On the Interaction of Belief Bias and Explanations

Ana Valeria González[1], **Anna Rogers**[2], and **Anders Søgaard**[1]

University of Copenhagen
[1]Department of Computer Science
[2]Copenhagen Centre for Social Data Science
{ana,soegaard}@di.ku.dk
arogers@sodas.ku.dk

Abstract

A myriad of explainability methods have been proposed in recent years, but there is little consensus on how to evaluate them. While automatic metrics allow for quick benchmarking, it isn't clear how such metrics reflect human interaction with explanations. Human evaluation is of paramount importance, but previous protocols fail to account for *belief biases* affecting human performance, which may lead to misleading conclusions. We provide an overview of belief bias, its role in human evaluation, and ideas for NLP practitioners on how to account for it. For two experimental paradigms, we present a case study of gradient-based explainability introducing simple ways to account for humans' prior beliefs: models of varying quality and adversarial examples. We show that *conclusions about the highest performing methods change when introducing such controls*, pointing to the importance of accounting for belief bias in evaluation.

1 Introduction

Machine learning has become an integral part of our lives; from everyday use (e.g., search, translation, recommendations) to high-stake applications in healthcare, law, or transportation. However, its impact is controversial: neural models have been shown to make confident predictions relying on artifacts (McCoy et al., 2019; Wallace et al., 2019) and have shown to encode and amplify negative social biases (Manzini et al., 2019; Caliskan et al., 2017; May et al., 2019; Tan and Celis, 2019; González et al., 2020; Rudinger et al., 2018).

Explainability aims to make model decisions transparent and predictable to humans; it serves as a tool for model diagnosis, detecting failure modes and biases, and more generally, to increase trust by providing transparency (Amershi et al., 2019). While automatic metrics have been proposed to evaluate various properties of explanations such as faithfulness, consistency and agreement with human explanations (Atanasova et al., 2020; Robnik-Šikonja and Bohanec, 2018; DeYoung et al., 2020), these metrics do not inform us about human interaction with explanations.

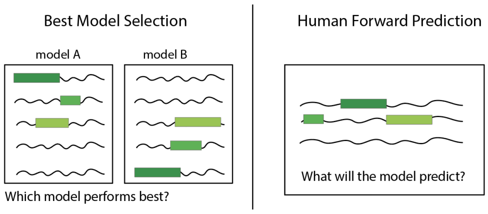


Figure 1: Evaluation protocols considered in this work

Doshi-Velez and Kim (2017) suggested *human forward prediction*, a simulation task in which humans are given an input and an explanation, and their task is to predict the expected model output, regardless of the gold answer. Recent studies include Nguyen (2018); Lage et al. (2019); **?**); Poursabzi-Sangdeh et al. (2021). Such protocols are widely used and can provide valuable insight into human understanding of explanations. However, prior work has not accounted for how humans' prior beliefs (*belief biases*) interact with the evaluation; simulating model decisions becomes an easier task when the model being evaluated makes predictions which align with human expectations. We argue that not considering belief bias in such protocols *may lead to misleading conclusions about which explainability methods perform best.*

Other protocols have evaluated participant's ability to select the best model based on explanations offered by different interpretability methods (e.g. decide which model would generalize 'in the wild') (Ribeiro et al., 2016a). However, comparisons have been made between a model which is clearly in line with human beliefs, and another which exploits spurious correlations diverging from human expec-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2930–2942
August 1–6, 2021. ©2021 Association for Computational Linguistics

tations. When differences are less obvious, humans may not be able to leverage their belief biases, and conclusions may change.

This paper, which includes evaluations for both of the previously mentioned tasks, closes an important gap: to the best of our knowledge, no prior work in NLP addresses the interaction of belief bias with current human evaluations of explainability.

Contributions. We provide an overview of belief bias meant to highlight its role in human evaluation and provide some preliminary ideas for NLP practitioners on how to handle such cases. Using *human forward prediction* and *best model selection* (Figure 1), we present a case-study where we compare two gradient-based explainability methods in the context of reading comprehension (RC), introducing conditions to take into account belief bias. We find that both explainability methods are helpful to participants in the standard settings (in line with most previous work), but the *conclusions about the best performing models change when incorporating additional control conditions*, reinforcing the importance of accounting for such biases.

2 Belief Bias

Belief bias is a type of cognitive bias, defined in psychology as *the systematic (non-logical) tendency to evaluate a statement on the basis of prior belief rather than its logical strength* (Evans et al., 1983; Klauer et al., 2000; Barston, 1986). Cognitive biases are not necessarily bad; they help us filter and process a great deal of information (Bierema et al., 2020), and have been widely studied in real human-decision making (Tversky and Kahneman, 1974; Kahneman, 2003; Furnham and Boo, 2011). However, in evaluations involving human participants, such biases may alter results and affect conclusions (Anderson and Hartzler, 2014; Wall et al., 2017).

Classic psychology studies of belief bias have assessed how prior beliefs affect syllogistic reasoning (Newstead et al., 1992; Klauer et al., 2000; Evans et al., 1983; Markovits and Nantel, 1989; Evans and BT). Consider the following example by Anderson and Hartzler (2014):

(a) *If all birds are animals, and if no animals can fly, then no birds can fly.*

(b) *If all cats are animals, and if no animals can fly, then no cats can fly.*

In syllogistic reasoning, the task for humans is to assess the *logical* validity of such arguments while ignoring believability. While both arguments are logically valid, most work converges on the finding that humans will rate argument (a) as invalid more often than (b), biased by the fact that the premise in (a) is less believable.

In psychology, belief bias has been tied to the *dual-processing* theory, which assumes that reasoning is performed by two competing cognitive systems: (1) *system 1* which takes care of fast, heuristic processes and (2) *system 2* which handles slower, more analytical processes (Evans, 2003; Trippas and Handley, 2018; Evans and Curtis-Holmes, 2005; Croskerry, 2009). Generally, humans tend to have a cognitive preference for relying on fast, intuitive *system 1* processes, rather than engaging in the slow and more analytical *system 2* processes. Belief bias is attributed to system 1 (Evans and Curtis-Holmes, 2005; Evans, 2008; Evans and Frankish, 2009; Stanovich and West, 2008) due to several factors, reviewed in detail by Evans (2003); Caravona et al. (2019).

For the purposes of NLP studies relying on crowd workers, one relevant finding is that **time pressures exacerbate reliance on previous beliefs** (Evans and Curtis-Holmes, 2005). Since crowd workers generally are incentivized to work as quickly as possible to maximize their hourly pay, reliance on belief bias is to be expected.

Another relevant finding for NLP is that threatening or negatively charged arguments (e.g. content violating political correctness and social norms) leads to greater engagement of system 2, whereas **neutral content leads to increased reliance on belief bias** (Goel and Vartanian, 2011; Klaczynski et al., 1997). Since NLP studies tend to be performed on neutral content such as passages from Wikipedia – content which may not sufficiently engage participants' system 2 processes – belief bias is more likely to play a role in human performance.

This study aims to highlight the phenomenon of belief bias to encourage NLP practitioners to assess the role it plays in their evaluations, and introduce mechanisms to account for belief bias effects. We illustrate how belief bias effects can significantly affect the results of human evaluation of explainability for two paradigms: *human forward prediction* and *best model selection*.

3 Related Work

Human forward prediction. Human forward prediction experiments have been recently presented in the context of synthetic data (Poursabzi-Sangdeh et al., 2021; Lage et al., 2019; Slack et al., 2019) to evaluate explainability methods for *their ability to make model decisions predictable to humans*. In this paradigm, humans are presented with explanations and tasked with predicting the model's decision regardless of the ground truth (Doshi-Velez and Kim, 2017).[1]

In NLP, Nguyen (2018) introduced human forward prediction for LIME explanations (Ribeiro et al., 2016b) of sentiment analysis of product reviews and correlated the results with automatic evaluations. Unlike with synthetic data, participants have prior beliefs on what the *true* outcome is. Since participants in Nguyen (2018) had no training phase to learn how explanations correlate with predictions and the model being evaluated sufficiently matched human behavior, humans likely relied *exclusively* on their prior knowledge and beliefs to complete the task at hand.

? improved on this protocol by adding a training phase. This is something we also do in our experiments (section 5), but it is unlikely to solve the belief bias problem because even after training, humans will naturally opt for fast, heuristic mechanisms (e.g. belief bias) in order to simplify tasks (Wang et al., 2019); this is particularly true if the model is high performing (i.e. likely aligns with human beliefs).

The protocol by **?** had another key feature: they leave out the explanations for the test data points. This would seem like an advantage for evaluating explainability methods in the context of reading comprehension where explanations can, in theory, simply highlight the answer span, making it easy to guess the model output from the explanations. However, it is easy to control for the amount of explanation provided by the explanation methods we compare; in our experiments below, we highlight the top 10 tokens with highest attribution scores. This key feature in their protocol is problematic for two reasons:

- It makes the human learning problem much harder, and we argue it is infeasible to expose participants to enough examples to make human forward prediction learnable (unless the task is made very easy on purpose; again by only evaluating high performing models). If it is not learnable, participants fall back on belief bias.
- It introduces a systematic bias between the training and test scenarios.

The protocol in **?** also does not randomize the order in which participants are exposed to problems with or without explanations.

We improve on the above protocol by introducing a condition which can help account for belief bias effects: evaluating explainability methods on low-quality models, the predictions of which substantially differ from human beliefs. This means that in order to succeed in the task, humans cannot simply rely on their previous beliefs, therefore, helping us assess the ability of explanations in helping humans to *realign* their expectations of model behavior. The predictions of reading comprehension models can also be made different from human answers by introducing distractor sentences that fool machine reading models, but not humans (Jia and Liang, 2017). If in human forward prediction, participants predict the true answer rather than spans in the distractor sentences, this suggests participants may be relying on their belief biases.

Best model selection. Ribeiro et al. (2016b) presented an evaluation of explainability methods for text classification, where explanations for decisions of two different models on the same instance are presented side by side, and humans decide which model is likely to generalize better. With some exceptions (Lertvittayakumjorn and Toni, 2019), there has not been much follow up work on this task, but this scenario is important: it mimicks the decisions about what model is *safer for deployment*. Ribeiro et al. (2016b) and Lertvittayakumjorn and Toni (2019) both make a single comparison between a model which clearly diverges from human intuition, and a model that generalizes and *aligns with humans' beliefs*. Accounting for the extent to which belief biases are leveraged (e.g. by introducing additional model comparisons where differences are not so obvious or where models are of low quality) is important in such paradigms, and can allow us to better evaluate where explanation methods may fail.

In the following sections, we show that introducing conditions which take into account belief biases can have an effect on the conclusions for both *human forward prediction* and *best model se-*

[1] Using synthetic data from fictitious domains effectively controls for belief bias (Lage et al., 2019; Slack et al., 2019). Slack et al. (2019), for example, evaluate explanations in the domain of recommending recipes and medicines to aliens.

lection. We emphasize that many other potential strategies can be introduced and this is largely dependent on the goals of the evaluation protocol; we merely provide one example case with the following strategies:

(1) Introducing low quality models which considerably diverge from humans' prior beliefs (*human forward prediction*)

(2) Introducing evaluation problems with distractor sentences (*human forward prediction*)

(3) Introducing model comparisons where relying on belief bias is not enough to obtain high performance (*best model selection*)

4 Experimental Setup

This section introduces the general setup of the experiments, with details specific to each experimental paradigm described in section 5 and section 6.

4.1 Models

We evaluate explanations produced by three BERT-based (Devlin et al., 2019) models:

(a) a high performing model (**HIGH**): BERT-base, fine-tuned on SQuAD 2.0. *This model is more aligned with human beliefs.*

(b) a medium performing model (**MEDIUM**): tinyBERT, a 6-layer distilled version of BERT (Jiao et al., 2020), fine-tuned on SQuAD 2.0. It performs about 20 F_1 points below HIGH. *This model somewhat aligns with human intuition, but performs significantly lower.*

(c) a low performing model (**LOW**): BERT-base, fine-tuned to always choose the first occurrence of the last word of the question. This system mimicks a rule-based system[2]; however, we evaluate gradient-based methods requiring a neural model. *This model diverges significantly from human beliefs.*

4.2 Data

We use SQuAD 2.0 (Rajpurkar et al., 2018), a RC dataset consisting of 150k factoid question-answer pairs, with texts coming from Wikipedia articles. We opt for this data as it contains short passages that can be read by humans in a short time. In the human forward prediction experiments, we refer to experiments using this data as ORIG. As described in section 2, Wikipedia texts could by themselves induce people to rely on their belief bias, but this particular dataset allows us to also introduce controls for the bias: the adversarial version of the data (Jia and Liang, 2017), has been shown to distract models but not humans. This means that in order to perform the task with success, humans need disregard their belief biases, and in some cases align with distractor sentences. We refer to this data in our simulation experiments as ADV.

4.3 Explainability Methods

We focus on gradient-based approaches, as they require no modifications to the original network, and are considerably faster than perturbation-based methods. We compare two explainability methods:

Gradients. Computing the gradient of the prediction output with regard to the features of the input is a common way to interpret deep neural networks (Simonyan et al., 2013) and capture relevant information regarding the underlying model.

Integrated gradients. Integrated gradients approach (IG) (Sundararajan et al., 2017) attributes an importance score to each input feature by approximating the integral of gradients of the model's output with respect to the inputs along the path, from the references to the inputs. IG was introduced to address the sensitivity issues which are present in vanilla gradients and implementation invariance.

5 Experiment 1: Human Forward Prediction

Human forward prediction for evaluating explainability was proposed by Doshi-Velez and Kim (2017). They argue that if a human is able to simulate the model's behavior, they understand *why* the model predicts in that manner. For the reasons previously outlined, we suspect that belief biases may be affecting performance and the conclusions once can draw from this task. We investigate this by asking the following: *Can humans predict model decisions, if model behavior considerably diverges from their own beliefs?*

Stimuli presentation. We include: (i) HIGH, which is finetuned to solve SQuAD 2.0 and (ii) LOW, which is finetuned to select the first appearance in the context of the last word in the question. We evaluate each of the two models twice: with or without adversarial data. We contrast using

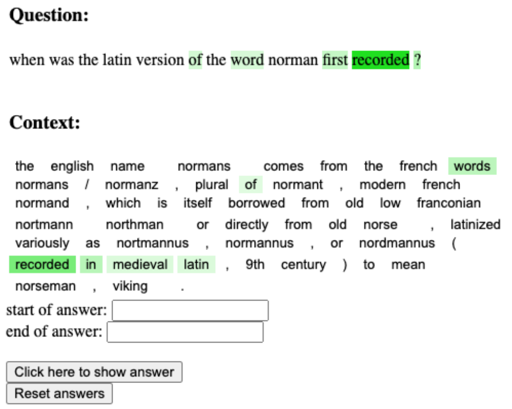
[2] This model achieves about 0.90 F1 for this task, but in the results we show its performance on the actual RC task

vanilla gradients and IG with a baseline condition, in which no explanations are shown (BASELINE).

We highlight the top-10 tokens[3] with the highest attribution scores wrt. the start and end positions of the predicted span, and zero out the rest.[4] The two sets of tokens often overlap.

Participants were provided with a question and a passage (with or without explanations) and were told to pick the *shortest* span of text which matched the model prediction. They saw the actual model answers before the next example (done for both baseline and explanation conditions), which was an important part of training to infer model behavior. Before the model prediction was shown, their answers were locked to prevent any further changes. An example of our interface can be found in Figure 2 and the instructions are shown in Appendix A.

We ran these experiments on Amazon Mechanical Turk, recruiting participants with approval ratings greater than 95%[5] and ensuring different groups of participants per condition by specifying that participation is only allowed once, otherwise risking rejection[6]. We paid participants \$5.25 for about 20 minutes of work (to ensure at least a \$15 hourly pay) and obtained at least three annotations per example. The data included 120 unique questions divided into small fixed batches (the same questions across conditions). About 75% of questions are accurate in the HIGH model, and around 15% are accurate for the LOW model. In total, we obtained 4,300 data points across 123 participants (35 data points per participant).



Question:

when was the latin version of the word norman first recorded ?

Context:

the english name normans comes from the french words normans / normanz , plural of normant , modern french normand , which is itself borrowed from old low franconian nortmann northman or directly from old norse , latinized variously as nortmannus , normannus , or nordmannus (recorded in medieval latin , 9th century) to mean norseman , viking .

start of answer:

end of answer:

Figure 2: Interface for Experiment 1 for LOW condition. To select model predictions, participants clicked on tokens to select the start and end of the span. Then they would see the actual model prediction.

Results. As humans often did not select the exact span that was provided as ground truth, we *manually* labeled the spans as correct or incorrect. We also inspected the impact of training in human forward prediction, e.g., the learning effect of multiple exposures on annotator accuracy. Both with vanilla gradients and integrated gradients, we observe an increase in the participants' accuracy at around 15 examples. In contrast, in our baseline condition, performance either stays constant or drops slightly. To reduce the noise introduced due to the training period, we remove the first 15 examples of each participant. The results without this preprocessing (Appendix A) suggest that **the effect of training differed across explainability methods**, as will be discussed later in the section.

Using the average human accuracies per example, we run a one-way ANOVA to test for significant differences across the groups. As we obtained statistically significant results, we then ran the Tukey honest significant difference (HSD) test (Tukey, 1949), comparing the means of every condition to the means of every other condition. The results are presented in Table 1.

As expected, in the absence of explanations (BASELINE), **humans rely on belief bias and predict the gold standard answer more often than the model prediction** (y in Table 1). Even with training (seeing the true model prediction), humans fail to catch onto the simple rule used by the LOW model, when no explanations are presented.

Overall, explanations derived from both of the gradient-based approaches lead to statistically significant improvements over the baseline. This indicates that the **explanations allow humans to realign their expectations of the model behavior**, better than with no explanations.

For HIGH-ORIG, the standard setting explored in previous evaluations, both IG gradients and vanilla gradients perform well, with IG gradients performing better. Given these results and the theoretical advantages of IG over vanilla gradients, one could arrive at the conclusion that IG are better for simulatability. However, **the differences between the two gradient-based methods are reversed in the conditions where humans cannot rely on their previous beliefs** (LOW). The gap be-

[3] Explanations should be *selective* (Mittelstadt et al., 2019)

[4] Ribeiro et al. (2016a) use the top 6 attributes; we opt for 10 given that our texts are slightly longer.

[5] Previous research has shown that proper filtering and selection of participants on Mechanical Turk, can be enough to ensure high quality data (Peer et al., 2014).

[6] We also remove such (few) repetitions at analysis

	MODEL	HUMAN		
CONDITION	F1	$\hat{y}$	y	SEC
		BASELINE		
LOW-ORIG	0.17	0.16	0.48	33.9
LOW-ADV	0.15	0.12	0.34	63.3
HIGH-ORIG	0.79	0.45	0.46	34.6
HIGH-ADV	0.66	0.38	0.48	36.1
		INTEGRATED (IG)		
LOW-ORIG	▨	*0.58	*0.22	*16.8
LOW-ADV		*0.63	*0.18	*22.3
HIGH-ORIG	▨	***0.84**	*0.88	36.1
HIGH-ADV		***0.52**	* 0.35	*18.9
		GRADIENTS		
LOW-ORIG	▨	***0.69**	*0.06	32.6
LOW-ADV		***0.72**	*0.15	*25.6
HIGH-ORIG	▨	*0.79	*0.81	47.4
HIGH-ADV		0.49	*0.60	48.4

Table 1: Human forward prediction results (HUMAN($\hat{y}$)) for LOW and HIGH models, compared to no explanations (BASELINE). Each experiment is run on vanilla SQuAD 2.0 data (ORIG) and adversarial SQuAD 2.0 data (ADV). HUMAN(y) is the dataset ground truth and an indicator of belief bias. Statistically significant results are indicated with an asterisk. Time is the average time per question. The best $\hat{y}$ results in each condition are bolded.

tween gradients and IG as large as 0.11, and being statistically significant. This finding is *surprising* and points again to the importance of not drawing incorrect conclusions about the best performing method using the standard paradigm.

Finally, in the HIGH conditions, model behavior decreases about 13% F1 score with the presence of **adversarial examples**, meaning that the model we used does get affected by adversarial inputs. We observe that human performance is considerably lower in HIGH-ADV as opposed to HIGH-ORIG. **With vanilla gradients, performance is more aligned with the ground truth labels than with model behavior,** showing that in this condition humans are also relying on their prior beliefs. **With IG, where performance is less aligned with prior beliefs (ground truth), the end performance increases**, but it seems that this condition is considerably more difficult for humans.

Effect of training. In BASELINE, training does not affect either the LOW or HIGH conditions (see Table 3 in Appendix A for the raw results). For the LOW model, multiple factors can be taking place (possibly at the same time): (1) the task is too far from the humans' beliefs and there is no mechanism to help participants realign their expectations, (2) participants may not be incentivized to seriously engage and look for patterns, (3) participants opt for a mixed strategy, where for some questions they go with their prior beliefs and for others, choice is random (as seen in their performance in y).

For HIGH conditions in BASELINE, performance remains higher than LOW but this is likely due to belief bias and not training, given that performance remains constant after removing the training data points. We hypothesize that for HIGH, instances where the model does not align to human intuition might be more detrimental than in explanation conditions. More specifically, if humans are aware that the model aligns with their beliefs after some examples but encounter instances where it doesn't (model is not 100% accurate), they will likely develop an expectation that the model is bound to make some errors, without any indication of when.

In addition, our raw results suggest IG required longer training. While this does not mean IG is a worse method than vanilla gradients, explanations derived from IG may have confused participants due to containing information which was irrelevant to them. It may be that experts (e.g. system engineers knowledgeable about neural networks) can take better advantage of such explanations; however, we leave this exploration of the interaction of human expertise with explanations as a direction for future work.

6 Experiment 2: Best Model Selection

This section presents the setup and results of our model selection experiments; a task where humans select the model that is more likely to succeed in the wild. We present the participants with the explanations from two models (HIGH vs LOW and HIGH vs MEDIUM), and ask them to decide which model is likely to perform better. As a follow-up, we also experimented with *soliciting explanations about what leads the worse model to fail*. Intuitively, comparative evaluation difficulty depends on how clear the difference is between the compared objects. Explanations should at least show the difference between a high-performing model and a low-performing one, enabling human participants to predict which is better (standard setting).

Stimuli presentation. We presented participants with saliency information from both models (a high performing model + one of the lower performing models), and their task was to determine which model performs best in the wild. We shuffled the order at random so that the best model would not remain in a fixed position. We obtain 120 samples (question-context pairs), and show the explanations next to each other as seen in Figure 1. The participants are told that the highlighted attributes are the words the model found important in making its decision. A screenshot of the UI is shown in Figure 4 in section B and the instructions provided to the participants are also shown in section B. These experiments were also ran on Amazon Mechanical Turk with the same general procedures and pay. The same subset of 120 examples is used in all conditions. We obtained at least three annotations per example and ended with a total of 1440 data points across 48 participants (30 examples each).

Results. For each example shown to annotators, we obtained the average accuracy scores and performed a standard T-test to compare the performance of the two methods. The results are shown in Table 2. Using explanations from both methods, when shown the HIGH and LOW model, humans are clearly able to correctly select the better one. With IG, humans achieve **0.95** accuracy on average, while with vanilla gradients they achieve **0.89**. The difference is *not* statistically significant. The fact that users are consistently able to discriminate between HIGH and LOW models is expected, and serves as a *sanity check* that these explanations are meaningful for humans.

Condition	Gradients	IG
HIGH VS LOW	0.89	0.95
HIGH VS MEDIUM*	0.85	0.52

Table 2: Both methods do well in (HIGH VS LOW). In HIGH VS MEDIUM, performance drops dramatically for IG. * = statistical significant difference ($\rho < 0.001$)

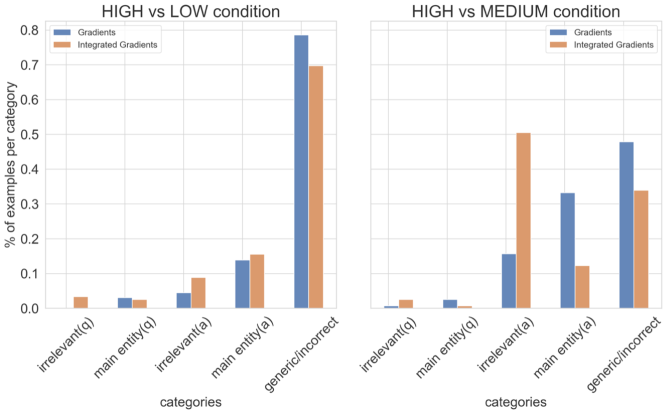
When the same experiment was repeated in the HIGH vs MEDIUM condition, we found clear and statistically significant differences between the two explainability methods. Using IG, participants reach only **0.52** accuracy, while with vanilla gradients their performance is **0.85**. This is *surprising*, given that the difference in performance between the two models is still quite large (about 20% F1); the expectation is that both methods would capture this difference relatively well. It appears that when both models *more or less* align with human beliefs, the task is much more difficult. To solve the task, humans now need to engage in more analytical thinking and cannot simply rely on belief biases to solve the task. We further investigate these differences through qualitative coding.

Qualitative analysis. After each instance, we asked participants to describe how the worse model will fail. We do not provide detailed guidelines in order to not further bias the participants by introducing specific criteria. The instructions given to the participants are shown in Appendix B.

We collected 1440 responses, which were all inspected manually to uncover categories (codes). After multiple iterations, we tagged each response with one code (categories are mutually exclusive, no response can be placed in two). A description of the categories and their distribution are shown in Figure 3, and examples of feedback per category are provided in the Appendix B.

In the HIGH vs LOW condition, feedback for both methods was generic (about 70-80% of the time), e.g., *model B is likely incorrect so it is worse*. This was expected: this task should be easy when model differences are large and humans can rely on their *system 1* processes to get through the task without thinking deeply about the explanations.

In the HIGH vs MEDIUM condition, the distribution of the feedback categories is very different. For IG, 50% of the time participants *felt* the highlighted tokens where irrelevant. This is not the case for gradients, where only about 15% of responses fell in that category. Additionally, for vanilla gradients, 50% of feedback is generic, signaling that in this condition, it may have been an easy task as well; explanations are making model behavior clear enough. It remains an open question whether IG explanations may in fact be more faithful to the model reasoning. In that case, *expert users* (e.g. a system engineer debugging a system) may not find IG attributions irrelevant and would be able take better advantage of the information provided. For this reason, other kinds of human participants may show different results. Nevertheless, as evaluating on non-experts (crowdsourced workers for example) is common, this preliminary result is important: it shows that **conclusions can shift dramatically when introducing additional model comparisons** which reduce the participants' ability to rely on prior knowledge.



Category	Description
irrelevant(q)	The model puts emphasis on irrelevant entities in the **question**.
Main entity (q)	The model does not put emphasis on the main entity in the **question**
irrelevant(a)	The model puts emphasis on irrelevant entities in the **answer**.
Main entity (a)	The model does not put emphasis on the main entity in the **answer**
generic/ incorrect	"Model is worse, because it is incorrect."

Figure 3: Feedback categories and their distribution. We observed that the HIGH vs MEDIUM condition results are considerably different from the HIGH vs LOW condition, with more participants giving generic answers for vanilla gradients, and emphasizing the irrelevant terms highlighted in the IG condition.

7 Discussion: Mitigating Belief Bias

This study introduced *additional conditions* in which the human participants could not rely on their belief biases to facilitate the task at hand. We presented a case study on evaluating reading comprehension models in model selection and human forward prediction paradigms, and we showed that this simple addition led to different conclusions in the evaluation and a better understanding of how humans interacted with explanations. Other tasks and paradigms might call for different setups, but generally including conditions with models of varying quality would be helpful both for the purposes of bias control, and for simulation of real-life use of explainability techniques to support decisions about which model is safer to deploy.

To conclude, we will briefly mention other directions for mitigating belief biases that can also be explored in future work and which should be kept in mind when developing evaluation protocols for explainability.

Reducing ambiguity. Ambiguity of task instructions leads humans to align interpretations to their own prior beliefs (Heath and Tversky, 1991); this may lead to misinterpretation and results which do not reflect the intended interaction with explanations. Ambiguity may also be present in other parts of the evaluation setup. For example, Lamm et al. (2020) evaluate the effectiveness of explanations in helping humans detect model errors for open-domain QA, but the data they use contains questions where multiple answers can be true. Users may deem an answer to be correct or incorrect based on their understanding of the question, which makes the effect of explanations blurry. Removing ambiguous instances from the data can be a way of reducing such confounds.

Removing time constraints. Time constraints exacerbate reliance of system 1 processes, which leads to humans relying on belief biases. In crowdsourced evaluations, it is common practice to to provide workers with enough time to perform tasks, but workers may have intrinsic motivations for performing tasks quickly. A major challenge for evaluation research with crowd workers is creating better incentives for engaging in system 2 processes, e.g. pay schemes which encourage workers to be more analytical and accurate (Bansal et al., 2019).

Include fictitious domains. Using data from domains from which subjects have no prior beliefs e.g. fictitious domains, may be an efficient way of controlling for belief bias in some tasks[7]. This strategy has been used outside of NLP (Poursabzi-Sangdeh et al., 2021; Lage et al., 2019; Slack et al., 2019), where subjects are asked to imagine alternative worlds such as scenarios involving aliens. In QA for example, one could introduce context-question pairs that describe facts about fictitious scenarios that sufficiently differ from human reality.

8 Conclusion

The main contribution of this paper is bringing the discussion of belief bias from psychology into the context of evaluating explainability methods in

[7] Again, we emphasize that some strategies are task dependent; fictitious domains may not be relevant in some tasks.

NLP. Belief bias is a phenomenon which plays a role in human decision making and which interacts with previous evaluations in a way which may affect the conclusions we draw from these paradigms. We provide an overview of belief bias, making a connection between findings in psychology and the field of NLP, and present a case study of evaluating explanations for BERT-based reading comprehension models. We show that introducing models of various quality and adversarial examples can help to account for belief bias, and that introducing such conditions affects the conclusions about which explainability method works better. Finally, we provide additional insights and ideas for how to account for belief bias effects in human evaluation.

9 Broader Impact Statement

The work presented here makes strides towards a better understanding about the interaction of humans with explanations of model decisions. We have highlighted a phenomenon studied in psychology with hope that this opens the door to more NLP research involving a wider and more interdisciplinary understanding of humans, and the effect of explainability.

This study involved human participants recruited on Mechanical Turk platform. No personally identifiable data was collected from the participants, they were made aware that the data would only be used for research, and they were not exposed to any emotionally traumatizing or offensive stimuli. We ensured a minimum $15 hourly wage.

Acknowledgements

We thank the reviewers for their insightful feedback for this and previous versions of this paper. This work is partly funded by the Innovation Fund Denmark.

References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *CHI*.

Richard B Anderson and Beth M Hartzler. 2014. Belief bias in the perception of sample size adequacy. *Thinking & Reasoning*, 20(3):297–314.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437.

Julie Linda Barston. 1986. An investigation into belief biases in reasoning.

Andrea Bierema, Anne-Marie Hoskinson, Rosa Moscarella, Alex Lyford, Kevin Haudek, John Merrill, and Mark Urban-Lurain. 2020. Quantifying cognitive bias in educational researchers. *International Journal of Research & Method in Education*, pages 1–19.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Laura Caravona, Laura Macchi, Francesco Poli, Michela Vezzoli, Miriam AG Franchella, and Maria Bagassi. 2019. How to get rid of the belief bias: boosting analytical thinking via pragmatics. *Europe's Journal of Psychology*, 15(3):595–613.

Pat Croskerry. 2009. A universal model of diagnostic reasoning. *Academic medicine*, 84(8):1022–1028.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

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

J St Evans and Handley BT. Sj and harper, c. 2001. *Necessity, possibility and belief: A study of syllogistic reasoning. Quarterly Journal of Experimental Psychology A*, 54:935–958.

J St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

Jonathan St BT Evans. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59:255–278.

Jonathan St BT Evans and Jodie Curtis-Holmes. 2005. Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11(4):382–389.

Jonathan St BT Evans and Keith Ed Frankish. 2009. *In two minds: Dual processes and beyond.* Oxford University Press.

Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42.

Vinod Goel and Oshin Vartanian. 2011. Negative emotions can attenuate the influence of beliefs on logical reasoning. *Cognition and Emotion*, 25(1):121–131.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Chip Heath and Amos Tversky. 1991. Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of risk and uncertainty*, 4(1):5–28.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. pages 4163–4174.

Daniel Kahneman. 2003. A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697.

Paul A Klaczynski, David H Gordon, and James Fauth. 1997. Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology*, 89(3):470.

KC Klauer, J Musch, and B Naumer. 2000. On belief bias in syllogistic reasoning. *Psychological Review*, 107.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. 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)*, pages 5195–5205, Hong Kong, China. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. pages 615–621.

Henry Markovits and Guilaine Nantel. 1989. The belief-bias effect in the production and evaluation of logical conclusions. *Memory & cognition*, 17(1):11–17.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288.

Stephen E Newstead, Paul Pollard, Jonathan St BT Evans, and Julie L Allen. 1992. The source of belief bias effects in syllogistic reasoning. *Cognition*, 45(3):257–284.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4):1023–1031.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. pages 1–52.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016a. " 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*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016b. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Marko Robnik-Šikonja and Marko Bohanec. 2018. Perturbation-based explanations of prediction models. In *Human and machine learning*, pages 159–175. Springer.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. pages 8–14.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

D Slack, SA Friedler, CD Roy, and C Scheidegger. 2019. Assessing the local interpretability of machine learning models.

Keith E Stanovich and Richard F West. 2008. On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology*, 94(4):672.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. pages 3319–3328.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13230–13241.

Dries Trippas and Simon J Handley. 2018. The parallel processing model of belief bias: Review and extensions.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 104–115. IEEE.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15.

A Experiment 1: Human Forward Prediction

Below we show the instructions provided to the participants, as well as an example of the saliency maps presented to participants for adversarial examples.

Instructions. Question-answering systems are a particular form of artificial intelligence. The task here is for you to learn to predict how the system answers questions. In other words, when in a bit, you are presented with questions, the task is not to provide the right answer, but to guess the answer the system provided. For each question, you will also see a context paragraph. The answer is a span of text in this paragraph. Instead of writing out the answer, you can simply mark the relevant span.

If you want to select a new answer, please click *reset answer*, if you are ready to see the model answer, please click *show answer*. Note that your answer will lock at that time.

Raw Results. In our evaluation, we use the first 15 points as training, therefore, we discard them from the main evaluation but show them in this section. Overall, we see that training, for the most part has a positive effect, or not so much of an effect. These scores can be seen in Table 3.

	MODEL	HUMAN		
		BASELINE		
LOW-ORIG	0.17	0.14	0.52	52.27
LOW-ADV	0.15	0.10	0.36	54.36
HIGH-ORIG	0.79	0.53	0.58	37.12
HIGH-ADV	0.66	0.35	0.48	47.64
		INTEGRATED (IG)		
LOW-ORIG		*0.34	0.35	41.68
LOW-ADV		*0.36	0.28	44.38
HIGH-ORIG		*0.71	0.76	46.87
HIGH-ADV		0.46	0.47	42.99
		GRADIENTS		
LOW-ORIG		***0.64**	*0.09	*32.16
LOW-ADV		***0.63**	0.23	*30.05
HIGH-ORIG		***0.82**	*0.84	44.65
HIGH-ADV		***0.57**	*0.62	*52.30

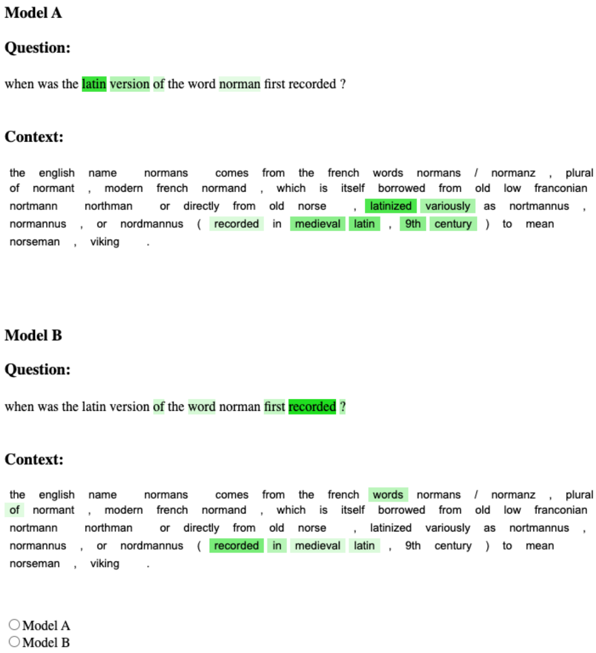
Table 3: Raw scores, before removing data points on training session

B Experiment 2: Best Model Selection

Below we show the instructions given to the participants, and more details about the qualitative analysis of the feedback we obtained.

Instructions. Question-answering (QA) systems are a particular form of artificial intelligence. We have trained two QA systems and have extracted the most important words the model uses to make its final decision. Based on these highlighted words, your task is to select the model that you think is more likely to perform best. Additionally, please write how the low-performing model fails and/or how it could be better (try to be detailed)

User Interface. An example instance, as shown to the participants, can be seen in Figure 4.



Model A

Question:

when was the latin version of the word norman first recorded ?

Context:

the english name normans comes from the french words normans / normanz , plural of normant , modern french normand , which is itself borrowed from old low franconian nortmann northman or directly from old norse , latinized variously as nortmannus , normannus , or nordmannus (recorded in medieval latin , 9th century) to mean norseman , viking .

Model B

Question:

when was the latin version of the word norman first recorded ?

Context:

the english name normans comes from the french words normans / normanz , plural of normant , modern french normand , which is itself borrowed from old low franconian nortmann northman or directly from old norse , latinized variously as nortmannus , normannus , or nordmannus (recorded in medieval latin , 9th century) to mean norseman , viking .

○ Model A
○ Model B

Figure 4: Experiment 1 UI: LOW(bottom) vs HIGH(top) condition.

Qualitative analysis of feedback. In Table 4, we include a few examples of the sentence that were categorized using the qualitative codes. Unsurprisingly, once participants found a strategy for giving feedback , they mostly stuck to it.

After categorizing all the feedback into each category, we visualize the distribution per condition. This can be found in Figure 3. We find that for the HIGH vs LOW conditions, the distribution is very similar between gradients and integrated gradients. Many participants gave very generic feedback , for example by simply saying that ”model A is better because it is correct, and model B is wrong”. This was not surprising, as here the differences were supposed to be clear and it is likely most participants did not have to think too hard before making

QUALITATIVE CODES	EXAMPLES
Irrelevant (q)	**1.** Model A only extracted some important words but also some punctuations in the question which is insufficient to derive to a good answer. Model B extracted a number of key important words that would lead to the correct answer. **2.** Option b chose quantitative statements, while option A seems confused about what it's looking for since it highlights all sorts of things in the question.
main entity (q)	**1.** The words "year" and "norman" in the question were not extracted by Model A. The Model will not be able get the correct answer without knowing what to look for. **2.** The question was asking about the year lavoisier's work was published but neither of the key words in this question were highlighted. Model A had no idea where to locate the answer without considering those key words.
main entity (a)	**1.** The answer requires a year; it hasn't highlighted any years as part of the answer. **2.** Answer needed to be a name and option A chose nothing that could be a name.
Irrelevant (a)	**1.** Model B has highlighted many extra words in the answer **2.** Both models selected the correct terms, but model A selected more irrelevant terms in the answer too, so it's less likely to choose the correct one from those numerous options. **3.** B highlighted the answer but also too much unneeded info.
Generic/correctness	**1.** Model A does not highlight the right answer **2.** Model B is wrong and model A is correct

Table 4: Examples of some of the feedback categorized into these classes

a decision. However, the distribution is very different for the HIGH vs MEDIUM conditions. Here, for standard gradients, the feedback followed a similar pattern as in the previous condition, but about 30% less examples received generic feedback than before. For integrated gradients, most examples received feedback regarding the irrelevant terms being highlighted, showing that even when the difference in performance between models is large (20 F1 points), this method makes the distinction difficult for the best model selection task.