Epistemic closure filters for natural language inference

Michael Cohen
micohen@stanford.edu
Stanford CS224U, Spring 2021, project report

Abstract

Epistemic closure refers to the assumption that humans are able to recognize what entails or contradicts what they believe and know, or more accurately, that humans' epistemic states are closed under logical inferences. Epistemic closure is part of a larger *theory of mind* ability, which is arguably crucial for downstream NLU tasks, such as inference, QA and conversation. In this project, we introduce a new automatically constructed natural language inference dataset that tests inferences related to epistemic closure. We test and further fine tune the model RoBERTa-large-mnli on the new dataset, with limited positive results.

1 Introduction

In this project, we introduce and study a new natural language inference (NLI) dataset about *epistemic closure* inferences. The NLU task of natural language inference (NLI) is the task of predicting, given two sentences, a *premise* X and a *hypothesis* Y, whether X entails Y, contradicts Y or is neutral with respect to Y, under common sense use (Storks et al., 2020).

In epistemology, *epistemic closure* refers to the property of certain epistemic states (and epistemic verbs) to be *closed under* the inference relations of *entailment* and *contradiction* (Luper, 2020). In simpler words, epistemic closure captures the assumption that the content of our epistemic states (the propositions we believe or know) is subject to the same inference relations that govern propositions in general. For instance, if we assume that premise X entails the hypothesis Y, and that Ann believes X, then according to the closure of belief under entailment (epistemic closure for belief), it follows that Ann believes Y. Likewise, if X contradicts Y, the claim that *Ann believes* X, would contradict the claim that *Ann believes* Y, given epistemic closure.[1]

The human tendency to attribute epistemic closure is part of a larger *theory of mind* ability. Theory of mind, also known as *folk psychology* within philosophy (Ravenscroft, 2019) or *intuitive psychology* in artificial intelligence (Storks et al., 2020), is a term that describes the human ability to reason about other humans' mental states, thus recognizing them as rational agents, having beliefs, knowledge, intentions and emotions of their own. Theory of mind reasoning is considered an important developmental test within cognitive science and philosophy of mind, used for evaluating the cognitive capacities of young children and non-human animals (Ravenscroft, 2019). According to this line of research, the ability of a thinking system to accurately track the thoughts of a different system is an important mark of intelligence.

Within the broader area of theory of mind, epistemic closure inferences test the specific ability to recognize others as rational agents who are aware of the intuitive logical relations between their own thoughts. Epistemic closure might seem like a relatively insignificant linguistic and cognitive phenomenon, but we believe that it plays a significant part in everyday communication. The following examples highlight the implicit epistemic closure reasoning that occurs in everyday conversations. Consider the following fictitious conversation:

> **A:** Why isn't James here?
> **B:** James thinks that the event was cancelled.
> **A:** Didn't he get the update about moving the event's date?
> **B:** He thinks that the event is tomorrow.

Here, **B**'s last utterance is inconsistent with **B**'s

[1] In philosophy, *epistemic closure* is often specifically reserved for closure under logical entailment. Here we use it liberally to apply both for entailments and contradictions.

earlier utterance. To recognize this oddity, competent speakers must reason as follows: first, recognize that the claim that *The event was cancelled* contradicts the claim that *The event is tomorrow*. Second, note that **B** is ascribing contradictory beliefs to James. This is a violation of epistemic closure. Now consider this conversation:

> **A:** Where is James?
> **B:** James assumes that today's meeting was moved to next Friday.
> **A:** But why isn't he in the meeting today?

Epistemic closure reasoning explains the oddity of **A**'s last utterance. The claim *Today's meeting was moved to next Friday* entails that *The meeting is not today*. By epistemic closure, if James assumes that *today's meeting was moved to next Friday*, then James assumes that *the meeting is not today*. This makes **A**'s last question redundant.

In everyday conversation, participants keep a mental model of the thoughts of other participants of the conversation (see, e.g. (Stalnaker, 1978; Lewis, 1979)). Such mental models make it easy for humans to implicitly recognize conversational entailments and contradictions related to the participants' mental states. It is far from obvious that current NLP models have the capacity to recognize similar entailments and contradictions.

In this project, with the task of NLI in mind, we focus on premise hypothesis pairs that involve epistemic closure reasoning, such as:

> ***premise:*** James thinks that the event was cancelled.
> ***hypothesis:*** He thinks that the event is tomorrow.
> ***label:*** contradiction

We do so by automatically generating examples such as the above one from the existing SNLI dataset, via a process we call epistemic closure *filters*. As we demonstrate later, some epistemic closure inferences are quite subtle, even if humans are quick to recognize them.

Our central hypothesis is that the state of the art model RoBERTa-large, finetuned on the MNLI dataset, struggles with simple epistemic closure inferences. We further hypothesize that RoBERTa-large can quickly learn epistemic closure inferences. We test the latter hypothesis using the method of *inoculation* (Liu et al., 2019a), in which we expose the model only to a small amount of data from the challenge dataset.

2 Related work

common sense entailments in NLI. Logicians distinguish between deductive (i.e. logical) and non-deductive inferences (Beall et al., 2019). Deductive inferences are those in which it is impossible for the premise to be true and the conclusion to be false. In non-deductive inferences, the premise does not necessitate the conclusion, rather just make it more likely.

Epistemic closure inferences, like other theory of mind inferences, are not logical inferences. It is possible that X logically entails Y, and for Ann to believe X, but not to believe Y, maybe because she does not recognize that X entails Y. Sometimes agents fail to see the logical connection between a set of sentences, and there is no contradiction in entertaining such a failure. Likewise, it is possible for X to logically contradict Y and for John to believe both X and Y. Sometimes agents hold contradictory beliefs.

It is a matter of debate whether NLI datasets should only include logical entailments and contradictions (Zaenen et al., 2005; Manning, 2006). Nevertheless, recent large NLI datasets, such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018b), combine both types of inferences, with an aim to capture a notion of 'common sense' entailment. As a result, recent work aims to explore to what degree NLI models that were trained on such datasets can explicitly recognize non-deductive inferences, such as *pragmatic* (Jeretic et al., 2020) and *abductive* (Bhagavatula et al., 2020) inferences. This project continues this line of work by focusing on a different type of non-deductive inference.

Epistemic closure in formal semantics. Standard semantic models for attitude verbs like *to know* and *to believe*, which are based on Hintikka's (Hintikka, 1962) and montague's (Janssen and Zimmermann, 2021) semantic frameworks, assume the property of epistemic closure (Pearson, 2020). Thus, the standard view on epistemic verbs in formal semantics vindicates the epistemic closure assumption made in this project.

However, it is not always made clear whether epistemic closure is a desirable feature of the formal systems of just a formal artifact. This confusion is illustrated in *the problem of logical omniscience* (Égré, 2020). The problem is that an unrestricted form of epistemic closure predicts that agents know (or believe, or assume, or think) every logical consequence of what they know (or believe, etc.). Since actual agents are not perfect logicians, epistemic closure (and the standard semantic models that assume it) has been criticized as unrealistic. If the actual logician Gottlob Frege (1848-1925) did not recognize that his axioms of formal arithmetic famously contradict each other (Irvine and Deutsch, 2021), why make it part of the meaning of *to believe* that every agent can recognize any logical entailment and contradiction of their beliefs?

While we don't have a solution to the problem of logical omniscience, we don't believe that it poses any serious obstacle in this project. In constructing our dataset, we modify existing examples of premise hypothesis pairs from the SNLI dataset (see data section); such pairs are not meant to offer logically challenging examples of entailments or contradictions (like the consistency of formal arithmetic), rather, simple, common sense examples of entailments and contradictions. Therefore, by assuming epistemic closure on such examples, we are not assuming logical omniscience.

Epistemic closure in symbolic AI. Within symbolic AI, *epistemic logic* is a logical system that allows for explicit reasoning about different forms of epistemic closure (Hintikka, 1962; Fagin et al., 1995). Given a sentence of the logic p, the epistemic logic expression $K_a(p)$ reads *Agent a knows p*. Different precise forms of epistemic closure can be then expressed in the formal language, for instance $K(p) \wedge K(p \rightarrow q) \rightarrow K(q)$ – if the agent knows that p and that p entails q, then the agent knows q. Most of the existing literature on epistemic closure involves some usage of epistemic logic (Luper, 2020; Rendsvig and Symons, 2021).

One epistemic closure principle that guides us in this project can be expressed in epistemic logic as follows:

If $X \rightarrow Y$ then $K_a(X) \rightarrow K_a(Y)$

It states that if X entails Y, then if a knows X, then a knows Y. Further epistemic closure principles that involve contradiction will be explained in the data section.

Theory of mind in NLP. In the intersection of cognitive science and NLP, (Nematzadeh et al., 2018) and (Le et al., 2019) explore the ability of language models to perform *the false belief task* (Ravenscroft, 2019), a classic theory of mind task, through question answering (QA).We note that the false belief task and epistemic closure are related but distinct theory of mind phenomena.

Theory of mind reasoning is also related to the NLP task of *speaker commitment* or *event factuality*. In that task, an NLP model has to predict to what extent a speaker is committed to the complement of a sentence (Jiang and de Marneffe, 2019; Ross and Pavlick, 2019) cf. (de Marneffe et al., 2012). For example, the model has to predict that the speaker of the sentence *Michael knows that there is milk in the fridge* is committed to the claim that there is milk in the fridge. Speaker commitment is related to theory of mind inferences since it includes mental state verbs (like *to know* and *to believe*.)

In a previous 224N project (Cohen, 2021), we constructed an NLI dataset that includes theory of mind reasoning that mixes both speaker commitment reasoning and epistemic closure reasoning. That project only included 4 explicit epistemic closure inference tasks, unlike the 12 that are tested here. Furthermore, that project did not distinguish between the neutral and contradictory labels (treating both of them as non-entailment), although this distinction plays an important role in epistemic closure reasoning (see the data section). In addition, the previous project did not try to inoculate existing models for epistemic closure performance.

Challenge NLI datasets. In recent years, many special NLI dataset have been constructed, with the aim of challenging existing NLI models trained on SNLI and MNLI on types of entailments that can be considered out-of-domain, or with minor manipulations to the existing datasets. Manipulations include testing NLI models only on hypotheses (Poliak et al., 2018), making small lexical changes on a single word in the example (Glockner et al., 2018), and inducing spelling errors (Naik et al., 2018). NLI tasks on specific out-of-domain datasets include datasets for defeasible reasoning (Rudinger et al., 2020), sentences with multiple quantifiers (Geiger et al., 2018), entailment with conjunctions (Saha et al., 2020), the transitivity of the entailment relation (Yanaka et al., 2021), entailments involving event veridicality (Jiang and de Marneffe, 2019), and inferences involving presuppositions and implicatures (Jeretic et al., 2020). From these challenge datasets one can study the heuristics that NLI models employ in their prediction (e.g. predicting entailment if there is a sub-string overlap between the premise and the hypothesis) (Naik et al., 2018; McCoy et al., 2019). This project continues this line of work: it offers a challenging out-of-domain type of inference (epistemic closure) in a way that only requires a minimal modification to existing examples (as explained in the data section). Furthermore, our experimental results reveal heuristics that models employ when encountering new types of examples.

3 Data

We generate data automatically, resulting in an entirely new dataset. Here we detail the generation of the data.

General procedure. To generate data, the following general procedure is used: a random $(X, Y, label)$ example is taken from the SNLI test set. It is then passed through one of four *filters*, that modify the strings X and Y. Then a new label is generated for the modified example, depending on the $label$ of the original example and the filter. Since there are four different filters and three possible labels for each example, we get $(4 \times 3 =)$ 12 types of examples, each of which we call a *template*. We generate 300 examples for each template.

Preprocessing and modifying X and Y. Although the exact modification for a given premise X or hypothesis Y depends on the filter, the general procedure is as follows:

1. Test that X is a full sentence and not a fragment, by checking that it is S-rooted in its syntax tree. Discard X if it is not S-rooted.
2. Lowercase the first letter of X.
3. Append the sub-string (schema) *name verb that* to the left of X.

The same process applies to Y. In step 3., the *name* placeholder is replaced by randomly choosing from an even gendered list of 20 names from the US census, or with the pronouns *He* or *She*. The *verb* placeholder is taken from the following list of strings: *believes,thinks, assumes, suspects, knows, sees, learns, understands, recognizes, remembers*.

Each one of these verbs is assumed to respect epistemic closure, and further acts semantically as a propositional attitude (Nelson, 2019). Syntactically, appending a declarative sentence X to a propositional attitude results in a grammatical declarative sentence. Note, therefore, that step 1. of the preprocessing guarantees that step 3. results in a semantically and syntactically valid sentence.

For a concrete example, let X be the SNLI example:
A land rover is being driven across a river.
Since this X is S-rooted, the above process may result in modifying X to:
Eva sees that a land rover is being driven across a river.

Labeling the modified examples. The labels for each new template are not generated manually by human annotators, rather automatically via a set of epistemic closure rules. Table 1 summarizes the modified labels (in the cells) given to each filter (in rows) and original label (in columns) pair. Here we elaborate on the reasoning behind these rules.

There are three ingredients that determine the label for a given epistemic closure filter: the original label of the X, Y pair, which agents are mentioned in the modified pair X, Y pair, and, if more than one agent is mentioned, whether the verb used in the modification is factive or not.

Starting with the most simple case, consider a situation where we have the same agent both in the premise and in the hypothesis. In that case, epistemic closure dictates that the new label for the modified X, Y is the same as the original label: if X entails Y, then the agent believes X entails that the agent believes Y; if X contradicts Y, then the agent believes X contradicts that the agent believes Y; if X and Y are neutral, then the claim that the agent believes X is neutral with respect to the claim the agent believes Y. In other words, we assume the agent recognizes the relation between X and Y the same as we do. Note that this holds both if we refer to the agent by name both in the premise and hypothesis (*Eva thinks* X, *Eva thinks* Y, the single agent filter), or if we refer to them using a pronoun anaphora in the hypothesis (*Eva thinks* X, *She thinks* Y, the anaphora filter). Although identical in meaning, the anaphora case requires the extra step of resolving co-reference during inference.

One aspect of epistemic closure is recognizing that given a particular agent, mental states are closed under entailment and contradiction; a second aspect is recognizing that this does not hold if we consider different agents with potentially independent mental lives. Recognizing the Independence of the mental states of different agents is a core theory of mind ability (Ravenscroft, 2019).

Even if X entails Y, and Eva knows X, it does *not* imply that John knows Y. After all, Eva's mental state is independent from that of John's.

Therefore, when we consider different agents, we will need to modify the original relation between X and Y. If X entails Y, or is neutral w.r.t Y, then one agent's beliefs or knowledge about X is independent (and thus neutral) w.r.t to another agent's beliefs or knowledge about Y. This reasoning justifies the choice of labels in the columns Neutral and Entailment for the two last rows in table 1.

The case of two agents, where X, Y are contradictory, is more complicated, and requires drawing a distinction between our epistemic verbs. Epistemic verbs are divided into factive and non-factive ones. In common sense use, factive verbs are assumed to represent reality truthfully, or accurately (Karttunen, 1971). Among the list of verbs we use, the verbs *knows, sees, learns, understands, recognizes, remembers* are factive verbs. In common usage, if Ann knows that it is raining, then it is in fact raining. Non-factive verbs are those verbs that are not assumed to accurately represent reality. The rest of the verbs we use, *believes, thinks, assumes,* and *suspects*, are non-factive. If Ann believes that it is raining, she might be wrong, and it might not be raining.

It is impossible for two agents to *know* contradictory claims. If Ann knows that it is raining here right now, then Bob cannot know that its not raining here right now, since the verb *to know* is assumed to represent reality accurately, and in reality its impossible for it to rain and not to rain at the same time. In general, if X contradicts Y, then for any factive verb V, the claim that Ann V's that X contradicts the claim that Bob V's that Y. This situation is captured in the multi agent factive filter.

On the other hand, it is possible, and quite common, for two agents two hold contradictory beliefs (or other non-factive attitudes). If X and Y are contradictory, there is no contradiction in assuming that Bob believes X while Ann believes Y, and the modified two sentences are judged neutral. See the multi agent non-factive filter row in table 1.

Note that it is only in the multi agent, contradiction case, that the difference between factive and non factive verbs plays a role. Therefore, the single agent and anaphora filter do not include this distinction.

3.1 Summary of the four filters

With the above information in mind, we can briefly describe the function of each filter. Each filter takes as an **input** an S-rooted (X, Y) example pair from the SNLI test set. The filters differ in their output and label.

Filter	Neut.	Ent.	Cont.
control	N	E	C
single agent filter	N	E	C
Anaphora filter	N	E	C
multi agent factive	N	N	C
multi agent non-factive	N	N	N

Table 1: Summary of the interaction between filters (rows) and the original label that is fed to them (cols). Underlined cells indicate that the filter has caused a label change. The Blue cells are templates which were used in the inoculation process.

Single agent filter.
Output: (*name1 verb X, name1 verb Y.*)
Here *name1* is taken from our list of names, and *verb* is taken from our list of (factive and non factive) verbs. **Label:** the same as that of X, Y. **Example output:** premise: Bob thinks that X. hypothesis: Bob thinks that Y.

Anaphora filter.
Output: (*name1 verb X, anaphora verb Y*).
Here *name1* is taken from our list of names, and *anaphora* is the pronoun he or she (depending on *name1*), and *verb* is taken from our list of (factive and non factive) verbs. **Label**: same as that of X, Y. **Example output:** premise: Eva sees that X. hypothesis: She sees that Y.

Multi agent factive filter.
Output: (*name1 factive verb X, name2 factive verb Y*).
Here *name1* and *name2* are different names taken from our list of names, and *factive verb* is taken from our list of factive verbs. **label:** the same as that of X, Y if the original label was neutral or contradiction, and neutral otherwise. **Example output:** premise: Eva sees that X. hypothesis: John sees that Y.

Multi agent non-factive filter.
Output: (*name1 non-factive verb X, name2 non-factive verb Y.*)
Here *name1* and *name2* are different names taken from our list of names, and *non-factive verb* is taken from our list of non-factive verbs.
label: always neutral. **Example output:** premise: Eva assumes that X. hypothesis: John assumes that Y.

4 Model and evaluations metrics

To evaluate our epistemic closure templates, we use the model RoBERTa-large finetuned on the MNLI dataset, available via Huggingface.[2] The MNLI dataset contains 433K crowd-sourced and labeled examples of premise hypothesis pairs, from multiple genres, including examples from written and spoken sources (Williams et al., 2018a). RoBERTa-large finetuned on the MNLI dataset achieves a score of 0.908 on the MNLI test set (Liu et al., 2019b).

We report the accuracy score with respect to each individual template, without averaging the results.

5 Experiments

We perform two main experiments: testing the model RoBERTa-large-mnli (as is from Huggingface) on our dataset, and testing the model after further fine-tuning it via inoculation.

5.1 RoBERTa-large-MNLI without inoculation

Results The accuracy results of the original RoBERTa-large-MNLI are reported in table 2, and can act as a baseline. Note that the first row of the table is a control filter, which tests the model's accuracy on 300 unchanged examples for each label, taken from SNLI.

Analysis First note that the model performs well on the control test, with the lowest score of 0.86 for the neutral examples. Comparing table 1 and table 2 reveals the following pattern: in every template in which the filter has not modified the label, and in which the original label is entailment or contradiction, accuracy score is high and close to the control. In the templates that do change the label of the original examples (underlined cells in table 1), the model's accuracy drops below 0.1. This suggest that RoBERTa-large-MNLI employs the following heuristic when encountering epistemic closure filters: the model ignores the modification to X, Y generated by the various filters, and predicts a label according to X, Y. This explains why the model receives high accuracy on every template that does not modify the original label, and extremely low accuracy on the tables that do.

Therefore, the model does not seem to perform any epistemic closure reasoning. We hypothesize that this is because the MNLI dataset does not include examples of epistemic closure inferences.

Filter	Neut.	Ent.	Cont.
control	0.86	0.923	0.903
single agent filter	0.806	0.946	0.94
Anaphora filter	0.783	0.933	0.93
multi agent factive	0.663	0.043	0.936
multi agent non-factive	0.6	0.033	0.05

Table 2: Accuracy results on RoBERTa-large-MNLI

Filter	Neut.	Ent.	Cont.
control	0.953	0.84	0.106
single agent filter	0.973	0.763	0.043
Anaphora filter	1.0	0.42	0.023
multi agent factive	0.996	0.87	0.016
multi agent non-factive	1.0	0.983	0.996

Table 3: RoBERTa-large-MNLI after 10 examples

Filter	Neut.	Ent.	Cont.
control	0.87	0.84	0.83
single agent filter	0.836	0.793	0.57
Anaphora filter	0.843	0.68	0.616
multi agent factive	0.78	0.3	0.953
multi agent non-factive	0.996	0.906	1.0

Table 4: RoBERTa-large-MNLI after 50 examples

5.2 RoBERTa-MNLI with inoculation

Since the original model seems to ignore epistemic closure inferences, we turn to see if this type of inference can be taught. We believe, however, that feeding to the model a large number of examples from all templates, using all the verbs from our verbs list, will be not very illuminating. As table 1 shows, the model just needs to memorize the modification of three templates (underlined in table 1) to reach overall high accuracy results.

Instead, we pick the method of *inoculation* (Liu et al., 2019a) in order to further finetune the model. In this method, we gradually finetune the model on more and more examples, starting with a very small number of examples. Furthermore, in the inoculation process, we only train the model on a limited subset of templates, with a limited subset of epistemic verbs. Inoculation will allow us to

[2] See https://huggingface.co/transformers/ for the package and https://huggingface.co/roberta-large-MNLI for the model.

Template	Neut.	Ent.	Cont.
control	0.0	0.0	1.0
single agent filter	0.333	0.0	0.583
Anaphora filter	0.363	0.0	0.59
multi agent factive	0.0	0.0	1.0
multi agent non-factive	0.996	0.993	0.993

Table 5: RoBERTa-large-MNLI after 250 examples

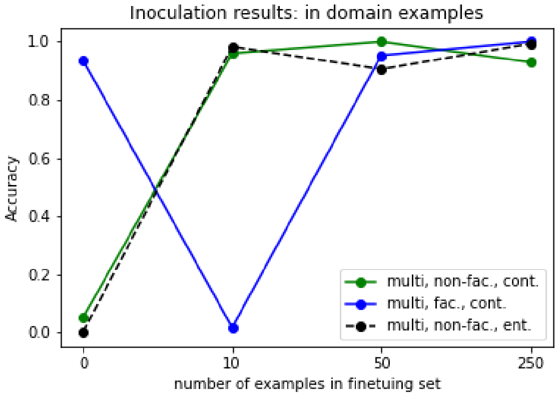

Figure 1: Inoculation results (10, 50, 250, examples) for in-domain template (blue cells in table 1).

answer the following questions:
1) Can the model quickly learn the structure of seen (in-domain) templates?
2) Can the model generalize from seen templates to held out (out-of domain) templates?
3) Can the model generalize from the behaviour of seen verbs to held out verbs?

In the inoculation process, we only train the model on the following templates (the in-domain templates, colored blue in table 1)
- Multi agent non-factive contradiction
- Multi agent non-factive entailment
- Multi agent factive contradiction
The first two templates were chosen since the original model has low accuracy on these templates. The third template was chosen since it requires realizing that contradictions can occur in multi agent cases (this is the only template in which a multi agent filter results in a non-neutral label). The training examples included a different list of names, a non-factive verb list of: believes, thinks (holding out: assumes, suspects) and with a factive list of: knows, sees (holding out: learns, understands, recognizes, remembers). We inoculated the model on training sets of sizes 10, 50, and 250 examples. For each inoculation training set, we fine-tune the model using the Trainer method of the Huggingface transformers package.[3] We analyze each individual inoculation step:

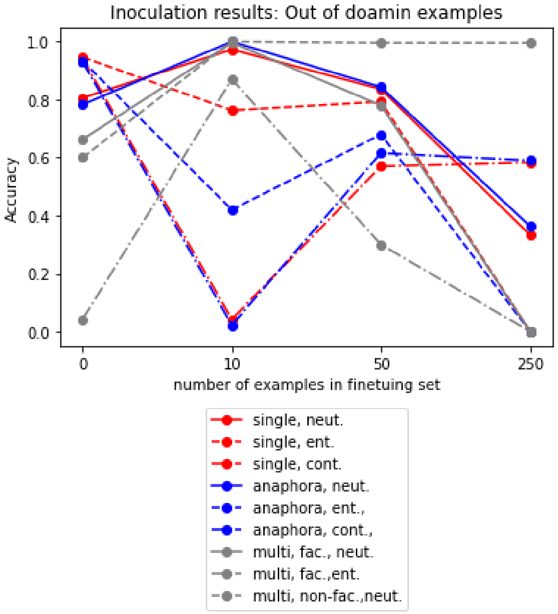

Figure 2: Inoculation results (10, 50, 250, examples) for out of domain examples (non-blue templates in table 1).

Analysis of the 10 examples inoculation (table 3) Even only after 10 examples, the model shows significant improvement in two of the three in-domain templates (see figure 2). The model improved on all the underlined examples of Table 1. It seems that the model avoids predicting contradictions (note the contradiction column in table 3), and is biased towards neutral (note the Neutral column), to the degree that both the control templates and the out-of-domain examples accuracy drops significantly.

Analysis of the 50 examples inoculation (table 4). After 50 examples, the model still avoids predicting contradictions (note the contradiction column of table 4), and is more biased towards neutral compared to the original model (Neutral col), but to a much lesser degree than the 10 examples model. As figure 1 shows, this model performs well on all three in-domain templates that appear in the inoculation dataset, while keeping the control tests above 0.8 acc. (control row, table 4). Further, although the results for out-of-domain templates range from

[3] See https://huggingface.co/transformers/training.html#trainer for that method. We finetune for 2 epochs, evaluating each epoch on the accuracy of the evaluation set, and using the AdamW optimizer and its default hyper-parameters.

0.3 to 1.0, this variance is the lowest among all other tests (see figure 2).

Analysis of the 250 examples inoculation (table 5). This model seems to associate factive verbs with contradictions, and non factive verbs with neutral. This pattern is consistent with the labels of the templates that were used in the inoculation process (note that in table 1, the one blue template with the modified label of contradiction contains factive verbs, while the other blue templates are non-factive verbs with a modified label of neutral). Of course, predicting a label just according to the factivity property of the verbs in the examples will results in bad accuracy in out-of-domain templates, where factivity does not play a role (like in the single agent filters that were not in the inoculation data). The bias for neutral and contradiction resulted in many 0.0 accuracy results (see ent. column).

This bias is further apparent in the drop of the single agent anaphora and single agent filter results, when looking at examples that include factive verbs. Here is an actual model prediction from the template of single agent neutral:

> **premise**: Joseph learns that four men are posing behind a cash register.
> **Hypothesis**: He learns that the men all know each other.
> **Model prediction**: contradiction≈ 0.99

The clauses of the premise and the hypothesis have gold label neutral (there is no reason to assume that the men know each other). Therefore, the modified examples should be neutral. We suspect that the model predicts contradiction because the filter modified the example with the factive verb *to learn*. Here is another actual example, now from the single agent entailment template:

> **Premise:** Sophia believes that a man puts his hands up while telling an amusing story to his friend with a beard.
> **Hypothesis:** Sophia believes that people share a conversation.
> **model prediction:** neutral ≈ 0.999

The gold label of this example is entailment: the content of Sophia's belief does imply that people are having a conversation. It seems, however, that the model predicts *neutral* because of the use of the non-factive verb *to believe*. This model heuristic explains the poor performance on the out-of domain templates (figure 2).

The model behaviour on control is completely destroyed, predicting contradictions constantly, although the control examples do not include any modifying filters. We have no explanation for this behaviour (see the appendix for a plot of the inoculated model results on the control tests).

5.3 Inoculation: General Analysis

We can now answer the three questions that motivated our inoculation process:
1) Yes, the model is able to quickly learn the structure of examples it sees in the inoculation process. As figure 1 shows, even after 50 examples the model reaches high accuracy scores on in-domain templates.
2) No, the model did not generalize from seen templates to unseen templates well. As figure 2 shows, at no step of the inoculation process does the model shows coherent overall improvement on all held-out templates. Moreover, the control test fails significantly after 250 examples (figure 3, appendix).
3) No, the model does not correctly generalize from seen verbs to unseen ones. As the analysis of the model after 250 examples showed, the model has settled on an incorrect correlation between the factivity of verbs and the label of the example

In retrospect, the inoculation design we chose was particularly challenging: holding out both templates and verbs, plus using very small data size. We hypothesise that increasing the number of templates in the inoculation process will result in better results. Picking a subset of templates for inoculation that does not allow the model to just memorize the label information in table 1, but still shows that the model has performed accurate generalizations, remains challenging.

6 Conclusions

Epistemic closure inferences constitute a very special type of inference, but, as we have argued, it plays an important role in coherent communication. Epistemic closure is also part of larger theory of mind ability, which is arguably crucial for many NLU tasks, such as inference, QA and conversation. In this project, we have demonstrated how to transform standard NLI data into epistemic closure inference data, and showed that the model RoBERTa-large-mnli struggles with such examples, including after limited attempt of inoculation.

Authorship Statement

This project was single-authored by Michael Cohen, without external collaborators. This project continues the author's line of work on theory of mind reasoning in NLI, which includes the CS224N Winter 2021 project *Exploring RoBERTa's theory of mind through textual entailment* (Cohen, 2021). This and the previous project are sufficiently different.

References

Jc Beall, Greg Restall, and Gil Sagi. 2019. Logical Consequence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2019 edition. Metaphysics Research Lab, Stanford University.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, S. Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *ArXiv*, abs/1908.05739.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Michael Cohen. 2021. Exploring roberta's theory of mind through textual entailment. *Stanford CS224n course project, Winter 2021*.

Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Vardi. 1995. *Reasoning About Knowledge*. MIT Press.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2018. Stress-testing neural models of natural language inference with multiply-quantified sentences.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Jaakko Hintikka. 1962. *Knowledge and Belief*. Ithaca: Cornell University Press.

Andrew David Irvine and Harry Deutsch. 2021. Russell's Paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.

Theo M. V. Janssen and Thomas Ede Zimmermann. 2021. Montague Semantics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213, Florence, Italy. Association for Computational Linguistics.

Lauri Karttunen. 1971. Some observations on factivity. *Paper in Linguistics*, 4(1):55–69.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. 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 5872–5877, Hong Kong, China. Association for Computational Linguistics.

David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1):339–359.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. Inoculation by fine-tuning: A method for analyzing challenge datasets.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Steven Luper. 2020. Epistemic Closure. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2020 edition. Metaphysics Research Lab, Stanford University.

Christopher D. Manning. 2006. Local textual inference: It's hard to circumscribe, but you know it when you see it – and nlp needs it.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

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.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Michael Nelson. 2019. Propositional Attitude Reports. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2019 edition. Metaphysics Research Lab, Stanford University.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Hazel Pearson. 2020. *Attitude Verbs*, pages 1–22. The Wiley Blackwell Companion to Semantics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Ian Ravenscroft. 2019. Folk Psychology as a Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2019 edition. Metaphysics Research Lab, Stanford University.

Rasmus Rendsvig and John Symons. 2021. Epistemic Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? 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 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online. Association for Computational Linguistics.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural nli models through veridicality.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Paul Égré. 2020. *Logical Omniscience*, pages 1–25. The Wiley Blackwell Companion to Semantics.

A Appendix: Control results for inoculation

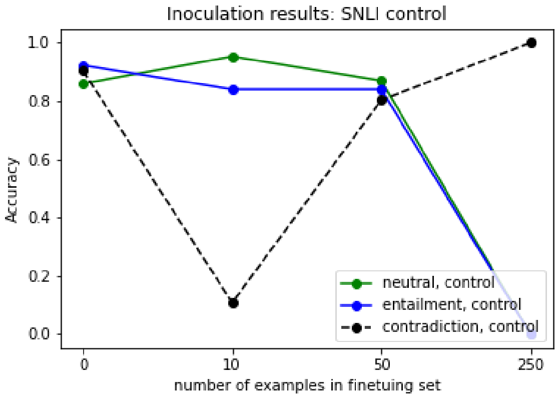

Figure 3: Inoculation results (10, 50, 250, examples) for SNLI control (unchanged) explaes.