, 2019.

[Sun *et al.*, 2021] Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Qiu Han, Guoyin Wang, Eduard Hovy, and Jiwei Li. Interpreting deep learning models in natural language processing: A review. *arXiv preprint arXiv:2110.10470*, 2021.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Int. Conf. on Machine Learning*, pages 3319–3328. PMLR, 2017.

[Tenney *et al.*, 2020] I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, and A. Yuan. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models, 2020.

[Tomasello and Farrar, 1986] Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child development*, pages 1454–1463, 1986.

[Vanhoucke, 2018] Vincent Vanhoucke. Interpretability and post-rationalization, Oct 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Vig and Belinkov, 2019] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

[Zhuang *et al.*, 2020] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proc. of the IEEE*, 109(1):43–76, 2020.