Aligning Faithful Interpretations with their Social Attribution

Alon Jacovi
Bar Ilan University
alonjacovi@gmail.com

Yoav Goldberg
Bar Ilan University and
Allen Institute for AI
yoav.goldberg@gmail.com

Abstract

We find that the requirement of model interpretations to be faithful is vague and incomplete. With interpretation by textual highlights as a case study, we present several failure cases. Borrowing concepts from social science, we identify that the problem is a misalignment between the causal chain of decisions (causal attribution) and the attribution of human behavior to the interpretation (social attribution). We reformulate faithfulness as an accurate attribution of causality to the model, and introduce the concept of aligned faithfulness: faithful causal chains that are aligned with their expected social behavior. The two steps of causal attribution and social attribution together complete the process of explaining behavior. With this formalization, we characterize various failures of misaligned faithful highlight interpretations, and propose an alternative causal chain to remedy the issues. Finally, we implement highlight explanations of the proposed causal format using contrastive explanations.

1 Introduction

When formalizing the desired properties of a quality interpretation of a model or a decision, the NLP community has settled on the key property of *faithfulness* (Lipton, 2018; Herman, 2017; Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020), or how ''accurately'' the interpretation represents the true reasoning process of the model.

A common pattern of achieving faithfulness in interpretation of neural models is via decomposition of a model into steps and inspecting the intermediate steps (Doshi-Velez and Kim, 2017, cognitive chunks). For example, neural modular networks (NMNs; Andreas et al., 2016) first build an execution graph out of neural building blocks, and then apply this graph to data. The graph structure is taken to be a faithful interpretation of the model's behavior, as it describes the computation precisely. Similarly, *highlight* methods (also called *extractive rationales*[1]), decompose a textual prediction problem into first *selecting* highlighted texts, and then *predicting* based on the selected words (*select-predict*, described in Section 2). The output of the selection component is taken to be a faithful interpretation, as we know exactly which words were selected. Similarly, we know that words that were not selected do not participate in the final prediction.

However, Subramanian et al. (2020) call NMN graphs *not faithful* in cases where there is a discrepancy between a building block's behavior and its name (i.e., expected behavior). Can we better characterize the requirement of faithful interpretations and amend this discrepancy?

We take an extensive and critical look at the formalization of faithfulness and of explanations, with textual highlights as an example use-case. In particular, the *select-predict* formulation for faithful highlights raises more questions than it provides answers: We describe a variety of curious failure cases of such models in Section 4, as well as experimentally validate that the failure cases are indeed possible and do occur in practice. Concretely, the behavior of the selector and predictor in these models do not necessarily line up with expectations of people viewing the highlight. Current literature in ML and NLP interpretability fails to provide a theoretical foundation to characterize these issues (Sections 4, 6.1).

[1]The term ''rationale'' (Lei et al., 2016) is more commonly used for this format of explanation in NLP, for historical reasons: Highlights were associated with human rationalization of data annotation Zaidan et al. (2007). We argue against widespread use of this term, as it refers to multiple concepts in NLP and ML (e.g., Zaidan et al., 2007; Bao et al., 2018; DeYoung et al., 2019; Das and Chernova, 2020), and importantly, ''rationalization'' attributes human intent to the highlight selection, which is not necessarily compatible with the model, as we show in this work.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 294–310, 2021. https://doi.org/10.1162/tacl_a_00367
Action Editor: Dan Gildea. Submission batch: 9/2020; Revision batch: 11/2020; Published 3/2021.

To remedy this, we turn to literature on the science of social explanations and how they are utilized and perceived by humans (Section 6): The social and cognitive sciences find that human explanations are composed of two, equally important parts: The attribution of a causal chain to the decision process (*causal attribution*), and the attribution of social or human-like intent to the causal chain (*social attribution*) (Miller, 2019), where ''human-like intent'' refers to a system of beliefs and goals behind or following the causal process. For example, ''she drank the water *because she was thirsty*.''[2]

People may also *attribute social intent to models*: In the context of NLP, when observing that a model consistently translates ''doctor'' with male morphological features (Stanovsky et al., 2019), the user may attribute the model with a ''belief'' that all doctors are male, despite the model lacking an explicit system of beliefs. Explanations can influence this social attribution: For example, a highlight-based explanation may influence the user to attribute the model with the intent of ''performing a summary before making a decision'' or ''attempting to justify a prior decision''.

Fatally, the second key component of human explanations—the social attribution of intent—has been missing from current formalization on the desiderata of artificial intelligence explanations. In Section 7 we define that a faithful interpretation —a causal chain of decisions—is *aligned* with human expectations if it is adequately constrained by the social behavior attributed to it by human observers.

Armed with this knowledge, we can now verbalize the issue behind the ''non-faithfulness'' perceived by Subramanian et al. (2020) for NMNs: The inconsistency between component names and their actual behavior causes a misalignment between the causal and social attributions. We can also characterize the more subtle issue underlying the failures of the *select-predict* models described in Section 4: In Section 8 we argue that for a set of possible social attributions, the *select-predict* formulations fails to guarantee *any* of them.

In Section 9 we propose an alternative causal chain for highlights explanations: *predict-select-verify*. Predict-select-verify does not suffer from the issue of misaligned social attribution, as the highlights can only be attributed as *evidence* towards the predictor's decision. As a result, predict-select-verify highlights do not suffer from the misalignment failures of Section 4, and guarantee that the explanation method does not reduce the score of the original model.

Finally, in Section 10 we discuss an implementation of *predict-select-verify*, namely, designing the components in the roles predictor and selector. Designing the selector is non-trivial, as there are many possible options to select highlights that evidence the predictor's decision, and we are only interested in selecting ones that are meaningful for the user to understand the decision. We leverage observations from cognitive research regarding the internal structure of (human-given) explanations, dictating that explanations must be *contrastive* to hold tangible meaning to humans. We propose a classification *predict-select-verify* model that provides contrastive highlights—to our knowledge, a first in NLP—and qualitatively exemplify and showcase the solution.

Contributions. We identify shortcomings in the definitions of faithfulness and plausibility to characterize what is useful explanation, and argue that the *social attribution* of an interpretation method must be taken into account. We formalize ''aligned faithfulness'' as the degree to which the causal chain is aligned with the social attribution of intent that humans perceive from it. Based on the new formalization, we (1) identify issues with current *select-predict* models that derive faithful highlight interpretations, and (2) propose a new causal chain that addresses these issues, termed *predict-select-verify*. Finally, we implement this chain with *contrastive explanations*, previously unexplored in NLP explanaibility. We make our code available online.[3]

2 Highlights as Faithful Interpretations

Highlights, also known as extractive rationales, are binary masks over a given input that imply some behavioral interpretation (as an incomplete description) of a particular model's decision process to arrive at a decision on the input. Given input sequence $x \in \mathbb{R}^n$ and model $m : \mathbb{R}^n \longrightarrow Y$, a highlight interpretation $h \in \mathbb{Z}_2^n$ is a binary mask over x that attaches a meaning to $m(x)$, where

[2]Note that coincidence (lack of intent) is also a part of this system.

[3]https://github.com/alonjacovi/aligned-highlights.

the portion of x highlighted by h was important to the decision.

This functionality of h was interpreted by Lei et al. (2016) as an implication of a behavioral process of $m(x)$, where the decision process is a modular composition of two unfolding stages:

1. **Selector** component $m_s : \mathbb{R}^n \longrightarrow \mathbb{Z}_2^n$ selects a binary highlight h over x.

2. **Predictor** component $m_p : \mathbb{R}^n \longrightarrow Y$ makes a prediction on the input $h \odot x$.

The final prediction of the system at inference is $m(x) = m_p(m_s(x) \odot x)$. We refer to $h := m_s(x)$ as the **highlight** and $h \odot x$ as the **highlighted text**.

What does the term ''faithfulness'' mean in this context? A highlight interpretation can be *faithful* or *unfaithful* to a model. Literature accepts a highlight interpretation as ''faithful'' if the highlighted text was provably the only input to the predictor.

Implementations. Various methods have been proposed to train select-predict models. Of note: Lei et al. (2016) propose to train the selector and predictor end-to-end via REINFORCE (Williams, 1992), and Bastings et al. (2019) replace REINFORCE with the reparameterization trick (Kingma and Welling, 2014). Jain et al. (2020) propose FRESH, where the selector and predictor are trained separately and sequentially.

3 Use-cases for Highlights

To discuss whether an explanation procedure is useful as a description of the model's decision process, we must first discuss what is considered useful for the technology.

We refer to the following use-cases:

Dispute: A user may want to dispute a model's decision (e.g., in a legal setting). They can do this by disputing the selector or predictor: by pointing to some non-selected words, saying: ''the model wrongly ignored *A*,'' or by pointing to selected words and saying: ''based on this highlighted text, I would have expected a different outcome.''

Debug: Highlights allow the developer to designate model errors into one of two categories: Did the model focus on the wrong part of the input, or did the model make the wrong prediction based on the correct part of the input? Each category implies a different method of alleviating the problem.

Advice: Assuming that the user is unaware of the ''correct'' decision, they may (1) elect to trust the model, and learn from feedback on the part of the input relevant to the decision; or (2) elect to increase or decrease trust in the model, based on whether the highlight is aligned with the user's prior on what the highlight should or should not include. For example, if the highlight is focused on punctuation and stop words, whereas the user believes it should focus on content words.

4 Limitations of Select-Predict Highlights as (Faithful) Interpretations

We now explore a variety of circumstances in which select-predict highlight interpretations are **uninformative** to the above use-cases. While we present these failures in context of highlights, they should be understood generally as symptoms of mistaken or missing formal perspective of explanations in machine learning. Specifically, that **faithfulness is insufficient to formalize the desired properties of interpretations.**

4.1 Trojan Explanations

Task information can manifest in the interpretation in exceedingly unintuitive ways, making it faithful, but functionally useless. We lead with an example, to be followed by a more exact definition.

Leading Example. Consider the following case of a faithfully highlighted decision process:

1. The selector makes a prediction y, and encodes y in a highlight pattern h. It then returns the highlighted text $h \odot x$.

2. The predictor recovers h (the binary mask vector) from $h \odot x$ and decodes y from the mask pattern h, *without* relying on the text.

The selector may choose to encode the predicted class by the location of the highlight (beginning vs. end of the text), or as text with more highlighted tokens than non-highlighted, and so forth. This is problematic: Although the model appears to make a decision based on the highlight's word content, the functionality of the highlight serves a different purpose entirely.

Evidently, this highlight ''explains'' nothing of value to the user. The role of the highlight is completely misaligned with the role expected by the user. For example, in the *advice* use-case, the highlight may appear random or incomprehensible. This will cause the user to lose trust in

Model	SST-2	AGNews	IMDB	Ev.Inf.	20News	Elec
Random baseline	50.0	25.0	50.0	33.33	5.0	50.0
Lei et al. (2016)	59.7	41.4	69.11	33.45		60.75
Bastings et al. (2019)	62.8	42.4			9.45	
FRESH	52.22	35.35	54.23	38.88	11.11	58.38

Table 1: The performance of an RNN classifier using h alone as input, in comparison to the random baseline. Missing cells denote cases where we were unable to converge training.

the prediction, even if the model was making a reasonable and informed decision.

This case may seem unnatural and unlikely. Nevertheless, it is *not explicitly avoided* by faithful highlights of *select-predict* models: This ''unintentional'' exploit of the modular process is a valid trajectory in the training process of current methods. We verify this by attempting to predict the model's decision based on the mask h alone via another model (Table 1). The experiment surprisingly succeeds at above-random chance. Although the result does not ''prove'' that the predictor uses this unintuitive signal, it shows that *there is no guarantee that it doesn't.*

Definition (Trojan Explanations). We term the more general phenomenon demonstrated by the example a *Trojan explanation*: The explanation (in our case, $h \odot x$) carries information that is encoded in ways that are ''unintuitive'' to the user, who observes the interpretation as an explanation of model behavior. This means that the user observing the explanation naturally expects the explanation to convey information in a particular way,[4] which is different from the true mechanism of the explanation.

The ''unintuitive'' information encoded in $h \odot x$ is not limited to h itself, and can be anything that is useful to predict y and that the user will be unlikely to easily comprehend. To illustrate, we summarize cases of Trojans (in highlight interpretations), which are reasonably general to multiple tasks:

1. **Highlight signal**: The label is encoded in the mask h alone, requiring no information from the original text it is purported to focus on.

2. **Arbitrary token mapping**: The label is encoded via some mapping from highlighted tokens to labels which is considered arbitrary to the user: for example, commas for one class, and periods for another; the quantity of capital letters; distance between dashes, and so on.

3. **The default class**: In a classification case, a class can be predicted by precluding the ability to predict all other classes and selecting it by default. As a result, the selector may decide that the absence of class features in itself defines one of the classes.

[4]The expectation of a ''particular way'' is defined by the attribution of intent, explored later (§7). Precise descriptions of this expectation depend on the model, task, and user.

Model	AGNews	IMDB	20News	Elec
Full text baseline	41.39	51.22	8.91	56.41
Lei et al. (2016)	46.83	57.83		60.4
Bastings et al. (2019)	47.69		9.66	
FRESH	43.29	52.46	12.53	57.7

Table 2: The performance of a classifier using quantities of the following tokens in $h \odot x$: comma, period, dash, escape, ampersand, brackets, and star; as well as the quantity of capital letters and $|h \odot x|$. Missing cells are cases where we were unable to converge training.

In Table 2 we experiment with type (2) above: we predict the decision (via MLP classifier) of a model from quantities of various characters, such as commas and dashes, in the highlighted texts generated by the models.[5] We compare to a baseline of predicting the decisions based on the same statistics from the full text. Surprisingly, all models show an *increased* ability to predict their decisions on some level compared to the baseline.

Conclusion. Trojan explanations are not merely possible, but *just as reasonable* to the model as any other option unless countered explicitly. However, explicit modeling of Trojans is difficult, as our definition depends on user perception and contains limitless possible exploits. Even more critical, our current formal definition of what constitutes a faithful explanation does not rule out trojan explanations, and we cannot point to a property that makes such trojan explanations undesirable.

4.2 The Dominant Selector

In another failure case, the selector makes an implicit decision, and proceeds to *manipulate* the

[5]We count the following characters for a feature vector of length 10: comma, period, dash, escape, ampersand, both brackets, quantity of capital letters, and length (by tokens) of the highlighted text.

Model	Text and Highlight	Prediction
(a)	i really don't have much to say about this book holder, not that it's just a book holder. it's a nice one. it does it's job . it's a little too expensive for just a piece of plastic. it's strong, sturdy, and it's big enough, even for those massive heavy textbooks, like the calculus ones. although, i would not recommend putting a dictionary or reference that's like 6'' thick (even though it still may hold). it's got little clamps at the bottom to prevent the page from flipping all over the place, although those tend to fall off when you move them. but that's no big deal. just put them back on. this book holder is kind of big, and i would not put it on a small desk in the middle of a classroom, but it's not too big. you should be able to put it almost anywhere when studying on your own time.	Positive
(b)	i really don't have much to say about this book holder, not that it's just a book holder it's a nice one. it does it's job . it's a little too expensive for just a piece of plastic. it's strong, sturdy, and it's big enough, even for those massive heavy textbooks, like the calculus ones. although, i would not recommend putting a dictionary or reference that's like 6'' thick (even though it still may hold). it's got little clamps at the bottom to prevent the page from flipping all over the place, although those tend to fall off when you move them. but that's no big deal. just put them back on. this book holder is kind of big, and i would not put it on a small desk in the middle of a classroom. but it's not too big. you should be able to put it almost anywhere when studying on your own time.	Positive

Table 3: Highlights faithfully attributed to two fictional select-predict models on an elaborate Amazon Reviews sentiment classification example. Although highlight (a) is easier to understand, it is also far less useful, as the selector clearly made hidden decisions.

predictor towards this decision (without necessarily manifesting as a ''Trojan''). This means that **the selector can dictate the decision with a highlight that is detached from the selector's inner reasoning process.**

Whereas in the case of Trojan explanations the highlight's explanatory power is misunderstood by the user (but nevertheless exists), in this failure case, the information in the highlight is unproductive as an explanation altogether.

Suppose that the selector has made some decision based on some span A in the input, while producing span B to pass to the predictor—confident that the predictor will make the same prediction on span B as the selector did on span A. Although span B may seem reasonable to human observers, it is a ''malicious'' manipulation of the predictor.

The dominant selector can realistically manifest when span A is more informative to a decision than span B, but the selector was incentivized, for some reason, to prefer producing span B over span A. This is made possible because, while span B may not be a good predictor for the decision, it can become a good predictor *conditioned on the existence of span A in the input.* Therefore, **as far as the predictor is concerned, the probability of the label conditioned on span B is as high as the true probability of the label conditioned on span A.** We demonstrate with examples:

Example 1. Consider the case of the *binary sentiment analysis* task, where the model predicts the polarity of a particular snippet of text. Given this fictional movie review:

''Interesting movie about the history of *Iran*^{A}, only *disappointed*^{B} that it's so short.''

Assume that a select-predict model where the selector was trained to mimic human-provided rationales (DeYoung et al., 2019), and the predictor made a (mistaken) negative sentiment classification. Assume that *Iran* (A) is highly correlated with negative sentiment, more-so than *disappointed* (B)—as ''not disappointed'' and such are also common. Changing the word ''Iran'' to ''Hawaii'', for example, will change the prediction of the model from negative to positive. However, this correlation may appear controversial or unethical, and thus, humans tend to avoid rationalizing with it explicitly. The selector will be incentivized to make a negative prediction because of *Iran* while passing *disappointed* to the predictor.

Because the choice of span B is conditioned on the choice of span A (meaning that the selector will choose *disappointed* only if it had a priori decided on the negative class thanks to *Iran*), span B is just as informative to the predictor as span A is in predicting the negative label.

This example is problematic not only due to the ''interpretable'' model behaving unethically, but **due to the inherent incentive of the model to lie and pretend it had made an innocent mistake** of overfitting to the word ''disappointed''.

Example 2. Assume that two fictional select-predict models attempt to classify a complex, mixed-polarity review of a product. Table 3 describes two fictional highlights faithfully attributed

to these two models, on an example selected from the Amazon Reviews dataset.

The models make the same decision, yet their implied reasoning process is wildly different, thanks to the different highlight interpretations: Model (a)'s selector made some decision and selected a word, ''nice'', which trivially supports that decision. The predictor, which can only observe this word, simply does as it is told. Comparatively, the selector of model (b) performed a very different job: as a summarizer. The predictor then made an informed decision based on this summary. How the predictor made its decision is unclear, but the *division of roles* in (b) is significantly easier to comprehend—since the user expects the predictor to make the decision based on the highlight.

This has direct practical implications: In the *dispute* use-case, given a dispute claim such as ''the sturdiness of the product was not important to the decision'', the claim appears impossible to verify in (a). The true decision may have been influenced by words which were not highlighted. The claim *appears to be* safer in (b). But why is this the case?

4.3 Loss of Performance

It is common for select-predict models to perform worse on a given task in comparison to models that classify the full text in ''one'' opaque step (Table 4). Is this phenomenon a reasonable necessity of interpretability? Naturally, humans are able to provide highlights of decisions without any loss of accuracy. Additionally, while interpretability may sometimes be prioritized over state-of-the-art performance, there are also cases that will disallow the implementation of artificial models unless they are strong *and* interpretable.[6]

We can say that there is some expectation for whether models *can* or *cannot* surrender performance in order to explain themselves. This expectation may manifest in one way or the other for a given interaction of explanation. And regardless of what this expectation may be in this scenario, select-predict models do follow the former (loss of performance exists) and human rationalization follows the latter (loss of performance does not exist), such that there is a clear mismatch between the two. How can we formalize if, or whether, this behavior of select-predict models is reasonable? What is the differentiating factor between the two situations?

[6]For example, in the case of a doctor or patient seeking life-saving advice—it is difficult to quantify a trade-off between performance and explanation ability.

5 Explanatory Power of Highlights

We have established three failure cases of select-predict highlights: Trojan explanations (§4.1) cause a misinterpretation of the highlight's functionality, and in dominant selectors (§4.2), the highlight does not convey any useful information. Finally, loss of performance (§4.3) shows an inherent, unexplained mismatch between the behavior of select-predict explainers and human explainers.

All of these cases stem from a shared failure in formally characterizing the information to be conveyed to the user. For example, Trojan explanations are a symptom of the selector and predictor communicating through the highlight interface in an ''unintended'' manner; dominant selectors are a symptom of the selector making the highlight decision in an ''unintended'' manner, as well—but this is entirely due to the fact that we did not define what is intended, to begin with. Further, this is a general failure of interpretability, not restricted to select-predict highlight interpretations.

6 On Faithfulness, Plausibility, and Explanainability from the Science of Human Explanations

The mathematical foundations of machine learning and natural language processing are insufficient to tackle the underlying issue behind the symptoms described in Section 4. In fact, formalizing the problem itself is difficult. What enables a faithful explanation to be ''understood'' as accurate to the model? And what causes an explanation to be perceived as a Trojan?

In this section, we attempt to better formalize this problem on a vast foundation of social, psychological and cognitive literature about human explanations.[7]

6.1 Plausibility is not the Answer

Plausibility[8] (or *persuasiveness*) is the property of an interpretation being convincing towards the model prediction, regardless of whether the model was correct or whether the interpretation is

[7]Refer to Miller (2019) for a substantial survey in this area, which was especially motivating to us.

[8]Refer to the Appendix for a glossary of relevant terminology from the human explanation sciences.

	SST-2	SST-3	SST-5	AG News	IMDB	Ev. Inf.	MultiRC	Movies	Beer
Lei et al. (2016)	*22.65*	*7.09*	9.85	33.33	*22.23*	36.59	31.43	160.0	37.93
Bastings et al. (2019)	*3.31*	*0*	2.97	199.02	*12.63*	85.19	75.0		13.64
FRESH	90.0	*17.82*	*13.45*	50.0	*14.66*	9.76	0.0	20.0	

Table 4: The percentage *increase* in error of selector-predictor highlight methods compared to an equivalent architecture model which was trained to classify complete text. We report the numbers reported in previous work whenever possible (*italics* means our results). Architectures are *not* necessarily consistent across the table, thus they do not imply performance superiority of any method. The highlight lengths chosen for each experiment were chosen with precedence whenever possible, and otherwise chosen as 20% following Jain et al. (2020) precedence.

faithful. It is inspired by human-provided explanations as post hoc stories generated to plausibly justify our actions (Rudin, 2019). Plausibility is often quantified by the degree that the model's highlights resemble gold-annotated highlights given by humans (Bastings et al., 2019; Chang et al., 2020) or by querying for the feedback of people directly (Jain et al., 2020).

Following the failure cases in Section 4, one may conclude that plausibility is a desirable, or even necessary, condition for a good interpretations: after all, Trojan explanations are by default implausible. We strongly argue this is not the case.

Plausibility should be viewed an incentive of the explainer, and not as a property of the explanation: Human explanations can be categorized by utility across multiple axes (Miller, 2019), among them are (1) *learning* a better internal model for future decisions and calculations (Lombrozo, 2006; Williams et al., 2013); (2) *examination* to verify the explainer has a correct internal prediction model; (3) *teaching*[9] to modify the internal model of the explainer towards a more correct one (can be seen as the opposite end of (1)); (4) *assignment of blame* to a component of the internal model; and finally, (5) *justification* and *persuasion*. These goals can be trivially mapped to our case, where the explainer is artificial.

Critically, goal (5) of justification and persuasion by the explainer may not necessarily be the goal of the explainee. Indeed, in the case of AI explainability, it is *not* a goal of the explainee to be persuaded that the decision is correct (even when it is), but to understand the decision process. If plausibility is a goal of the artificial model, this perspective outlines a game theoretic mismatch of incentives between the two players. And specifically in cases where the model is incorrect, it is interpreted as the difference between an innocent mistake and an intentional lie—of course, lying is considered more unethical. As a result, we conclude that **modeling and pursuing plausibility in AI explanations is an ethical issue**.

The failures discussed above do not stem from how (un)convincing the interpretation is, but from *how well the user understands the reasoning process of the model*. If the user is able to comprehend the *steps* that the model has taken towards its decision, then the user will be the one to decide whether these steps are plausible or not, based on how closely they fit the user's prior knowledge on whatever correct steps should be taken—regardless of whether the user knows the correct answer or whether the model is correct.

6.2 The Composition of Explanations

Miller (2019) describes human explanations of behavior as a social interaction of knowledge transfer between the explainer and the explainee, and thus *they are contextual*, and can be perceived differently depending on this context. Two central pillars of the explanation are *causal attribution* —the attribution of a causal chain[10] of events to the behavior—and *social attribution*—the attribution of intent to others (Heider et al., 1958).

Causal attribution describes faithfulness: We note a stark parallel between causal attribution and faithfulness—for example, the select-predict composition of modules defines an unfolding causal chain where the selector hides portions of the input, causing the predictor to make a decision based on the remaining portions. In fact, recent

[9] Although (1) and (3) are considered one-and-the-same in the social sciences, we disentangle them as that is only the case when the explainer and explainee are both human.

[10] See Hilton et al. (2005) for a breakdown of types of causal chains; we focus on unfolding chains in this work, but others may be relevant as well.

work has connected an accurate (faithful) attribution of causality with increased explainability (Feder et al., 2020; Madumal et al., 2020).

Additionally, **social attribution is missing:** Heider and Simmel (1944) describe an experiment where participants attribute human concepts of emotion, intentionality, and behavior to animated shapes. Clearly, the same phenomenon persists when humans attempt to understand the predictions of artificial models: We naturally attribute social intent to artificial decisions.

Can models be constrained to adhere to this attribution? Informally, prior work on highlights has considered such factors before. Lei et al. (2016) describe desiderata for highlights as being ''short and consecutive'', and Jain et al. (2020) interpreted ''short'' as ''around the same length as that of human-annotated highlights''. We assert that the nature of these claims is an attempt to constrain highlights to the social behavior implicitly attributed to them by human observers in the select-predict paradigm (discussed later).

7 (Socially) Aligned Faithfulness

Unlike human explainers, artificial explainers can exhibit a misalignment between the causal chain behind a decision and the social attribution attributed to it. This is because the artificial decision process may not resemble human behavior.

By presenting to the user the causal pipeline of decisions in the model's decision process as an interpretation of this process, the user naturally conjures social intent behind this pipeline. In order to be considered comprehensible to the user, the attributed social intent must match the actual behavior of the model. This can be formalized as a set of constraints on the possible range of decisions at each step of the causal chain.

We claim that this problem is the root cause behind the symptoms in Section 4. Here we define the general problem independently from the narrative of highlight interpretations.

Definition. We say that an interpretation method is *faithful* if it accurately describes causal information about the decision process of the decision. We say that the faithful method is *human-aligned* (short for ''aligned with human expectations of social intent'') if the model and method adhere to the social attribution of intent by human observers.

Conversely, ''*misaligned*'' interpretations are interpretations whose mechanism of conveying causal information is different from the mechanism that the user utilizes to glean causal information from the interpretation, where the mechanism is defined by the user's social attribution of intent towards the model. Furthermore, we claim that this attribution (expected intent) is heavily dependent on the model's task and use-case: Different use cases may call for different alignments.

#	Claim
1	The marked text is supportive of the decision.
2	The marked text is selected after making the decision.
3	The marked text is sufficient for making the decision.
4	The marked text is selected irrespective of the decision.
5	The marked text is selected prior to making the decision.
6	The marked text includes all the information that informed the decision.
7	The marking pattern alone is sufficient for making the decision by the predictor.
8	The marked text provides no explanation whatsoever.

Table 5: A list of claims that attribute causality or social intent to the highlight selection.

8 Alignment of *Select-Predict* Highlight Interpretations

In contrast to the human-provided explanations, in the ML setup, our situation is unique in that we have control over the causal chain but not the social attribution. Therefore, the social attribution must *lead* the design of the causal chain. In other words, we argue that we must first identify the behavior expected of the decision process, and then constrain the process around it.

8.1 Attribution of Highlight Interpretations

In order to understand what ''went wrong'' in the failure cases above, we need to understand what are *possible* expectations—potential instances of social attribution—to the ''rationalizing'' select-predict models. Table 5 outlines possible claims that could be attributed to a highlight explanation. Claims 1–6 are claims that could reasonably be attributed to highlights, while claims 7 and 8 are *not* likely to manifest.

These claims can be packaged as two high-level ''behaviors'':

***Summarizing* (3–6)**, where the highlight serves as an extractive summary of the most important and useful parts of the complete text. The highlight is merely considered a compression of the text, with sufficient information to make informed decisions

in a different context, towards some concrete goal. It is not selected with an answer in mind, but in anticipation that an answer will be derived in the future, for a question that has not been asked yet. And ***evidencing*** **(1–3)**, in which the highlight serves as supporting evidence towards a *prior* decision that was not necessarily restricted to the highlighted text.

Claims 7–8 are representative of the examples of failure cases discussed in Section 4.

8.2 Issues with *Select-Predict*

We argue that select-predict is inherently misleading. **Although claims 1–6 are plausible attributions to select-predict highlights, *none of them can be guaranteed* by a select-predict system**, in particular for systems in which the selector and predictor are exposed to the end task during training.

If the select-predict system acts ''as intended'', selection happens before prediction, which is incompatible with claims 1–2. However, as we do not have control over the selector component, it cannot be guaranteed that the selector will not perform an implicit decision prior to its selection, and once a selector makes an implicit decision, the selected text becomes disconnected from the explanation. For example, the selector decided on class A, and chose span B because it ''knows'' this span will cause the predictor to predict class A (see Section 4.2).

In other words, the advertised select-predict chain may implicitly become a ''predict-select-predict'' chain. The first and hidden prediction step makes the final prediction step disconnected from the cause of the model's decision, because the second prediction is conditioned on the first one. This invalidates attributions 3–6. It also allows for 7–8.

Select-predict models are closest to the characteristics of highlights selected as summaries by humans—therefore they can theoretically be aligned with summary attribution if the selector is designed as a truly summarizing component, and has no access to the end-task. This is hard to achieve, and no current model has this property.

The issues of Section 4 are direct results of the above conflation of interests: Trojan highlights and dominant selectors result from a selector that makes hidden and unintended decisions, so they serve as neither summary nor evidence towards the predictor's decision. Loss of performance is due to the selector acting as an imperfect summarizer—whether summary is relevant to the task to begin with, or not (as is the case in agreement classification, or natural language inference).

9 *Predict-Select-Verify*

We propose the *predict-select-verify* causal chain as a solution that can be constrained to provide highlights as evidence (i.e., guarantee claims 1–3). This framework solves the misalignment problem by allowing the derivation of faithful highlights aligned with evidencing social attribution.

The decision pipeline is as follows:

1. The predictor m_p makes a prediction $\hat{y} := m_p(x)$ on the full text.

2. The selector m_s selects $h := m_s(x)$ such that $m_p(h \odot x) = \hat{y}$.

In this framework, the selector provides evidence that is verified to be useful to the predictor towards a particular decision. Importantly, the final decision has been made on the full text, and the selector is constrained to provide a highlight that adheres to this exact decision. The selector does not purport to provide a highlight which is comprehensive of all evidence considered by the predictor, but it provides a *guarantee* that the highlighted text is supportive of the decision.

Causal Attribution. The selector highlights are provably faithful to the predict-select-verify chain of actions. They can be said to be *faithful by construction* (Jain et al., 2020), similarly to how select-predict highlights are considered faithful—the models undergo the precise chain of actions that is attributed to their highlights.

Social Attribution. The term ''rationalization'' fits the current causal chain, unlike in select-predict, and so there is no misalignment: The derived highlights adhere to the properties of highlights as evidence described in Section 8.1. The highlight selection is made under constraints that the highlight serve the predictor's prior decision, which is *not* caused by the highlighted text. The constraints are then verified at the *verify* step.

Solving the Failure Cases (§4). As a natural but important by-product result of the above, predict-select-verify addresses the failures of Section 4:

Trojan highlights and dominant selectors are impossible, as the selector is constrained to only provide ''retroactive'' selections towards a specific priory-decided prediction. The selector cannot cause the decision, because it was made without its intervention. Finally, the highlights inherently cannot cause loss of performance, because they merely support a decision which was made based on the full text.

10 Constructing a *Predict-Select-Verify* Model with Contrastive Explanations

In order to design a model adhering to the predict-select-verify chain, we require solutions for the predictor and for the selector.

The **predictor** is constrained to be able to accept both full-text inputs and highlighted inputs. For this reason, we use masked language modeling (MLM) models, such as BERT (Devlin et al., 2018), fine-tuned on the downstream task. The MLM pre-training is performed by recovering partially masked text, which conveniently suits our needs. We additionally provide randomly highlighted inputs to the model during fine-tuning.

The **selector** is constrained to select highlights for which the predictor made the same decision as it did on the full text. However, there are likely many possible choices that the selector may make under these constraints, as there are many possible highlights that all result in the same decision by the predictor. **We wish for the selector to select *meaningful* evidence to the predictor's decision.**[11] What is meaningful evidence? To answer this question, we again refer to cognitive science on necessary attributes of explanations that are easy to comprehend by humans. We stress that selecting meaningful evidence is critical for predict-select-verify to be useful.

10.1 Contrastive Explanations

An especially relevant observation in the social science literature is of *contrastive explanations* (Miller, 2019, 2020), following the notion that the question ''why P?'' is followed by an addendum: ''why P, rather than Q?'' (Hilton, 1988). We refer to P as the *fact* and Q as the *foil* (Lipton, 1990). The concrete valuation in the community is that in the vast majority of cases, the cognitive burden of a ''complete'' explanation, that is, where Q is $\overline{P}$, is too great, and thus Q is selected as a subset of all possible foils (Hilton and Slugoski, 1986; Hesslow, 1988), and often not explicitly, but implicitly derived from context.

For example, ''Elmo drank the water because he was thirsty,'' explains the fact ''Elmo drank water'' without mentioning a foil. But while this explanation is acceptable if the foil is ''not drinking'', it is not acceptable if the foil is ''drinking tea'': ''Elmo drank the water *(rather than the tea)* because he was thirsty.'' Similarly, ''Elmo drank the water because he hates tea'' only answers the latter foil. The foil is implicit in both cases, but nevertheless it is not $\overline{P}$, but only a subset.

In classification tasks, the implication is that an interpretation of a prediction of a specific class is hard to understand, and should be contextualized by *the preference of the class over another*—and the selection of the foil (the non-predicted class) is non-trivial, and a subject of ongoing discussion even in human explanations literature.

Contrastive explanations have many implications for explanations in AI as a vehicle for explanations that are easy to understand. Although there is a modest body of work on contrastive explanations in machine learning (Dhurandhar et al., 2018; Chakraborti et al., 2019; Chen et al., 2020), to our knowledge, the NLP community seldom discusses this format.

10.2 Contrastive Highlights

An explanation in a classification setting should not only addresses the fact (predicted class), but do so against a foil (some other class).[12] Given classes c and $\hat{c}$, where $m_p(x) = c$, we will derive a contrast explanation towards the question: ''why did you choose c, rather than $\hat{c}$?''.

We assume a scenario where, having observed c, the user is aware of some highlight h which should serve, they believe, as evidence for class $\hat{c}$. In other words, we assume the user believes a pipeline where $m_p(x) = \hat{c}$ and $m_s(x) = h$ is reasonable.

[11] For example, the word ''nice'' in Table 3a is likely not useful supporting evidence, since it is a rather trivial claim, even in a predict-select-verify setup.

[12] Selecting the foil, or selecting what to explain, is a difficult and interesting problem even in philosophical literature (Hesslow, 1988; Mcgill and Klein, 1993; Chin-Parker and Cantelon, 2017). In the classification setting, it is relatively simple, as we may request the foil (class) from the user, or provide separate contrastive explanations for each foil.

Procedure	Text and Highlight	Label y	Prediction $m_p(x)$	Foil Prediction $m_p(h \odot x)$	Contrast Prediction $m_p(h_c \odot x)$
Manual	Ohio Sues Best Buy, Alleging Used Sales (AP): AP - Ohio authorities **sued** Best Buy Co. Inc. on Thursday, alleging the electronics retailer engaged in unfair and deceptive business practices .	Business	Sci/Tech	Business	Sci/Tech
Manual	HK Disneyland Theme Park to Open in September: Hong Kong's Disneyland theme park will open on Sept. 12, 2005 and become the driving force for growth in the city's tourism industry , Hong Kong's **government** and Walt Disney Co.	Business	World	Business	World
Manual	Poor? Who's poor? Poverty is down: The proportion of people living on less than **$**1 a day decreased from 40 to 21 percent of the global population between 1981 and 2001, says the World Bank's latest annual report.	World	Business	World	Business
Manual	**Poor**? Who's poor? Poverty is down: The proportion of people living on less than $1 a day decreased from 40 to 21 percent of the global population between 1981 and 2001, says the World Bank's latest annual report.	World	Business	World	Business
Manual	Poor? Who's poor? Poverty is down: The proportion of people living on less than $1 a day **decreased** from 40 to 21 percent of the global population between 1981 and 2001, says the World Bank's latest annual report.	World	Business	World	Business
Automatic	**Siemens** Says Cellphone Flaw May Hurt Users and Its Profit: Siemens, the world's fourth-largest maker of mobile phones, said Friday that a software flaw that can create a piercing ring in its newest phone models might hurt earnings in its handset division.	Business	Sci/Tech	Business	Sci/Tech
Automatic	Siemens Says **Cell**phone Flaw May Hurt Users and Its Profit: Siemens the world's fourth-largest maker of mobile phones, said Friday that a software flaw that can create a piercing ring in its newest phone models might hurt earnings in its handset division.	Business	Sci/Tech	Business	Sci/Tech
Automatic	Siemens Says Cellphone Flaw May Hurt **Users** and Its Profit: Siemens the world's fourth-largest maker of mobile phones, said Friday that a software flaw that can create a piercing ring in its newest phone models might hurt earnings in its handset division.	Business	Sci/Tech	Business	Sci/Tech
Automatic	Siemens Says Cellphone Flaw May Hurt Users and Its Profit: **Siemens**, the world's fourth-largest maker of mobile phones, said Friday that a software flaw that can create a piercing ring in its newest phone models might hurt earnings in its handset division.	Business	Sci/Tech	Business	Sci/Tech

Table 6: Examples of contrastive highlights (§10) of instances from the AG News corpus. The model used for m_p is fine-tuned bert-base-cased. The foil highlight h is in standard yellow; the contrastive delta h_Δ is in **bold yellow**; and $h_c := h + h_\Delta$. All examples are cases of model errors, and the foil was chosen as the gold label.

If $m_p(h \odot x) \neq \hat{c}$, then the user is made aware that the predictor disagrees that h serves as evidence for $\hat{c}$.

Otherwise, $m_p(h \odot x) = \hat{c}$. We define:

$$h_c := \operatorname*{argmin}_{\substack{h+h_\Delta \\ \text{s.t. } |h_\Delta|>0 \\ \wedge\, m_p((h+h_\Delta)\odot x)=c}} |h + h_\Delta|.$$

h_c is the minimal highlight containing h such that $m_p(h_c \odot x) = c$. Intuitively, the claim by the model is as such: ''I consider h_Δ as a sufficient change from h (evidence to $\hat{c}$) to h_c so that it is evidence towards c.''

The final manual procedure is, given a model m_p and input x:

1. The user observes $m_p(x)$ and chooses a relevant foil $\hat{c} \neq m_p(x)$.
2. The user chooses a highlight h which they believe supports $\hat{c}$.
3. If $m_p(h \odot x) \neq \hat{c}$, the shortest h_c is derived such that $h \subset h_c$ and $m_p(h_c \odot x) = m(x)$ by brute-force search.

Automation. Although we believe the above procedure is most useful and informative, we acknowledge the need for automation of it to ease the explanation process. Steps 1 and 2 involve human input which can be automated: In place of step 1, we may simply repeat the procedure separately for each of all foils (and if there are too many to display, select them with some priority and ability to switch between them after-the-fact); and in place of step 2, we may heuristically derive candidates for h—e.g., the longest highlight for which the model predicts the foil:

$$h := \underset{\substack{h \\ m_p(h \odot x) = \hat{c}}}{\arg\max} |h|.$$

The automatic procedure is then, for each class $\hat{c} \neq m_p(x)$:

1. Candidates for h are derived, for example, the longest highlight h for which $m_p(h \odot x) = \hat{c}$.
2. The shortest h_c is derived such that $h \subset h_c$ and $m_p(h_c \odot x) = m_p(x)$.

We show examples of both procedures in Table 6 on examples from the AG News dataset. For illustration purposes, we selected incorrectly classified examples, and selected the foil to be the true label of the example. The highlight for the foil was chosen by us in the manual examples.

In the automatic example, the model made an incorrect *Sci/Tech* prediction on a *Business* example. The procedure reveals that the model would have made the correct prediction if the body of the news article was given without its title, and that the words ''Siemens'', ''Cell'', and ''Users'' in the title are independently sufficient to flip the prediction on the highlight from *Business* to *Sci/Tech*.

We stress that while the examples presented in these figures appear reasonable, the true goal of this method is not to provide highlights that seem justified, but to provide a framework that allows models to be meaningfully incorporated in use-cases of *dispute*, *debug*, and *advice*, with robust and proven guarantees of behavior.

For example, in each of the example use-cases:
Dispute: The user verifies if the model ''correctly'' considered a specific portion of the input in the decision: The model made decision c, where the user believes decision $\hat{c}$ is appropriate and is supported by evidence $h \odot x$. If $m_p(h \odot x) \neq c$, they may dispute the claim that the model interpreted $h \odot x$ with ''correct'' evidence intent. Otherwise the dispute cannot be made, as the model provably considered h as evidence for $\hat{c}$, yet insufficiently so when combined with h_Δ as $h_c \odot x$.
Debug: Assuming c is incorrect, the user performs error analysis by observing which part of the input is sufficient to steer the predictor away from the correct decision $\hat{c}$. This is provided by h_Δ.
Advice: When the user is unaware of the answer, and is seeking perspective from a trustworthy model: They are given explicit feedback on which part of the input the model ''believes'' is sufficient to overturn the signal in h towards $\hat{c}$. If the model is not considered trustworthy, the user may gain or reduce trust by observing whether $m(h \odot x)$ and h_Δ align with user priors.

11 Discussion

Causal Attribution of Heat-maps. Recent work on the faithfulness of attention heat-maps (Baan et al., 2019; Pruthi et al., 2019; Serrano and Smith, 2019) or saliency distributions (Alvarez-Melis and Jaakkola, 2018; Kindermans et al., 2019) cast doubt on their faithfulness as indicators to the significance of parts of the input (to a model decision). Similar arguments can be made regarding any explanation in the format of heat-maps, such as LIME and SHAP (Jacovi and Goldberg, 2020). We argue that this is a natural conclusion from the fact that, as a community, we have not envisioned an appropriate causal chain that utilizes heat-maps in the decision process, reinforcing the claims in this work on the parallel between causal attribution and faithfulness. This point is also discussed at length by Grimsley et al. (2020).

Social Attribution of Heat-maps. As mentioned above, the lack of a clear perception of a causal chain behind heat-map feature attribution explanations in NLP makes it difficult to discuss the social intent attributed by these methods. Nevertheless, it is possible to do so under two perspectives: (1) when the heat-map is discretized into a highlight, and thus can be analyzed along the list of possible attributions in Table 5; or (2) when the heat-map is regarded as a collection of pair-wise claims about which part the input is more important, given two possibilities. Perspective (1) can be likened to claims #1 and #2 in Table 5, namely, ''evidencing'' attributions sans sufficiency.

Contrastive Explanations in NLP. We are not aware of prior work that discusses or implements contrastive explanations explicitly in NLP, however this does not imply that existing explanation methods in NLP are not contrastive. To the contrary, the social sciences argue that *every* manner of explanation has a foil, and is comprehended by the explainee against some foil—including popular methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). The question then becomes *what* the foil is, and *whether this foil is intuitive* and thus useful. In the case of LIME, for example, the foil is defined by the aggregation of all possible perturbations admitted to the approximating linear model—where such perturbations may not be natural language, and thus less intuitive as foil; additionally, Kumar et al. (2020) have recently derived the foil behind general shapley value-based explanations, and have shown that this foil is not entirely aligned with human intuition. *We argue that making the foil explicit and intuitive is an important goal of any interpretation system.*

Inter-disciplinary Research. Research on explanations in artificial intelligence will benefit from a deeper interdisciplinary perspective on two axes: (1) literature on causality and causal attribution, regarding causal effects in a model's reasoning process; and (2) literature on the social perception and attribution of human-like intent to causal chains of model decisions or behavior.

12 Related Work

How interpretations are comprehended by people is related to *simulatability* (Kim et al., 2017)—the degree to which humans can simulate model decisions. Quantifying simulatability (Hase and Bansal, 2020) is decidedly different from social alignment, since, for example, it will not necessarily detect dominant selectors. We theorize that aligned faithful interpretations will increase simulatability.

Predict-select-verify is reminiscent of iterative erasure (Feng et al., 2018). By iteratively removing ''significant'' tokens in the input, Feng et al. show that a surprisingly small portion of the input could be interpreted as evidence for the model to make the prediction, leading to conclusions on the pathological nature of neural models and their sensitivity to badly-structured text. This experiment retroactively serves as a successful application of *debugging* using our formulation.

The approach by Chang et al. (2019) for class-wise highlights is reminiscent of contrastive highlights, but nevertheless distinct, since such highlights still explain a fact against all foils.

13 Conclusion

Highlights are a popular format for explanations of decisions on textual inputs, for which there are models available today with the ability to derive highlights ''faithfully''. We analyze highlights as a case study in pursuit of rigorous formalization of quality artificial intelligence explanations.

We redefine *faithfulness* as the accurate representation of the causal chain of decision making in the model, and *aligned faithfulness* as a faithful interpretation which is also aligned to the social attribution of intent behind the causal chain.

The two steps of causal attribution and social attribution *together* complete the process of ''explaining'' the decision process of the model to humans.

With this formalization, we characterize various failures in faithful highlights that ''seem'' strange, but could not be properly described previously, noting they are not properly constrained by their social attribution as summaries or evidence. We propose an alternative which can be constrained to serve as evidence. Finally, we implement our alternative by formalizing *contrastive explanations* in the highlight format.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation program, grant agreement no. 802774 (iEXTRACT).

References

David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the robustness of interpretability methods. *CoRR*, abs/1806.08049.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. *CoRR*, abs/1601.01705. **DOI:** https://doi.org/10.18653/v1/N16-1181

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke.

2019. Do transformer attention heads provide transparency in abstractive summarization? *CoRR*, abs/1907.00570.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1903–1913. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1216, **PMID:** 29551352

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977. Association for Computational Linguistics, Florence, Italy. **DOI:** https://doi.org/10.18653/v1/P19-1284

Tapabrata Chakraborti, Arijit Patra, and J. Alison Noble. 2019. Contrastive algorithmic fairness: Part 1 (theory). *ArXiv*, abs/1905.07360.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2019. A game theoretic approach to class-wise selective rationalization. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 10055–10065.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2020. Invariant rationalization. *CoRR*, abs/2003.09772.

Sheng-Hui Chen, Kayla Boggess, and Lu Feng. 2020. Towards transparent robotic planning via contrastive explanations. *ArXiv*, abs/2003.07425.

Seth Chin-Parker and Julie Cantelon. 2017. Contrastive constraints guide explanation-based category learning. *Cognitive Science*, 41(6): 1645–1655. **DOI:** https://doi.org/10.1111/cogs.12405, **PMID:** 27564059

Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 510–518. Association for Computing Machinery, New York, NY, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. Eraser: A benchmark to evaluate rationalized NLP models. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.408

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. Causalm: Causal model explanation through counterfactual language models. *CoRR*, abs/2005.13407.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan L. Boyd-Graber. 2018. Pathologies of neural models make interpretation difficult. Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3719–3728. Association for Computational Linguistics, **DOI:** https://doi.org/10.18653/v1/D18-1407

Christopher Grimsley, Elijah Mayfield, and Julia R. S. Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1780–1790. European Language Resources Association, Marseille, France.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *CoRR*, abs/2005.01831. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.491

Fritz Heider, American Psychological Association, and Ovid Technologies Inc. 1958. *The Psychology of Interpersonal Relations*. New York: Wiley ; Hillsdale, NJ: Lawrence Erlbaum.

Fritz Heider and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2): 243–259. **DOI:** https://doi.org/10.2307/1416950

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *CoRR*, abs/1711.07414. Withdrawn.

Germund Hesslow. 1988, The problem of causal selection, Denis J. Hilton, editor, *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, New York University Press.

Hilton, Denis J. 1988. Logic and causal attribution. In *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*. 33–65, New York University Press.

Denis J. Hilton, 1966 Mandel, David R., and Patrizia Catellani. 2005. *The Psychology of Counterfactual Thinking*, London ; New York : Routledge. Includes bibliographical references (p. [217]-244) and indexes.

Denis J. Hilton and Ben R. Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1):75. **DOI:** https://doi.org/10.1037/0033-295X.93.1.75

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *CoRR*, abs/2004.03685. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.386

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. *CoRR*, abs/2005.00115. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.409

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019, The (un)reliability of saliency methods, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, 11700, Springer, pages 267–280. **DOI:** https://doi.org/10.1007/978-3-030-28954-6_14

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. Yoshua Bengio and Yann LeCun, editors, In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. 2020. Problems with shapley-value-based explanations as feature importance measures. *CoRR*, abs/2002.11097.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. *CoRR*, abs/1606.04155. **DOI:** https://doi.org/10.18653/v1/D16-1011

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*:247–266. **DOI:** https://doi.org/10.1017/S1358246100005130

Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of ACM*, 61(10): 36–43. **DOI:** https://doi.org/10.1145/3233231

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470. **DOI:** https://doi.org/10.1016/j.tics.2006.08.004, **PMID:** 16942895

Scott M. Lundberg and Su-In Lee. 2017, A unified approach to interpreting model predictions,

I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pages 4765–4774.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2493–2500. AAAI Press. **DOI:** https://doi.org/10.1609/aaai.v34i03.5631

Ann Mcgill and Jill Klein. 1993. Contrastive and counterfactual thinking in causal judgment. *Journal of Personality and Social Psychology*, *64*:897–905. **DOI:** https://doi.org/10.1037/0022-3514.64.6.897

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*:1–38. **DOI:** https://doi.org/10.1016/j.artint.2018.07.007

Tim Miller. 2020. Contrastive explanation: A structural-model approach.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2019. Learning to deceive with attention-based explanations. *CoRR*, abs/1909.07913. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.432

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ‘‘Why Should I Trust You?’’: Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 1135–1144. ACM, New York, NY, USA. **DOI:** https://doi.org/10.1145/2939672.2939778

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. **DOI:** https://doi.org/10.1038/s42256-019-0048-x

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2931–2951.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684. Association for Computational Linguistics, Florence, Italy. **DOI:** https://doi.org/10.18653/v1/P19-1164

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining faithful interpretations from compositional neural networks. *CoRR*, abs/2005.00724. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.495

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, 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 2019, Hong Kong, China, November 3-7, 2019*, pages 11–20. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1002

Joseph Jay Williams, Tania Lombrozo, and Bob Rehder. 2013. The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, 142(4):1006. **DOI:** https://doi.org/10.1037/a0030996, **PMID:** 23294346

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256. **DOI:** https://doi.org/10.1007/BF00992696

Omar Zaidan, Jason Eisner, and Christine D. Piatko. 2007. Using ‘‘annotator rationales’’ to

improve machine learning for text categorization. Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 260–267. The Association for Computational Linguistics.

A Glossary

This work is concerned with formalization and theory of artificial models' explanations. We provide a (non-alphabetical) summary of terminology and their definitions as we utilize them. We stress that these definitions are *not* universal, as the human explanation sciences describe multiple distinct perspectives, and explanations in AI are still a new field.

Unfolding causal chain: A path of causes between a set of events, in which a cause from event C to event E indicates that C must occur before E.

Human intent: An objective behind an action. In our context, reasoning steps in the causal chain are actions that can be attributed with intent.

Interpretation: A (possibly lossy) mapping from the full reasoning process of the model to a human-readable format, involving some implication of a causal chain of events in the reasoning process.

Faithful interpretation: An interpretation is said to be faithful if the causal chain it describes is accurate to the model's full reasoning process.

Explanation: A process of conveying causal information about a model's decision to a person. We assume that the explainee always attributes intent to the actions of the explainer.

Plausibility: Incentive of the explainer to provide justifying explanation that appears convincing.