DEPARTMENT OF COMPUTER SCIENCE
SERIES OF PUBLICATIONS C
REPORT C-2011-1

Cooperative Replies to Unbelievable Assertions:

A Dialogue Protocol Based on Logical Interpolation

Matti Nykänen, Satu Eloranta, Olli Niinivaara, Raul Hakli

UNIVERSITY OF HELSINKI
FINLAND

Cooperative Replies to Unbelievable Assertions: A Dialogue Protocol Based on Logical Interpolation

(Extended version of a paper with the same title presented at the Third International Conference on Agents and Artificial Intelligence (ICAART'11))

Matti Nykänen, Satu Eloranta, Olli Niinivaara, Raul Hakli

Department of Computer Science
P.O. Box 68, FIN-00014 University of Helsinki, Finland
satu.eloranta@cs.helsinki.fi
olli.niinivaara@cs.helsinki.fi

School of Computing, University of Eastern Finland
P.O. Box 1627, FIN-70211 Kuopio, Finland
matti.nykanen@uef.fi

Helsinki Institute for Information Technology (HIIT)
P.O. Box 68, FIN-00014 University of Helsinki, Finland
raul.hakli@cs.helsinki.fi

Technical report, Series of Publications C, Report C-2011-1
Helsinki, January 2011, 18 pages

Abstract

We propose a dialogue protocol for situations in which an agent makes to another agent an assertion that the other agent finds impossible to believe. In this interaction, unbelievable assertions are rejected using explanations formed by logical interpolation and new assertions are being made such that all previous rebuttals are taken into account.

Computing Reviews (1998) Categories and Subject Descriptors:

I.2.3 [Artificial Intelligence]: Deduction and Theorem Proving—nonmonotonic reasoning and belief revision

I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence—multiagent systems

F.4.1 [Mathematical Logic and Formal Languages]: Mathematical Logic—proof theory

General Terms:
Algorithms, Theory

Additional Key Words and Phrases:
Belief Revision, Dialogue Protocols, Argumentation

1 Introduction

When two agents carry out a conversation with each other, one of them may well assert something which the other cannot believe for some reason. Witness the following example (Hansson, 1991):

Conversation 1.

> **Amy:** Last summer I saw a three-toed woodpecker just outside my window. I could clearly see its red forehead and its red rump.
>
> **Bob:** You must be mistaken. A three-toed woodpecker does not have a red forehead or a red rump.
>
> **Amy:** You make me uncertain. Thinking about it, the only thing I am certain of is that the bird had a red forehead.

We study here such conversations: Both agents have beliefs that they are certain of and that they are not willing to give up during the conversation. Here, Bob's ornithological knowledge is one example, and Amy's certainty of seeing a bird with a red forehead is another. When hearing an unbelievable assertion, Bob faces the task of helping Amy by offering informative rebuttals. Amy then faces the task of generating another assertion while taking Bob's rebuttal into account. This interaction continues until either Amy comes up with an assertion which Bob can consider possible or she concludes that they have irreconcilable differences, at least as far as this conversation is concerned.

We consider these conversations in the context of *belief revision* (Alchourrón et al., 1985) in the presence of what we call *convictions*. By these convictions we mean those beliefs the agent refuses to give up, at least during the current conversation. In several fields there are important concepts that can be interpreted as convictions: In computer science, *integrity constraints* (Reiter, 1988) are needed to ensure consistency of databases. In philosophy, the properties of *knowledge* differ from those of belief (Hintikka, 1962), and people take a different stand on what they take to know and not merely believe. In nonprioritized belief revision, *core beliefs* are immune to revision (Hansson, 1999, gives a survey). In theories of argumentation, agents have *dark-side commitments*, which are their fundamental commitments that they find extremely hard to retract once stated in a conversation (Walton and Krabbe, 1995, pp. 11–12).

We encounter the problem of what an agent should do when another agent asserts something that conflicts with his convictions. In this paper we propose a solution, in which the agents carry out a conversation as an *interactive preparatory phase* before belief revision. In this phase they seek together a final assertion which does not conflict with either agent's convictions. In Conversation 1, Amy's second assertion might serve as something which they both might be able to believe. We do not consider what happens after this preparatory phase, that is, we do not concern ourselves whether the agents actually revise their epistemic states or not.

Focusing on these assertion-rebuttal conversations raises immediately three questions: First, how can an agent form his rebuttal to the unbelievable assertion? Second, how should the other agent form her[1] next assertion on the basis of her epistemic state while taking into account the new information in the rebuttal? And third, how can the conversation stay focused on its original subject?

Our answer to the first question is to use logical *interpolation*, since it gives a formula which is entailed by the convictions of the agent and entails the negation of the unbelievable assertion. Moreover, it can be read as a description how or an explanation why (Hintikka and Halonen,

[1] We adhere to the convention that the asserting agent (such as Amy in Conversation 1) is female, whereas the rebutting agent (such as Bob) is male, and refer to them as "she" and "he" as well as by name.

1999) the assertion conflicted with the convictions of the agent, thereby bringing some currently relevant part of his convictions into light. Our answer to the second question is to use *hypothetical* thinking, as if the agent thought: "if I were to believe this rebuttal, then I would have this belief about my topic instead of the one I expressed before". She will form her next assertion as the least disbelieved alternative to the topic given the new information. Our answer to the third question is to require that their utterances remain relevant to it in the *letter-sharing* sense (Makinson, 2009, Definition 1.1), which is guaranteed by interpolation.

In related work, there are some approaches in which agents have convictions but do not use interaction for conflict resolution. These include such approaches to nonprioritized belief revision that secure some beliefs from revision. For instance, in Accommodative Belief Revision (Eloranta et al., 2008), the agent tries to guess what the other would have said, had she had his knowledge. In our solution, the agent does not have to guess what the other agent would believe, instead he gives her a chance to tell it.

Then there are approaches in which interaction is used as a preprocessing step before belief revision, but the possibility of agents having their private convictions is not considered. These include such merging approaches as mutual belief revision (Jin et al., 2007) and belief negotiation (Booth, 2006) in which all the agents' beliefs are weakened until they no longer contradict each other. As opposed to that, our solution is asymmetric. We have one agent, who is eager to inform another agent about some of her beliefs, whereas the other agent is willing to reply and share some of his convictions in case he finds the original assertion unbelievable. Application areas with such a setting include knowledge base systems in which some agents (either human beings or software agents) collect information and send it to one agent acting as a knowledge base with integrity constraints.

Certain types of argumentation-based dialogues (Walton and Krabbe, 1995; Parsons et al., 2003) can also be viewed as preparatory phases for belief revision: They aim at finding out whether a particular assertion should be believed by exchanging information about arguments that either support or undermine it. In our approach, however, the goal is to find out *what* could be believed about the topic when the agents' convictions are taken into account, not *whether* a particular proposition should be believed or not.

For example, van Veenen and Prakken (2006) included asking "Why did you rebut my assertion?" among the moves in their negotiation protocol as an embedded persuasion game. However, their idea is to bring the grounds for the rebuttal to light so that they too can be subject to further scrutiny by the other agent within this conversation. In contrast, the purpose of our dialogues is not to persuade the other agent to accept the original assertion, but to find an alternative assertion that is acceptable.

Our aim is that the agents' assertions in the dialogues satisfy *the Cooperative Principle* presented by Grice (1989, Chapters 2 and 3) to govern conversations between cooperative agents. These maxims rule out the naive extremities to deal with an unacceptable input, that is, either to terminate the dialogue or to reply with all one knows about the subject. Something more is needed, thus we will propose the use of logical interpolation as a cooperative reply to an assertion that an agent is convinced to be false.

The paper is organized as follows. In section 2, we will introduce our notations and present the interpolation principle as a tool for generating cooperative replies to unbelievable assertions. In section 3, we will propose guidelines for generating a new modified assertion based on this reply. Section 4 presents the conversation protocol driven by these interpolants and shows that the conversations will always end with a rational outcome. In section 5, we will study how interpolants can be computed in this setting using sequent calculus (Negri and von Plato, 2001, Chapter 3.1). In section 6, we will develop two alternative ways to produce new assertions. In section 7, we will give conclusions and propose some directions for future research.

2 Cooperative replies

We will consider dialogues, that is, conversations between two agents. We assume that these two agents, named A and B, such as Amy or Bob in Conversation 1, have *epistemic states*, which we denote with $\mathcal{A}$ and $\mathcal{B}$ correspondingly. These states contain the *belief sets* consisting of all the beliefs they currently hold; these sets we denote with $\mathbb{B}(\mathcal{A})$ and $\mathbb{B}(\mathcal{B})$. In these belief sets, beliefs are expressed with formulas of classical propositional logic; that is, they are beliefs about the actual state of affairs, and not about for instance beliefs about each other's beliefs. On the one hand, each agent may be willing to give up some of these beliefs given new evidence to the contrary. On the other hand, (s)he may regard some of them as convictions which (s)he will hold on to, regardless of any such new evidence. We denote the sets of convictions for agents A and B with $\mathbb{C}(\mathcal{A})$ and $\mathbb{C}(\mathcal{B})$ correspondingly. We assume that the sets of beliefs and the sets of convictions are non-contradictory and deductively closed, that is, $\mathbb{B}(\mathcal{A}) = \mathrm{Cn}(\mathbb{B}(\mathcal{A}))$, etc. We also assume that what an agent is convinced of, (s)he also believes, that is, $\mathbb{C}(\mathcal{A}) \subseteq \mathbb{B}(\mathcal{A})$ and $\mathbb{C}(\mathcal{B}) \subseteq \mathbb{B}(\mathcal{B})$.

Agent A is the initiator of the conversation. We are focusing entirely on the situation, in which agent B is convinced that the assertion made by A is false. We are not concerned whether agent B will give priority to the assertion if it does not contradict his convictions.

We use lower-case Greek letters to denote propositional formulas. Agent A initiates the conversation with her initial assertion φ. Whenever the assertion is acceptable by agent B, the dialogue ends. Therefore we shall presume this initial assertion to be unbelievable, that is, $\mathbb{C}(\mathcal{B}) \models \neg\varphi$. As already mentioned in Section 1, we propose for agent B to use the interpolant to this entailment as a cooperative reply in this situation, since (i) it is entailed by $\mathbb{C}(\mathcal{B})$ and (ii) it entails $\neg\varphi$ and (iii) it only employs vocabulary that appears in the original assertion (in terms of propositional variables).

Let $\mathbb{V}$ denote all the propositional variables and $\mathrm{Voc}(\alpha) \subset \mathbb{V}$ those appearing in the formula α. Let us recall what interpolation is:

Theorem 1 (Craig interpolation, for propositional logic)**.** *Let* α *and* β *be two propositional formulas. If* $\alpha \models \beta$*, then there is some* interpolant θ *such that (i)* $\alpha \models \theta$*, (ii)* $\theta \models \beta$*, and (iii)* $\mathrm{Voc}(\theta) \subseteq \mathrm{Voc}(\alpha) \cap \mathrm{Voc}(\beta)$.

In section 5, we prove this theorem for the system G3cp of sequent calculus (Negri and von Plato, 2001, Chapter 3.1) in a way which provides an explicit algorithmic construction for the interpolant θ given the formulas α and β.

In Conversation 1, α is a formula representing (some relevant part of) Bob's convictions $\mathbb{C}(\mathcal{B})$, β is the negation $\neg\varphi$ of Amy's initial assertion φ, and θ is a suggestion for an explanation why α rules out β.

Example 1. Consider Conversation 1, and let the propositional variable p stand for "Amy saw a three-toed woodpecker", q stand for "Amy saw a bird with a red forehead", r stand for "Amy saw a bird with a red rump", s stand for "Amy saw a lark". The conversation starts when Amy asserts $p \wedge q \wedge r$. Assume Bob's convictions include $(p \vee s) \to (\neg q \wedge \neg r)$. Then $(p \vee s) \to (\neg q \wedge \neg r)$ entails an interpolant $p \to \neg q \wedge \neg r$, which again entails the negation of Amy's assertion, $\neg(p \wedge q \wedge r)$.

Let us consider how well this suggestion fares in light of *Grice's Maxims* (Grice, 1989, Chapters 2 and 3). These maxims elaborated his general Cooperative Principle into more specific conversational rules which the participants can be expected to observe:

Maxim of Quantity: (i) Make your contribution as informative as required (for the current purposes of the exchange). (ii) Do not make your contribution more informative than is required.

Maxim of Quality: Try to make a contribution which is true. More specifically: (i) Do not say what you believe to be false. (ii) Do not say that for which you lack adequate evidence.

Maxim of Relevance: Be relevant.

Maxim of Manner: Be clear.

Using an interpolant as a reply conforms to part (i) of the maxim of Quantity, because it conveys information that agent *A* supposedly was not aware of, since it entails the negation of what she said. In part (ii), the amount of informativeness can be controlled by the selection of a suitable interpolation formula.

Using an interpolant as the reply conforms also to the maxim of Quality: not only does agent *B* reply with a belief of his, but with a conviction, and we assume here that a rational agent does not obtain convictions without proper evidence.

According to the maxim of Relevance, a reply should somehow be related to the preceding conversation. A natural syntactic concept is letter-sharing: two formulas are relevant to each other, if they share some propositional variable (Makinson, 2009, Definition 1.1). In this regard, an interpolant is an extremely relevant reply, since it consists only of variables in both $\mathbb{C}(\mathcal{B})$ and the current suggestion by agent *A*, by property (iii) of Theorem 1. Makinson (2009) notes that although letter-sharing is not wholly unproblematic as a notion of relevance, it does have its uses in computational contexts. Hence we define here the *topic* of the conversation to be the variables $\mathrm{Voc}(\varphi)$ about which agent *A* wants to have a conversation with agent *B*.

Another more refined concept of relevance could be to split the beliefs of an agent into disjoint parts, where each part consists of the agent's beliefs about a particular subject matter (Parikh, 1999). Since Theorem 1 can be extended to such split theories (Kourosias and Makinson, 2007, Theorem 1.1), we anticipate our approach to apply in this setting as well.

There are also more semantic accounts of relevance. In the theory of Sperber and Wilson (2004), the main determinant of relevance of a reply is how much positive cognitive effect (such as learning, settling doubt, or correcting mistaken assumptions) it creates in the recipient. In this respect too, a reply by interpolant fares well since it contradicts with what agent *A* supposedly believes and should thus invoke a process of belief revision.

Regarding the maxim of Manner, one could argue that it does not concern conversations such as ours, where the messages exchanged are formulated in logic instead of in a natural language. We note, however, that even in our conversations agent *B* can tailor the form of his interpolant to enhance its clarity to agent *A*.

3 On generating new assertions

Let us contemplate on what agent *A* should do when she learns a formula θ that tells her why her previous assertion was unbelievable.

If θ conflicts with the convictions of agent *A*, we take that this dialogue should fail. If the rebuttal is not unbelievable to agent *A*, then she can either (i) accept the input as a conviction (since it was *B*'s conviction), (ii) accept the input as a belief, or (iii) treat the input conditionally. The treatment may depend, e.g., on how reliable the agent considers the other agent. By the maxim of Quality, agent *A* then continues the dialog with a new assertion ψ, which she accordingly

(i) believes (now that she is convinced that θ), (ii) believes (now that she has come to believe that θ), or (iii) would believe, if she were to believe that θ.

Denoting the doxastic conditional “if I were to believe θ, then I would also believe ψ” as $\theta \mathrel{\circ\!\!\rightarrow} \psi$, our requirement of the maxim of Quality becomes $\mathcal{A} \models \theta \mathrel{\circ\!\!\rightarrow} \psi$, where this entailment is defined through the *Ramsey test:* if agent A were to revise her epistemic state $\mathcal{A}$ with θ, then would ψ be believed in the resulting state $\mathcal{A} \circ \theta$? Thus in all three alternatives, A may anwer ψ if and only if $\psi \in \mathbb{B}(\mathcal{A} \circ \theta)$. Note that this revision might be only tentative: the actual epistemic state of agent A might still be $\mathcal{A}$.

Now let us assume that the revision operator '$\circ$' that agent A uses (either when revising her epistemic state or when evaluating conditionals) satisfies the basic rationality criteria (R1)–(R4) for belief revision (Alchourrón et al., 1985) and the rationality criterion (IR1) for iterated belief revision (Darwiche and Pearl, 1997). That is:

$$\alpha \in \mathbb{B}(\mathcal{A} \circ \alpha). \tag{R1}$$

$$\text{If } \neg\alpha \notin \mathbb{B}(\mathcal{A}) \text{ then } \mathbb{B}(\mathcal{A} \circ \alpha) = \mathrm{Cn}(\mathbb{B}(\mathcal{A}) \cup \{\alpha\}). \tag{R2}$$

$$\text{If } \alpha \text{ is satisfiable then } \mathbb{B}(\mathcal{A} \circ \alpha) \text{ is consistent.} \tag{R3}$$

$$\text{If } \alpha \equiv \beta \text{ then } \mathbb{B}(\mathcal{A} \circ \alpha) = \mathbb{B}(\mathcal{A} \circ \beta). \tag{R4}$$

$$\text{If } \alpha \models \beta \text{ then } \mathbb{B}(\mathcal{A} \circ \alpha) = \mathbb{B}((\mathcal{A} \circ \beta) \circ \alpha). \tag{IR1}$$

Postulate (R1) says that the new piece of information is accepted, that is, the insertion succeeds. Postulate (R2) says that if the new piece of information is compatible with the old beliefs, neither is any of them discarded nor is anything not entailed by the old beliefs and the new information added to the belief set. Postulate (R3) says that adding a satisfiable formula to the belief set must not make it inconsistent. Postulate (R4) calls for syntax independence. Postulate (IR1) says that if $\alpha \models \beta$, then the beliefs in the epistemic state obtained when learning first β and then α are the same as when learning just α in the first place.

By the maxim of Relevance, agent A must bear in mind (and take into account) all the rebutting formulas during the dialog. Let γ denote the conjunction of those formulas. Thus in the scenarios, we must use γ instead of θ.

By the basic postulates for belief revision, we have $\mathcal{A} \models \theta \mathrel{\circ\!\!\rightarrow} \psi$ if and only if $\mathcal{A} \circ \theta \models \theta \mathrel{\circ\!\!\rightarrow} \psi$. Now assume that after n rebuttals, $\gamma_n = \theta_1 \wedge \theta_2 \wedge \ldots \wedge \theta_n$. Then by postulate (IR1), we have $\mathcal{A} \models \gamma_n \mathrel{\circ\!\!\rightarrow} \psi$ if and only if $\mathcal{A} \circ \gamma_1 \circ \gamma_2 \circ \ldots \circ \gamma_n \models \gamma_n \mathrel{\circ\!\!\rightarrow} \psi$. Thus the truth value of the conditional does not depend on whether the agent has actually revised her epistemic state on the way or not: all three alternatives for A's actions remain equivalent in this respect.

Our framework allows for several methods for constructing a formula ψ satisfying this requirement. Some methods are discussed in section 6.

4 A conversation protocol

We will now give a conversation protocol in which agent B uses interpolation to create rebuttals. In our protocol, agent A starts with an assertion φ, which also fixes the topic of the conversation. When agent B receives an assertion which conflicts with $\mathbb{C}(\mathcal{B})$, he answers with an interpolant θ. In the protocol, ψ contains the most recent assertion made by agent A, and γ is the conjunction of all the rebuttals made by agent B to her previous assertions, as above. The protocol is depicted as follows:

Conversation Protocol

```
1  $\psi \leftarrow \varphi$; $\gamma \leftarrow \top$
2  $A$ asserts $\psi$
3  while $\mathbb{C}(\mathcal{B}) \models \neg\psi$ with some interpolant $\theta$
4      do $B$ replies that he is convinced of $\theta$
5         $\gamma \leftarrow \gamma \wedge \theta$
6         if $\mathbb{C}(\mathcal{A}) \models \neg\gamma$
7            then $A$ says that their convictions conflict with each other
8                 return FAIL
9         $\psi \leftarrow$ some formula chosen by $A$ such that $\mathcal{A} \models \gamma \mathrel{\circ\!\!\rightarrow} \psi$
10        $A$ asserts $\psi$
11 $B$ replies that he too considers this $\psi$ believable[a]
12 return SUCCESS with $\psi$
```

[a] By ψ being believable to agent B we mean that ψ is consistent with his convictions $\mathbb{C}(\mathcal{B})$ (but not necessarily with his beliefs $\mathbb{B}(\mathcal{B})$).

Selecting the next assertion ψ on line 9 is possible if and only if $\mathbb{C}(\mathcal{A}) \cup \{\gamma\}$ is consistent, and this is guaranteed by line 6. If our conversation protocol terminates successfully (on line 12) then we do have an agreement: both agents could believe the final assertion ψ. For agent B, this follows by line 3. For agent A, this is an invariant of the **while** loop: it holds before the loop by line 1 and the maxim of Quality (i), and it continues to hold after each execution of the loop body by line 9.

Notice the algorithm uses the current epistemic states of the agents. As to the convictions the agents have, this causes no problems, because the agents do not give up their convictions. The beliefs of agent B are not used in the protocol. As to the beliefs of agent A, whether she has actually revised her epistemic state with γ or not does not affect the truth value of the conditional, as discussed in section 3.

The protocol could generate a conversation like the following:

Conversation 2.

> **Amy:** I saw a bird with a red forehead and a red rump. It was a three-toed woodpecker!
>
> **Bob:** A three-toed woodpecker does not have a red forehead.
>
> **Amy** *(thinking to herself):* In that case I would have to give up believing either that I saw a three-toed woodpecker or a red forehead. I prefer to keep believing the former. Hence I can take Bob's rebuttal into account by giving up the latter. I can also keep believing in the red rump, since he did not challenge that part.
>
> *(aloud to Bob):* In that case I think I saw a three-toed woodpecker, which had a red rump but no red forehead.
>
> **Bob:** A three-toed woodpecker does not have a red rump either.
>
> **Amy** *(after similar thinking):* Well, in that case I think I saw a three-toed woodpecker, but it had neither a red rump nor a red forehead.
>
> **Bob:** Now *that* is something I could believe!

If our protocol terminates with failure instead (on line 8) then this outcome is warranted as well: On the one hand, $\mathbb{C}(\mathcal{B}) \models \gamma$ by property (i) of Theorem 1 and lines 3 and 5. On the other hand, $\mathbb{C}(\mathcal{A}) \models \neg\gamma$ by line 6. Agent A can even explain this conflict between their convictions with an interpolant corresponding to the entailment on line 6, thereby expressing her own convictions about the topic. This could be useful if the agents attempt to reconcile their convictions somehow, but we do not consider such attempts here.

We would also like to show that our protocol does not run forever. However, this is not entirely clear at the outset:

Example 2. Let s_m denote the proposition "$1+1=m$". Chas believes s_3 and holds the doxastic conditional $\neg s_3 \wedge \cdots \wedge \neg s_n \circ\!\!\rightarrow s_{n+1}$ for every $n \in \mathbb{N}$. He contacts his teacher Dave:

Chas: I assert that s_3.

Dave: No, since I know that $\neg s_3$. (Dave has no other interpolant at his disposal, by property (iii) of Theorem 1.)

Chas: In that case, s_4.

Dave: No.

Chas: In that case, s_5. *(ad infinitum.)*

The trouble in Example 2 is that Chas keeps introducing new propositional variable symbols into it. If such rampant conversational behaviour is precluded by assuming that agent A keeps her assertions ψ relevant to the topic in our chosen letter-sharing sense, then our protocol becomes finite, albeit $O(2^{|\mathrm{Voc}(\varphi)|})$ in the worst case.

Theorem 2 (Finiteness)**.** *Assume that the assertions* ψ *by agent A in our conversation protocol satisfy* $\mathrm{Voc}(\psi) \subseteq \mathrm{Voc}(\varphi)$. *Then the maximum number of times its* ***while*** *loop can be executed is*[2]

$$|\{w \restriction \mathrm{Voc}(\varphi) \colon w \in \mathrm{Mod}(\mathbb{C}(\mathcal{A})) \setminus \mathrm{Mod}(\mathbb{C}(\mathcal{B}))\}|.$$

Proof. By the assumption and property (iii) of Theorem 1, $\mathrm{Voc}(\gamma) \subseteq \mathrm{Voc}(\varphi)$ is yet another invariant of the **while** loop. Hence it suffices to show that each execution of line 5 eliminates at least one element from $\mathrm{Mod}(\gamma)$. This would not happen if we had $\gamma \models \theta$. We also have $\theta \models \neg\psi$ by property (ii) of interpolation in line 3. Taken together they yield $\gamma \models \neg\psi$, but this contradicts $\gamma \circ\!\!\rightarrow \psi$, and so A could not have selected such an assertion ψ on line 9.

The worst case is to eliminate as few elements from $\mathrm{Mod}(\gamma)$ at each round as possible. This happens as follows: first agent A guesses some $w \in \mathrm{Mod}(\mathbb{C}(\mathcal{A})) \cap \mathrm{Mod}(\gamma)$ and asserts a formula ψ corresponding to this one value combination $w \restriction \mathrm{Voc}(\varphi)$, and then agent B chooses $\neg\psi$ as his rebuttal θ. This eliminates just this one combination. The stated upper bound is in turn the maximal number of all combinations available for agent A. □

However, we expect actual conversations to terminate in much fewer rounds than the pessimal upper bound in Theorem 2, since the more precise explanations θ agent B gives for his rejections, the fewer assertions agent A needs. And indeed, maxim of Quantity (i) directs agent B towards such precise explanations θ. However, we leave constructing precise interpolants θ for later study.

5 Calculating the interpolant

Let us now consider how agent B can construct an interpolant θ for the entailment $\mathbb{C}(\mathcal{B}) \models \neg\psi$ rebutting the most recent assertion ψ from agent A on line 3 of our protocol in Section 4. At the end of Section 2, we furthermore stipulated that we wish this construction to follow the proof which agent B used to establish this entailment.

[2] Here $f \restriction D = \{\langle x, f(x)\rangle : x \in D\}$ denotes the restriction of the function f into the domain D.

When $\mathbb{C}(\mathcal{B})$ can be expressed as a formula, we can use a standard construction for θ; one is given in Section 5.1 below. The case where no such formula is available will in turn be considered in Section 5.2.

5.1 When the convictions are a formula

We shall present a construction on the *sequent calculus G3cp* (Negri and von Plato, 2001, Chapter 3.1). In this sequent calculus, the left and right contexts Γ and Δ are multisets (that is, sets with multiple distinct copies of the same element) of propositional formulas. The notations φ, Γ and Γ, φ denote such a multiset $\{\varphi\} \cup \Gamma$ where this copy of φ is singled out and Γ consists of all the other elements. We omit writing the empty multiset explicitly. The axioms take the form

$$p, \Gamma \Rightarrow \Delta, p \tag{Ax}$$

where $p \in \mathbb{V}$. The symbol $\bot$ stands for logical falsity, and its only inference rule is

$$\frac{}{\bot, \Gamma \Rightarrow \Delta}\ (L\bot).$$

Negation $\neg\varphi$ is then a shorthand for $\varphi \to \bot$, and logical truth $\top$ for $\neg\bot$. The rules for the other three connectives are given in Table 1.

	conjunction	disjunction	implication
left	$\dfrac{\varphi, \psi, \Gamma \Rightarrow \Delta}{\varphi \wedge \psi, \Gamma \Rightarrow \Delta}\ (L\wedge)$	$\dfrac{\varphi, \Gamma \Rightarrow \Delta \quad \psi, \Gamma \Rightarrow \Delta}{\varphi \vee \psi, \Gamma \Rightarrow \Delta}\ (L\vee)$	$\dfrac{\Gamma \Rightarrow \Delta, \varphi \quad \psi, \Gamma \Rightarrow \Delta}{\varphi \to \psi, \Gamma \Rightarrow \Delta}\ (L\to)$
right	$\dfrac{\Gamma \Rightarrow \Delta, \varphi \quad \Gamma \Rightarrow \Delta, \psi}{\Gamma \Rightarrow \Delta, \varphi \wedge \psi}\ (R\wedge)$	$\dfrac{\Gamma \Rightarrow \Delta, \varphi, \psi}{\Gamma \Rightarrow \Delta, \varphi \vee \psi}\ (R\vee)$	$\dfrac{\varphi, \Gamma \Rightarrow \Delta, \psi}{\Gamma \Rightarrow \Delta, \varphi \to \psi}\ (R\to)$

Table 1: Rules of the G3cp system.

Agent B establishes $\mathbb{C}(\mathcal{B}) \models \neg\psi$ by constructing a corresponding derivation $\mathcal{D}$ in G3cp, and then extracts his interpolant θ from this $\mathcal{D}$ using the method explained in the proof for Theorem 3 below, which reformulates Theorem 1 for G3cp.

Example 3. In Example 1, Amy asserted that $p \wedge q \wedge r$, but Bob's was convinced that $p \to \neg q \wedge \neg r$. Based on this conviction, Bob can reason against Amy's assertion using G3cp as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{\overset{\text{(Ax)}}{p, q, r \Rightarrow \bot, p} \qquad \dfrac{\dfrac{\overset{\text{(Ax)}}{\neg r, p, q, r \Rightarrow \bot, q} \quad \dfrac{}{\bot, \neg r, p, q, r \Rightarrow \bot}\ (L\bot)}{\neg q, \neg r, p, q, r \Rightarrow \bot}\ (L\to)}{\neg q \wedge \neg r, p, q, r \Rightarrow \bot}\ (L\wedge)}{p \to \neg q \wedge \neg r, p, q, r \Rightarrow \bot}\ (L\to)}{p \to \neg q \wedge \neg r, p, q \wedge r \Rightarrow \bot}\ (L\wedge)}{p \to \neg q \wedge \neg r, p \wedge q \wedge r \Rightarrow \bot}\ (L\wedge)}{p \to \neg q \wedge \neg r \Rightarrow \neg(p \wedge q \wedge r)}\ (R\to).$$

Theorem 3. *Let* α *and* β *be two formulas such that* $\alpha \Rightarrow \beta$ *has a derivation* $\mathcal{D}$ *in G3cp. Then we can construct an interpolant* θ *such that (i)* $\alpha \Rightarrow \theta$ *has a derivation, (ii)* $\theta \Rightarrow \beta$ *has a derivation, and (iii)* $\mathrm{Voc}(\theta) \subseteq \mathrm{Voc}(\alpha) \cap \mathrm{Voc}(\beta)$.

Proof. We will construct θ satisfying properties (i) and (ii) by induction on the height of $\mathcal{D}$ for $\alpha \Rightarrow \beta$, similarly to Troelstra and Schwichtenberg (2000, Chapter 4.4.2).

We can attach to each formula occurrence φ within $\mathcal{D}$ a unique *label of origin* which tells whether φ originated from α (label 1) or β (label 0). This is possible because G3cp has the subformula property (Negri and von Plato, 2001, Corollary 3.2.4) and it uses multisets. We denote these labels ℓ with superscripts. The given sequent gets the labelling $\alpha^1 \Rightarrow \beta^0$, and if $(\varphi \oplus \psi)^\ell$ is the principal formula of a step in $\mathcal{D}$, then its subformulas φ^ℓ and ψ^ℓ will inherit the same label ℓ in the premiss(es). With this labelling in place, we can state our generalized inductive hypothesis as follows:

> Let $\Gamma \Rightarrow \Delta$ have a derivation. Then we can construct for it an interpolant θ such that both (i) $\Gamma^1 \Rightarrow \Delta^1, \theta$ and (ii) $\theta, \Gamma^0 \Rightarrow \Delta^0$ have derivations. (1)

Here Γ^1 consists of those elements from Γ which are labelled with 1, and so on.

Axiom (Ax)**.** The given derivation has the form $p^l, \Gamma \Rightarrow \Delta, p^r$. The corresponding interpolant can be formed as follows:

l	r	θ
1	1	$\bot$
1	0	p
0	1	$\neg p$
0	0	$\top$

(2)

Consider for instance the case where $l = 0$ and $r = 1$. Then Eq. (2) yields $\neg p$ as θ. This satisfies property (i), because we do have the required derivation

$$\dfrac{\overset{\text{(Ax)}}{p, \Gamma^1 \Rightarrow \Delta^1, p^1, \bot}}{\Gamma^1 \Rightarrow \Delta^1, p^1, p \to \bot}\,(R\to)$$

while property (ii) is satisfied by having

$$\dfrac{\overset{\text{(Ax)}}{p^0, \Gamma^0 \Rightarrow \Delta^0, p} \qquad \dfrac{}{\bot, p^0, \Gamma^0 \Rightarrow \Delta^0}\,(L\bot)}{p \to \bot, p^0, \Gamma^0 \Rightarrow \Delta^0}\,(L\to).$$

The required derivations for the other three combinations of l and r can be obtained similarly.

Property (iii) will also be satisfied, because Eq. (2) will be the only place which introduces a predicate symbol occurrence into the interpolant being constructed, and it does so only when $l \neq r$, which guarantees that the introduced p originates from both α and β.

The zero-premiss rule $(L\bot)$**.** The given derivation is of the form

$$\dfrac{}{\bot^\ell, \Gamma \Rightarrow \Delta}\,(L\bot). \tag{3}$$

If $\ell = 1$, then we can choose $\bot$ as θ; otherwise $\top$. These choices can be verified as in Eq. (2).

The single-premiss rules $(L\wedge)$, $(R\vee)$ and $(R\rightarrow)$. We can choose the interpolant θ' obtained by induction from the single subderivation as our θ for the whole given derivation.

Let for instance the given derivation have the form

$$\frac{\begin{array}{c}\vdots\\ \varphi^0,\Gamma\Rightarrow\Delta,\psi^0\end{array}}{\Gamma\Rightarrow\Delta,(\varphi\rightarrow\psi)^0}\;(R\rightarrow). \tag{4}$$

Induction provides us with some θ' which has derivations for (i) $\Gamma^1\Rightarrow\Delta^1,\theta'$ and (ii) $\theta',\varphi^0,\Gamma^0\Rightarrow\Delta^0,\psi^0$. The former is directly what property (i) requires. The latter is almost what property (ii) requires; we only need to reintroduce the final $(R\rightarrow)$ step, as in Eq. (4).

When the principal formula is $(\varphi\rightarrow\psi)^1$ instead, these roles change: property (ii) follows directly by induction, while property (i) needs the final step.

The other rules can be handled in the same way, with $(L\wedge)$ or $(R\vee)$ instead of $(R\rightarrow)$ in Eq. (4).

The two-premiss rules $(L\vee)$, $(R\wedge)$ and $(L\rightarrow)$. We can choose

$$\theta=\begin{cases}\theta_\varphi\wedge\theta_\psi & \text{if the principal formula } \phi\oplus\psi \text{ has label 0}\\ \theta_\varphi\vee\theta_\psi & \text{if it has label 1 instead}\end{cases} \tag{5}$$

where $\oplus$ denotes the main connective of the principal formula, and θ_φ and θ_ψ are the two interpolants obtained inductively for the two subderivations.

Let for instance the principal formula be $(\varphi\rightarrow\psi)^1$. Then the given derivation has the form

$$\frac{\begin{array}{c}\vdots\\ \Gamma\Rightarrow\Delta,\varphi^1\end{array}\qquad\begin{array}{c}\vdots\\ \psi^1,\Gamma\Rightarrow\Delta\end{array}}{(\varphi\rightarrow\psi)^1,\Gamma\Rightarrow\Delta}\;(L\rightarrow). \tag{6}$$

Induction on the leftmost subderivation provides an interpolant θ_φ which has a derivation $(\mathrm{i})_\varphi$ for $\Gamma^1\Rightarrow\Delta^1,\varphi^1,\theta_\varphi$ and a derivation $(\mathrm{ii})_\varphi$ for $\theta_\varphi,\Gamma^0\Rightarrow\Delta^0$. Induction on the rightmost subderivation provides an interpolant θ_ψ which has a derivation $(\mathrm{i})_\psi$ for $\psi^1,\Gamma^1\Rightarrow\Delta^1,\theta_\psi$ and a derivation $(\mathrm{ii})_\psi$ for $\theta_\psi,\Gamma^0\Rightarrow\Delta^0$. The chosen interpolant is their disjunction $\theta=\theta_\varphi\vee\theta_\psi$. We can construct the derivation required for property (i) as

$$\cfrac{\cfrac{\overbrace{\Gamma^1\Rightarrow\Delta^1,\varphi^1,\theta_\varphi,\theta_\psi}^{\text{derivation (i)}_\varphi}\quad\overbrace{\psi^1,\Gamma^1\Rightarrow\Delta^1,\theta_\psi,\theta_\varphi}^{\text{derivation (i)}_\psi}}{(\varphi\rightarrow\psi)^1,\Gamma^1\Rightarrow\Delta^1,\theta_\varphi,\theta_\psi}\;(L\rightarrow)}{(\varphi\rightarrow\psi)^1,\Gamma^1\Rightarrow\Delta^1,\theta_\varphi\vee\theta_\psi}\;(R\vee) \tag{7}$$

where the leftmost subderivation is obtained by weakening (Negri and von Plato, 2001, Theorem 3.2.1) derivation $(\mathrm{i})_\varphi$ with the missing θ_ψ, and the rightmost by weakening $(\mathrm{i})_\psi$ with θ_φ. We can also construct the derivation required for property (ii) by

$$\frac{\text{derivation (ii)}_\varphi\quad\text{derivation (ii)}_\psi}{\theta_\varphi\vee\theta_\psi,\Gamma^0\Rightarrow\Delta^0}\;(L\vee). \tag{8}$$

When the principal formula is $(\varphi\rightarrow\psi)^0$ instead, then the roles change: The derivation for property (i) can be constructed similarly to Eq. (8) with the chosen interpolant $\theta=\theta_\varphi\wedge\theta_\psi$ on the right side of '$\Rightarrow$', while the derivation for property (ii) can be constructed similarly to Eq. (7) with θ on the left. That is, we employ the symmetry that disjunction on the right behaves like conjunction on the left and vice versa.

The other rules can be handled in the same way, with $(L\lor)$ or $(R\land)$ instead of $(L\to)$ in Eq. (7). □

Example 4. Adding the labels of origin (0 for Amy and 1 for Bob) to the proof in Example 3 yields the following:

the sequent on top of the proof	interpolant
$p^0, q^0, r^0 \Rightarrow \bot, p^1$	$\neg p$, by Eq. (2)
$\neg r^1, p^0, q^0, r^0 \Rightarrow \bot, q^1$	$\neg q$, by Eq. (2)
$\bot^1, \neg r^1, p^0, q^0, r^0 \Rightarrow \bot$	$\bot$, by Eq. (3)

(The $\bot$ on the right caused by Bob negating what Amy said does not need a label, because it cannot take part in any inference steps.) These three interpolants are then connected together with the disjunctions corresponding to the $(L\to)$ steps as the overall interpolant $\neg p \lor (\neg q \lor \bot)$, which is equivalent to $\neg p \lor \neg q$, which in turn corresponds to Bob replying "either you did not see a three-toed woodpecker at all or you did not see a red forehead on it".

The interpolant θ constructed in the proof of Theorem 3 given above represents the branching structure of the given derivation $\mathcal{D}$ in the sense that each '$\lor$' represents branching due to α and each '$\land$' due to β, by Eq. (5). Negations can appear only in front of proposition symbols; that is, θ will be in *Negation Normal Form (NNF)*.

However, we may wish to follow the maxim of Manner even more closely, and obtain an alternative interpolant θ' which represents the branching structure of $\mathcal{D}$ in even more detail. In particular, we may wish to represent an $(L\to)$ step whose principal formula is $(\varphi \to \psi)^1$, as in Eq. (6), with an implication instead of turning it into a disjunction. We may namely wish our θ' to represent the portions originating from α as faithfully as we can, because in our setting α represents the convictions $\mathbb{C}(\mathcal{B})$ of agent B. We can indeed create such a more faithful θ' if we prove Theorem 3 in a slightly more elaborate way.[3]

An alternative proof sketch for Theorem 3. We introduce into the given proof a new parameter, namely the *polarity* π of the subformula being constructed. Informally, this π tells whether or not it will appear negated within the whole interpolant being constructed. Given that negation is a shorthand in G3cp, the formal definition of π is to count those implications '$\delta \to \dots$' where it will appear in the left side δ, modulo 2. Its initial value will thus be $\pi = 0$. (This and rule $(R\to)$ also explain why we chose the label 1 for α and 0 for β in our original proof.) Furthermore, we replace label 0 with π and label 1 with $1-\pi$ in our inductive hypothesis (1) and in the cases of our proof for it.

The two-premiss rules now have an alternative way to proceed besides the one already given: choose $\theta = \theta_\varphi \to \theta_\psi$ where θ_ψ is obtained by induction with the current value of π as before, but θ_φ with *the opposite value* $1-\pi$ instead.

Consider for instance the case where the principal formula is $(\varphi \to \psi)^{1-\pi}$. We are given a

[3] This way was originally mandated by intuitionistic logics, where implications cannot be considered as shorthands as in classical logics (Troelstra and Schwichtenberg, 2000, Chapter 4.4.2).

derivation

$$\dfrac{\begin{array}{cc}\vdots & \vdots\\ \Gamma\Rightarrow\Delta,\varphi^{1-\pi} & \psi^{1-\pi},\Gamma\Rightarrow\Delta\end{array}}{(\varphi\to\psi)^{1-\pi},\Gamma\Rightarrow\Delta}\;(L\to).$$

Induction on the rightmost subderivation with π provides an interpolant θ_ψ which has a derivation $(\mathrm{i})_\psi$ for $B^{1-\pi},\Gamma^{1-\pi}\Rightarrow\Delta^{1-\pi},\theta_\psi$ and a derivation $(\mathrm{ii})_\psi$ for $\theta_\psi,\Gamma^\pi\Rightarrow\Delta^\pi$. Induction on the leftmost subderivation with $1-\pi$ provides an interpolant θ_φ which has a derivation $(\mathrm{i})_\varphi$ for $\Gamma^\pi\Rightarrow\Delta^\pi,\theta_\varphi$ and a derivation $(\mathrm{ii})_\varphi$ for $\theta_\varphi,\Gamma^{1-\pi}\Rightarrow\Delta^{1-\pi},\varphi^{1-\pi}$; note how the labels got reversed. This allows us to construct the derivation required by property (i) similarly to Eq. (7) as

$$\dfrac{\dfrac{\overbrace{\theta_\varphi,\Gamma^{1-\pi}\Rightarrow\Delta^{1-\pi},\varphi^{1-\pi}}^{\text{derivation (ii)}_\varphi},\theta_\psi \qquad \theta_\varphi,\overbrace{\psi^{1-\pi},\Gamma^{1-\pi}\Rightarrow\Delta^{1-\pi},\theta_\psi}^{\text{derivation (i)}_\psi}}{\theta_\varphi,(\varphi\to\psi)^{1-\pi},\Gamma^{1-\pi}\Rightarrow\Delta^{1-\pi},\theta_\psi}\;(L\to)}{(\varphi\to\psi)^{1-\pi},\Gamma^{1-\pi}\Rightarrow\Delta^{1-\pi},\theta_\varphi\to\theta_\psi}\;(R\to).$$

The derivation required for property (ii) can in turn be constructed as

$$\dfrac{\text{derivation (i)}_\varphi \quad \text{derivation (ii)}_\psi}{\theta_\varphi\to\theta_\psi,\Gamma^\pi\Rightarrow\Delta^\pi}\;(L\to)$$

similarly to Eq. (8).

The other rules stem from the replacements above. For instance, Eq. (3) becomes "choose $\bot$ as θ if $\ell = 1-\pi$, otherwise $\top$" and so on. □

Example 5. If we apply this alternative proof for Theorem 3 to the lower $(L\to)$ step in Example 4, we get $p\to\neg q$ as the interpolant. This in turn has an intuitive reading where Bob says "a three-toed woodpecker does not have a red forehead".

5.2 When the convictions are a theory

Let us now turn to the case where the convictions $\mathbb{C}(\mathcal{B})$ of agent B are not expressed with a single formula α, but are given as some infinite theory $\mathbb{T}$ instead. This does not pose any problems to interpolation as such: When $\mathbb{T}\models\beta$, there still exists some interpolant formula θ such that (i) $\mathbb{T}\models\theta$, (ii) $\theta\models\beta$, and (iii) $\mathrm{Voc}(\theta)\subseteq\mathrm{Voc}(\mathbb{T})\cap\mathrm{Voc}(\beta)$, where $\mathrm{Voc}(\mathbb{T})\subseteq\mathbb{V}$ denotes all the non-logical symbols which appear in $\mathbb{T}$.

However, property (iii) must be viewed with some suspicion in our setting. For instance, if the propositional variable $x\in\mathbb{V}$ makes its only appearance in $\mathbb{T}$ within the tautology $x\vee\neg x$, then $\mathbb{T}$ doesn't really "say" anything about x; that is, its appearance is just a syntactic quirk. And now that agent B is explaining something to agent A based on $\mathbb{T}=\mathbb{C}(\mathcal{B})$, then this explanation should use only those variables about which this $\mathbb{T}$ does really say something; that is, such quirks should be ruled out.

Formula-variable independence (Lang et al., 2003, Definition 4) rules out such quirks from a single formula: The formula α is independent from x if and only if there exists an equivalent formula α' such that $x\notin\mathrm{Voc}(\alpha')$. Conversely, it turns out that α is dependent of x if and only if x appears in some prime implicate of α (Lang et al., 2003, Proposition 9).

Recall at this point the following standard logical notions: A *literal* is an occurrence of a propositional symbol x or its negation $\neg x$. A *clause* $\mathcal{C}$ is a finite set of literals, representing their disjunction. This $\mathcal{C}$ is an *implicate* of α if $\alpha \models \mathcal{C}$. It is a *prime* implicate, if it has no proper subset $\mathcal{C}' \subsetneq \mathcal{C}$ which would also be an implicate of α.

This suggests an immediate extension to the infinite: The theory $\mathbb{T}$ is dependent on x if and only if $x \in \mathrm{Voc}(\mathcal{C})$ for some prime implicate $\mathcal{C}$ of $\mathbb{T}$ — but now $\mathbb{T}$ can have infinitely many such $\mathcal{C}$. Then we can replace the suspicious property (iii) with the modified property that (iii′) $\mathbb{T}$ must be dependent on every variable symbol in $\mathrm{Voc}(\theta) \subseteq \mathrm{Voc}(\beta)$.

Since we want to actually compute such an explicit interpolant θ given $\mathbb{T}$ and β, we must assume something about the computational properties of $\mathbb{T}$ itself. It suffices to assume that we can compute the YES or NO answers to questions of the form "Is this clause $\mathcal{C}$ an implicate of $\mathbb{T}$?" or equivalently "Can $\mathbb{T}$ rule out those states of affairs where every literal in $\mathcal{C}$ is false?"

With this assumption, our computation can proceed as follows: We start proving $\Rightarrow \beta$ in G3cp using $\mathbb{T}$ as the *background theory*. That is, we proceed by eliminating connectives from β. At each sequent $\Gamma \Rightarrow \Delta$, we can not only use all the rules in G3cp but also ask questions "Is this clause $\mathcal{Q}$ an implicate of $\mathbb{T}$?" where

$$\mathcal{Q} \subseteq \{\neg x : x \in \Gamma \cap \mathbb{V}\} \cup (\Delta \cap \mathbb{V}). \tag{9}$$

If we get the answer YES, then we can consider this branch of the proof to be completed by the background theory $\mathbb{T}$, because $\mathbb{T} \models \mathcal{Q}$ and $\mathcal{Q} \models \Gamma \Rightarrow \Delta$ by construction. When we have completed every branch of our proof, we can choose θ to be the conjunction of all such $\mathcal{Q}$ answered YES during the proof. This choice satisfies properties (i) and (ii) by construction. For property (iii′), we furthermore stipulate that when we are asking these questions, we ask about $\mathcal{Q}$ only after we have asked about all its proper subsets $\mathcal{Q}' \subsetneq \mathcal{Q}$ but received the answer NO for them, since this guarantees that this implicate $\mathcal{Q}$ is indeed prime.

Moreover, we can rewrite each clause (9) as

$$\bigwedge(\Gamma \cap \mathbb{V}) \to \bigvee(\Delta \cap \mathbb{V}) \tag{10}$$

and then read our θ as "Agent B is convinced of these rules (10), among others, and he used them to rebut ψ".

Example 6. Reconsider Example 1 and assume that Bob's convictions $\mathbb{C}(\mathcal{B})$ are treated as a background theory which can answer questions like (9). Now Bob's proof begins with

$$\dfrac{\dfrac{\dfrac{p, q, r \Rightarrow \bot}{p, q \wedge r \Rightarrow \bot}\,(L\wedge)}{p \wedge q \wedge r \Rightarrow \bot}\,(L\wedge)}{\Rightarrow \neg(p \wedge q \wedge r)}\,(R\to)$$

at which point the following questions can be posed[4]:

clause $\mathcal{Q}$	the question in words	answer
$\{\neg p\}$	"Does $\mathbb{C}(\mathcal{B})$ rule out Amy having seen a three-toed woodpecker?"	NO
$\{\neg q\}$	"Does $\mathbb{C}(\mathcal{B})$ rule out Amy having seen a red forehead?"	NO
$\{\neg p, \neg q\}$	"Does $\mathbb{C}(\mathcal{B})$ rule out Amy having seen a three-toed woodpecker and a red forehead?"	YES!

Hence the interpolant θ can be constructed as a conjunction containing only the clause $\{\neg p, \neg q\}$; that is, $\neg p \vee \neg q$ as before. Its reading as a rule like (10) is in turn $(p \wedge q) \to \bot$, which is $\neg(p \wedge q)$. These questions served to extract from $\mathbb{C}(\mathcal{B})$ enough information for rebutting what Amy says.

[4] We omit the question $\mathcal{Q} = \emptyset$ corresponding to $\mathbb{C}(\mathcal{B}) \models \bot$ or "Is $\mathbb{C}(\mathcal{B})$ inconsistent?" as redundant.

6 Calculating new assertions

We now consider how agent A could continue the conversation after agent B has rejected the previous assertion, that is, how to produce the new assertion ψ on line 9 of our protocol in section 4. Two restrictions limit the choices of agent A: On the one hand, the doxastic conditional $\gamma \circ\!\!\rightarrow \psi$ says that agent A must choose the next assertion ψ from the beliefs $\mathbb{B}(\mathcal{A} \circ \gamma)$ she would have after revising tentatively her epistemic state $\mathcal{A}$ with all the rejections γ from B so far. On the other hand, Theorem 2 requires that her choice must satisfy $\mathrm{Voc}(\psi) \subseteq \mathrm{Voc}(\varphi)$ where φ is the original topic of their conversation.

The first method described in section 6.1 emphasizes the maxim of Quantity by describing the beliefs in $\mathbb{B}(\mathcal{A} \circ \gamma)$ regarding the topic as accurately as possible. The other method described in section 6.2 emphasizes in turn the maxim of Relevance by focusing on the rebuttals γ given by agent B.

6.1 A truth-table-based method

One method to choose this ψ is to enumerate the true rows of the truth table for $\mathbb{B}(\mathcal{A} \circ \gamma)$ restricted to the topic $\mathrm{Voc}(\varphi)$. That is, first construct the set

$$\Lambda_\varphi = \left\{ \bigwedge_{x \in \mathrm{Voc}(\varphi)} \ell_x : \ell_x \in \{x, \neg x\} \right\}$$

of all the different conjunctions of those literals ℓ_x which mention exactly the variables x in the topic. Then retain only those which remain consistent with the beliefs:

$$\psi = \bigvee \left\{ \lambda \in \Lambda_\varphi : \lambda \text{ is consistent with } \mathbb{B}(\mathcal{A} \circ \gamma) \right\}. \tag{11}$$

This result will be in *Disjunctive Normal Form (DNF)*. Note that this ruling out can be applied also when $\mathbb{B}(\mathcal{A} \circ \gamma)$ is an infinite theory, as in section 5.2.

Reconsider Conversation 2, now using the propositional symbols introduced in Example 1:

Conversation 3.

Amy: $p \wedge q \wedge r$.

Bob: $\neg(p \wedge q)$.

Amy: $p \wedge \neg q \wedge r$.

> (Amy constructs this reply by cycling through all the eight possible conjunctions of literals built from the variables p, q and r in the topic (that is, $p \wedge q \wedge r$, $p \wedge q \wedge \neg r$, $p \wedge \neg q \wedge r, \ldots, \neg p \wedge \neg q \wedge \neg r$) and retaining the one(s) which are not ruled out by $\mathbb{B}(\mathcal{A} \circ \neg(p \wedge q))$ — in this case, just one.)

Bob: $\neg(p \wedge r)$.

Amy: $p \wedge \neg q \wedge \neg r$.

> (Again, Amy cycles through the eight possibilities and reports the one not ruled out by $\mathbb{B}(\mathcal{A} \circ \neg(p \wedge q) \wedge \neg(p \wedge r))$.)

Bob: OK.

6.2 An interpolation-based method

Another method for constructing ψ can be obtained by using interpolation on φ and γ instead of enumeration.

The AGM success postulate (R1) for the revision '$\circ$' in the doxastic conditional yields

$$\mathbb{B}(\mathcal{A} \circ \gamma) \models \gamma \tag{12}$$

so it would be natural to choose any corresponding interpolant θ_γ as the next assertion ψ by A, since this would yield us (i) $\mathbb{B}(\mathcal{A} \circ \gamma) \models \theta_\gamma$ and (ii) $\theta_\gamma \models \gamma$. Property (i) would namely ensure the doxastic conditional. Property (ii) accounts in turn for one aspect of the maxim of Relevance: the next assertion by A does include what B has told her. So far, so good.

The problem with θ_γ is that A is now explaining why she could believe what she just heard, whereas we would prefer A to explain instead what she would believe about the original topic φ in light of this new information γ from B. This shortcoming can be seen already on a technical level: $\theta_\gamma \subseteq \mathrm{Voc}(\gamma)$, but it may happen that this $\mathrm{Voc}(\gamma) \subsetneq \mathrm{Voc}(\varphi)$.

One solution is to replace Eq. (12) with

$$\mathbb{B}(\mathcal{A} \circ \gamma) \models \varphi * \gamma \tag{13}$$

where '$*$' is some operator for revising the formula φ on its left with the other formula γ on its right (Bienvenu et al., 2008; Bittencourt et al., 2004; Dalal, 1988, among others). That is, A asks introspectively "what would my original topic φ be like in view of this new information γ, and why would I believe that?" Then choosing a corresponding interpolant $\theta_{\varphi*\gamma}$ as the next assertion ψ by A yields us (i) $\mathbb{B}(\mathcal{A} \circ \gamma) \models \theta_{\varphi*\gamma}$ and (ii) $\theta_{\varphi*\gamma} \models \varphi * \gamma$. Again, property (i) ensures the doxastic conditional, and property (ii) accounts for the maxim of Relevance, since the AGM success postulate for '$*$' (Darwiche and Pearl, 1997, Postulate (R1)) yields $\varphi * \gamma \models \gamma$.

Example 7. Reconsider Conversation 3 after Bob's first reply. If Amy now formed her second assertion by Eq. (12), she would be able to state only her preferred choice(s) among $p \wedge \neg q$, $\neg p \wedge q$ or $\neg p \wedge \neg q$. In particular, she encounters the aforementioned shortcoming in not being able to say anything about r, since Bob omitted it from his first reply. This shortcoming would then continue to plague Bob's second reply as well.

But since Amy forms her second assertion by Eq. (13) instead, she can state her preference(s) among $p \wedge \neg q \wedge r$ or $\neg p \wedge q \wedge r$. The other choices, such as $\neg p \wedge \neg q \wedge r$ or those with $\neg r$, are eliminated by '$*$' because they are more distant from her original message than the first two (assuming e.g. the Hamming distance and the operator by Dalal (1988)).

However, Eq. (13) has a problem too: it might not necessarily hold in general! One solution to this problem would be to choose

$$\psi = \begin{cases} \theta_{\varphi*\gamma} & \text{if Eq. (13) does hold} \\ \theta_\gamma & \text{otherwise} \end{cases} \tag{14}$$

because Eq. (12) guarantees that at least the latter does exist. This solution can be developed further into choosing as ψ the interpolant $\theta_{(\varphi*\gamma)\vee\gamma}$ for $\mathbb{B}(\mathcal{A} \circ \gamma) \models (\varphi * \gamma) \vee \gamma$ and stipulating that we always prefer the left over the right disjunct in its construction. This develoment gives our proposed method for generating the next assertion by A:

1. First agent A revises her beliefs $\mathbb{B}(\mathcal{A})$ tentatively with the comments γ made by B to form $\mathbb{B}(\mathcal{A} \circ \gamma)$.

2. Then agent A tries to prove the corresponding sequent $\mathbb{B}(\mathcal{A} \circ \gamma) \Rightarrow (\varphi * \gamma) \vee \gamma$ without touching its right disjunct γ, using the methods from sections 5.1 and 5.2, as appropriate.

3. If this proof does not go through, then the resulting still incomplete proof tree contains some still unproved sequents of the form $\Gamma \Rightarrow \Delta, \gamma$ where $\Gamma \cup \Delta$ contains no more connectives to eliminate. Agent A proves them too by eliminating (some of) the connectives of γ, as appropriate.

4. Finally agent A chooses the interpolant corresponding to the now completed proof tree as her ψ.

This methods improves over Eq. (14) because it falls back to using γ for only those parts of the desired interpolant which could not be derived already from $\varphi * \gamma$, if any.

This interpolation-based method seems to be computationally superior to the truth-table-based method in section 6.1, because the latter involves many satisfiability checks with respect to $\mathbb{B}(\mathcal{A} \circ \gamma)$ in its Eq (11). Granted, computing $\varphi * \gamma$ might also involve hidden satisfiability checks (but not necessarily: for instance Bittencourt et al., 2004, manage with just one check) but this computation involves only formulas uttered during the conversation, and not $\mathbb{B}(\mathcal{A} \circ \gamma)$ as a whole. However, if $\mathbb{B}(\mathcal{A} \circ \gamma)$ is an infinite theory, as in section 5.2, then step 2 of this method does involve satisfiability checks involving $\mathbb{B}(\mathcal{A} \circ \gamma)$ as well.

In addition, the requirement $\mathrm{Voc}(\psi) \subseteq \mathrm{Voc}(\varphi)$ from Theorem 2 now hinges on having $\mathrm{Voc}(\varphi * \gamma) \subseteq \mathrm{Voc}(\varphi) \cup \mathrm{Voc}(\gamma)$, but this is a mild and natural assumption about '$*$'.

7 Conclusions and future work

We considered dialogues as a preparatory phase for belief revision and we presented a dialog protocol for resolving conflicts resulting from unbelievable assertions. Depending on the result of the dialogue, the agents either have found out that their convictions are in conflict with each other, or they have found a formula that neither of them finds unbelievable. During the dialogue, the unbelievable assertions do not cause the rebutting agent to change his epistemic state, but the asserting agent might (or might not) change her epistemic state due to the rebuttals. In general, we allow the agents to change their epistemic states during the dialog.

Our protocol can terminate successfully even when there is a conflict, since it may never surface during the protocol. Suppose for instance that A asserts some $a \vee b$, where A can believe the first disjunct but not the second disjunct, and vice versa for B. We leave such pseudoagreements to further study, since avoiding them would require continuing the conversation further even after finding this first mutually believable formula.

The work can be extended to several directions, such as adding the possibility to extend the topic with new literals, adding the possibility to agree to restrict the topic, adding new utterance types to the agents (for instance, for making the protocol symmetric), and considering more expressive languages as is done e.g. in cooperative query answering (Gaasterland et al., 1992).

References

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

Meghyn Bienvenu, Andreas Herzig, and Guilin Qi. Prime implicate-based belief revision operators. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikos Avouris, editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008)*, pages 741–742, 2008.

Guilherme Bittencourt, Laurent Perrussel, and Jerusa Marchi. A syntactical approach to revision. In Ramon Lopez de Mántaras and Lorenza Saitta, editors, *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, pages 788–792. IOS Press, 2004.

Richard Booth. Social contraction and belief negotiation. *Information Fusion*, 7:19–34, 2006.

Mukesh Dalal. Investigations into a theory of knowledge base revision: Preliminary report. In Paul Rosenbloom and Peter Szlovits, editors, *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-88)*, volume 2, pages 475–479. AAAI Press, 1988.

Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89(1-2):1–29, 1997.

Satu Eloranta, Raul Hakli, Olli Niinivaara, and Matti Nykänen. Accommodative belief revision. In Stefan Hölldobler, Carsten Cutz, and Heinrich Wansing, editors, *11th European Conference on Logics in Artificial Intelligence (JELIA 2008)*, number 5293 in Lecture Notes in Artificial Intelligence (LNAI), pages 180–191. Springer, 2008.

Terry Gaasterland, Parke Godfrey, and Jack Minker. An overview of cooperative answering. *Journal of Intelligent Information Systems*, 1:123–157, 1992.

Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1989.

Sven Ove Hansson. Belief contraction without recovery. *Studia Logica*, 50:251–260, 1991.

Sven Ove Hansson. A survey of non-prioritized belief revision. *Erkenntnis*, 50(2-3):413–427, 1999.

Jaakko Hintikka. *Knowledge and Belief.* Cornell University Press, 1962.

Jaakko Hintikka and Ilpo Halonen. Interpolation as explanation. *Philosophy of Science*, 66 (Proceedings):S414–S423, 1999.

Yi Jin, Michael Thielscher, and Dongmo Zhang. Mutual belief revision: semantics and computation. In *Proceedings of the 22nd national conference on Artificial Intelligence (AAAI'07)*, pages 440–445. AAAI Press, 2007. ISBN 978-1-57735-323-2.

Kourosias, G. and Makinson, D. (2007). Parallel interpolation, splitting, and relevance in belief change. *The Journal of Symbolic Logic*, 72(3):994–1002.

Jérôme Lang, Paolo Liberatore, and Pierre Marquis. Propositional independence: Formula-variable independence and forgetting. *Journal of Artificial Intelligence Research*, 18:391–443, 2003.

David Makinson. Propositional relevance through letter-sharing. *Journal of Applied Logic*, pages 377–387, 2009.

Sara Negri and Jan von Plato. *Structural Proof Theory*. Cambridge University Press, 2001.

Parikh, R. (1999). Beliefs, belief revision, and splitting languages. In Moss, L., Ginzburg, J., and de Rijke, M., editors, *Logic, Language, and Computation*, volume 2, pages 266–278. CSLI Publications.

Simon Parsons, Michael Wooldridge, and Leila Amgoud. Properties and complexity of some formal inter-agent dialogues. *Journal of Logic and Computation*, 13:347–376, 2003.

Raymond Reiter. On integrity constraints. In *2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 97–111. Morgan Kaufmann, 1988.

Dan Sperber and Deirdre Wilson. Relevance theory. In *The Handbook of Pragmatics*, pages 607–632. Blackwell, Oxford, 2004.

Anne S. Troelstra and Helmut Schwichtenberg. *Basic Proof Theory*. Cambridge University Press, 2nd edition, 2000.

Jelle van Veenen and Henry Prakken. A protocol for arguing about rejections in negotiation. In S. Parsons, N. Maudet, P. Moraitis, and I. Rahwan, editors, *Second International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2005)*, number 4049 in Lecture Notes in Artificial Intelligence (LNAI), pages 138–153. Springer, 2006.

Douglas N. Walton and Erik C. W. Krabbe. *Commitment in Dialogue: Basic Commitments in Interpersonal Dialogue*. SUNY series in logic and language. State University of New York Press, 1995.