Annals of Mathematics and Artificial Intelligence manuscript No.
(will be inserted by the editor)

How to Deal with Unbelievable Assertions

Matti Nykänen · Raul Hakli · Satu Eloranta · Olli Niinivaara

the date of receipt and acceptance should be inserted later

Abstract We tackle the problem that arises when an agent receives unbelievable information. Information is unbelievable if it conflicts with the agent's convictions, that is, what the agent considers knowledge. We propose two solutions based on modifying the information so that it is no longer unbelievable. In one solution, the source and the receiver of the information cooperatively resolve the conflict. For this purpose we introduce a dialogue protocol in which the receiver explains what is wrong with the information by using logical interpolation, and the source produces a new assertion accordingly. If such cooperation is not possible, we propose an alternative solution in which the receiver revises the new piece of information by its own convictions to make it acceptable.

Keywords belief revision · argumentation · convictions · dialogue protocols

Mathematics Subject Classification (2010) 03B41 · 68T42 · 03C40

1 Introduction

When two agents are engaged in a dialogue, one of them may at some point assert something that the other is not willing to believe. Witness the following dialogue (Hansson 1991):

Example 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.

Matti Nykänen
School of Computing, University of Eastern Finland, P.O. Box 1627, FI-70211 Kuopio, Finland. E-mail: matti.nykanen@uef.fi

Raul Hakli
Department of Culture and Society, Section of Philosophy and History of Ideas, Aarhus University, Jens Chr. Skous Vej 7, DK-8000 Aarhus C, Denmark. E-mail: raul.hakli@cas.au.dk

Satu Eloranta and Olli Niinivaara
Department of Computer Science, P.O. Box 68 (Gustaf Hällströmin katu 2B), FI-00014 University of Helsinki, Finland. E-mail: firstname.lastname@cs.helsinki.fi

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 such situations: An assertion is made which contradicts something the receiver is not willing to give up during the dialogue. In the example, Bob is not willing to give up his ornithological knowledge just because Amy claims to have evidence contradictory to it, and Amy might be clinging to seeing *some* bird, no matter what Bob may assert.

We call the subset of beliefs that an agent is not willing to give up *convictions*. 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 non-prioritized belief revision, *core beliefs* are immune to revision (for a survey, see Hansson 1999). 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 he receives an *unbelievable* assertion, that is, something that conflicts with his convictions. A trivial solution would be to reject the unbelievable input, but we propose two other solutions. In the first solution, called *accommodative belief revision* (Eloranta et al. 2008), the receiver fixes the input by revising it by his convictions. In that way he tries to guess what the asserter would have asserted, "had she known better". In our second, cooperative solution, the receiver informs the asserter about the mismatch and lets the asserter herself tell what she would have asserted, "had she known better".

In our cooperative solution, we let the agent not only to reject an unbelievable assertion, but also to make an irretractable assertion to the contrary. Such an irretractable assertion acts both as a signal that challenging the rejection any further would be futile, and also as an explanation why the assertion conflicted with the convictions of the agent. We show how the receiver can formulate this irretractable assertion so that it stays on the topic of their dialogue and provides relevant information about the conflict. This is accomplished by using logical *interpolation*, which gives a formula that is entailed by the convictions of the agent and entails the negation of the unbelievable assertion. When the asserter receives an irretractable assertion to the contrary, say φ, she then produces her new assertion, say ψ, based on some conditional belief "if I were to believe that φ, I would believe that ψ" she has. We call this type of conditionals *doxastic conditionals*. For the cooperative solution we give a general subdialogue protocol that can be used as part of any type of dialogue and is guaranteed to solve any conflict situation in a limited number of dialogue moves.

The outline of the paper is as follows. In section 2, we will introduce our notations and present some preliminaries concerning belief revision theory in the presence of convictions. In section 3, we examine two ways how the receiving agent can react to unbelievable assertions: In section 3.1 we propose how it can nevertheless revise its beliefs by renewing the assertion to be consistent with its convictions, while in section 3.2 we assume instead that the two agents can communicate, which enables the receiving agent to explain this inconsistency to the asserting agent who can then reconstruct the assertion. In section 4, we will propose guidelines for generating new assertions in response to such explanations and present our dialogue protocol. In section 5, we will show that the dialogues will always lead to a rational outcome: If the agents end up in a disagreement, then they have mutually exclusive convictions. Otherwise, they will have found something that both agents can revise their beliefs with. The extension is in turn in section 6, where we will study how the

asserting agent could restrict the original topic of the discussion in order to continue the dialogue even if some conflicts between the agents' convictions are found. In section 7, we will discuss some related work. In section 8, we will give conclusions.

The dialogue protocol presented in section 4.3 expands an earlier one developed by Nykänen et al. (2011). That preliminary version formalized neither these guidelines proposed in section 4.2 nor the subsequent analysis and extension of the protocol in sections 5 and 6, respectively.

2 Agents with convictions

Let us start by introducing the basic setup under consideration in section 2.1. In section 2.2 we will recall some rationality criteria for belief revision. We discuss doxastic conditionals in section 2.3. Then in section 2.4 we will explain what we mean by agents having convictions in addition to ordinary beliefs. In section 2.5 we will discuss the effect of convictions on belief revision, and in section 2.6 we will look at rationality criteria for changing convictions.

2.1 Epistemic states of agents

We assume that agents have *epistemic states*. The epistemic state of an agent determines (among other things) his *belief set* consisting of all the beliefs the agent currently holds.

We will denote the belief set of an epistemic state $\mathscr{S}$ by $\mathbb{B}(\mathscr{S})$. Beliefs are expressed in language $\mathscr{L}$ consisting of formulas of classical propositional logic, that is, we shall not deal here with quantifiers nor modalities (like beliefs about each other's beliefs). We assume that the agents are competent in the sense that they believe all the formulas entailed by their beliefs.

We do not assume any particular structure for epistemic states. However, Example 2 in section 2.4 provides one possibility. Instead, we treat epistemic states $\mathscr{S}$ as an abstract data type which must provide two operations: First, we must be able to access the information contained in $\mathscr{S}$, albeit not directly but through inference by asking whether or not $\mathbb{B}(\mathscr{S}) \models \varphi$ holds for a given formula φ. This sidesteps for instance the issue whether $\mathscr{S}$ contains finitely or infinitely many propositional symbols, since φ contains only finitely many symbols. And second, we must be able to construct a new state by revising an existing state $\mathscr{S}$ with a given formula φ. The rest of this section discusses this revision operation further.

2.2 Belief revision

Let us recall the theory of belief change, in which the agent evolves his epistemic state due to incoming information called epistemic input (Gärdenfors 1988). When epistemic states are changed, at first the input is classified. The input may contain new information about a static world or a contraction of some belief. It may also record a change in the world, in which case the operation is called an update. The way the epistemic state is changed depends on the result of the classification. On the meta level, the change is guarded by rationality criteria, such as the AGM-postulates (Alchourrón et al. 1985) for belief revision and the four additional postulates by Darwiche and Pearl (1997) for iterated belief revision.

Using $\mathscr{S}$ to denote an epistemic state, φ and ψ to denote epistemic inputs, and $\circ$ to denote a belief-revision operator, the AGM-postulates and the postulates for iterated belief revision can be rephrased as follows:

$$\mathbb{B}(\mathscr{S} \circ \varphi) \models \varphi. \tag{R1}$$
$$\text{If } \mathbb{B}(\mathscr{S}) \not\models \neg\varphi \text{ then } \mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}(\mathscr{S}) \cup \{\varphi\}. \tag{R2}$$
$$\text{If } \varphi \not\models \bot \text{ then } \mathbb{B}(\mathscr{S} \circ \varphi) \not\models \bot. \tag{R3}$$
$$\text{If } \varphi \equiv \varphi', \text{ then } \mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}(\mathscr{S} \circ \varphi'). \tag{R4}$$
$$\mathbb{B}(\mathscr{S} \circ \varphi) \cup \{\psi\} \models \mathbb{B}(\mathscr{S} \circ (\varphi \wedge \psi)). \tag{R5}$$
$$\text{If } \mathbb{B}(\mathscr{S} \circ \varphi) \not\models \neg\psi, \text{ then } \mathbb{B}(\mathscr{S} \circ (\varphi \wedge \psi)) \models \mathbb{B}(\mathscr{S} \circ \varphi) \cup \{\psi\}. \tag{R6}$$
$$\text{If } \varphi \models \psi \text{ then } \mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}((\mathscr{S} \circ \psi) \circ \varphi). \tag{IR1}$$
$$\text{If } \varphi \models \neg\psi \text{ then } \mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}((\mathscr{S} \circ \psi) \circ \varphi). \tag{IR2}$$
$$\text{If } \mathbb{B}(\mathscr{S} \circ \psi) \models \varphi, \text{ then } \mathbb{B}((\mathscr{S} \circ \varphi) \circ \psi) \models \varphi. \tag{IR3}$$
$$\text{If } \mathbb{B}(\mathscr{S} \circ \psi) \not\models \neg\varphi, \text{ then } \mathbb{B}((\mathscr{S} \circ \varphi) \circ \psi) \not\models \neg\varphi. \tag{IR4}$$

Postulate (R1) says that the new piece of information is accepted, that is, the insertion of the new belief into the belief set of the agent 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 irrelevance of syntax.

According to Alchourrón et al. (1985), an operator may be called a revision operator if it satisfies postulates (R1)–(R4); postulates (R5) and (R6) are considered supplementary. Postulates (R5) and (R6) may be thought to guard iterated change. Together they say that if learning φ does not contradict ψ, then learning first φ and then ψ gives the same belief set than learning $\varphi \wedge \psi$ in the first place.

According to postulate (IR1), if we obtain two pieces of information with the latter being more accurate, the resulting belief set should be the same as if we had learned only the latter piece of information. If we receive two opposite pieces of information, then according to postulate (IR2), the resulting belief set should again be the same as having received only the latter piece of information. According to postulate (IR3), believing φ should not be prevented by learning φ, if φ were otherwise believed, and (IR4) says that an insertion should not cause its own negation to be believed (Darwiche and Pearl 1997). Postulates (R5) and (R6) may be thought as special cases of postulate (IR1).

As a whole, the main principles in the rationality criteria are minimality of change and maintaining consistency of the beliefs.

2.3 Doxastic conditionals

If epistemic states $\mathscr{S}$ were determined by their belief sets $\mathbb{B}(\mathscr{S})$, the combined set of postulates would result in triviality of logic (Gärdenfors 1988, chapter 7), that is, no three satisfiable but pairwise contradictive formulas could exist. Hence epistemic states $\mathscr{S}$ must have some richer structure. If a revision operator is to satisfy both sets of postulates, (R1)–(R6) and (IR1)–(IR4), epistemic states $\mathscr{S}$ must contain doxastic conditionals (Eloranta 2013, Chapter 7). The semantics of doxastic conditionals can be defined by applying the Ramsey

test (Lewis 1973): a doxastic conditional "if φ then ψ" is true in an epistemic state $\mathscr{S}$, if revising $\mathscr{S}$ by φ results in an epistemic state in which ψ is believed, that is, if $\mathbb{B}(\mathscr{S} \circ \varphi) \models \psi$.

The AGM-postulates guard changing the set $\mathbb{B}(\mathscr{S} \circ \varphi)$, whereas the postulates (IR1)–(IR4) guard changing doxastic conditionals. Example 2 in the next section will give a concrete construction for epistemic states containing conditionals.

2.4 Epistemic states with convictions

On one hand, the agent may be willing to give up some of his beliefs given new evidence to the contrary. On the other hand, he may regard some of his beliefs as convictions which he will hold on to, regardless of any new evidence. We therefore assume that each epistemic state contains also a *conviction set*, which consists of all the convictions the agent currently holds. The conviction set of the state $\mathscr{S}$ is denoted with $\mathbb{C}(\mathscr{S})$. As in the case of beliefs, we assume that the agent takes as convictions all the formulas entailed by his convictions. Thus $\mathbb{B}(\mathscr{S}) = \{\varphi \in \mathscr{L} \mid \mathbb{B}(\mathscr{S}) \models \varphi\}$ and $\mathbb{C}(\mathscr{S}) = \{\varphi \in \mathscr{L} \mid \mathbb{C}(\mathscr{S}) \models \varphi\}$. We take these assumptions as static rationality criteria for epistemic states.

We will make two further assumptions (Eloranta 2013, Chapter 6) that can be taken as additional static rationality criteria for epistemic states. As we draw parallels between convictions and knowledge, we will require that what an agent is convinced of, it also believes (Hintikka 1962, Chapter 3), that is,

$$\mathbb{B}(\mathscr{S}) \models \mathbb{C}(\mathscr{S}). \tag{S1}$$

We will also require that belief sets are non-contradictory, that is,

$$\mathbb{B}(\mathscr{S}) \not\models \bot. \tag{S2}$$

We require that the epistemic states produced using any change operator satisfy these static rationality criteria.

We do wish to point out that we do not take the sets $\mathbb{B}(\mathscr{S})$ and $\mathbb{C}(\mathscr{S})$ to constitute the epistemic state $\mathscr{S}$. In addition to those, agents are assumed to have doxastic conditionals from section 2.3. Hence we assume some representation for $\mathscr{S}$ which includes all these three components. Although we do not fix any particular such representation, here is one possibility:

Example 2 An epistemic state can be represented by an *epistemic base* (Eloranta 2013, Chapter 6), which is a linearly ordered structure $\langle T, R\rangle$, where $T \subseteq \mathscr{L}$ is a finite nonempty set of satisfiable propositional formulas that are pairwise inconsistent with each other, and $R \subseteq T \times T$ is a linear ordering on T. Because T is finite, any of its subsets has a minimal element in any linear ordering on T.

The intuition behind epistemic bases goes like this. The ordering R is an *ordering of disbelief* on the formulas in T (it compares to the ordering of disbelief on the possible worlds (Spohn 1988) that are models of the formulas). The minimal element in the ordering is the most plausible one of all the formulas in T, and so on. A propositional formula φ is then believed in the epistemic state, if it is entailed by the most plausible formula in T, and it is known, or more precisely, taken as a conviction in the epistemic state, if it is entailed by all the formulas in T. Thus the minimal element in T represents the belief set, whereas the disjunction of all the elements in T represents the conviction set, that is, $\mathbb{B}(\mathscr{S}) \equiv \min(T, R)$ and $\mathbb{C}(\mathscr{S}) \equiv \bigvee T$. Then by definition, every epistemic base satisfies the static rationality criteria (S1) and (S2).

Each formula in $t \in T$ has a rank $\text{ord}(t) = |\{t' \in T : (t', t) \in R\}|$. If $|T| = n$, we may refer to the epistemic state by using list formulation $[t_0, t_1, \ldots, t_{n-1}]$, where each t_i, $0 \leq i < n$, is a formula $t \in T$ such that $\text{ord}(t) = i$. At the top of the list is the most plausible formula t_0.

Given a formula φ such that $\bigvee_{i=0}^{n-1} t_i \not\models \neg\varphi$, there is a minimal rank m such that $t_m \not\models \neg\varphi$ and $t_i \models \neg\varphi$ for all i, $0 \leq i < m$. Given a revision operator $\circ$ that satisfies the postulates for belief revision, then $\mathbb{B}(\mathscr{S} \circ \varphi) \equiv t_m \wedge \varphi$ and the doxastic conditional "if φ then ψ" is evaluated true in the state, if $t_m \wedge \varphi \models \psi$.

2.5 Changing beliefs in the presence of convictions

The main principles in the rationality criteria for belief change are minimality of change and maintaining consistency of the beliefs. Yet according to postulate (R1), even contradictions are accepted into beliefs. We wish to consider belief change, in which contradictions are not accepted, and in which convictions are never given up.

In belief revision convictions can be treated as integrity constraints: They express those properties that should always hold. When defining the effect of integrity constraints on belief change, Katsuno and Mendelzon (1992) have required that the result entails the integrity constraints. Using $\mathscr{S}$ to denote an epistemic state, φ to denote an epistemic input, $\circ$ to denote a belief-revision operator, and *IC* to denote a propositional formula expressing the integrity constraints, the definition by Katsuno and Mendelzon (1992) concerning the effect of integrity constraints on belief revision is rephrased as

$$\mathscr{S} \circ^{IC} \varphi =_{def} \mathscr{S} \circ (\varphi \wedge IC). \tag{1}$$

By Equation (1), convictions affect belief revision and thus doxastic conditionals.

Now that we have convictions in epistemic states, some adjustments are needed in the postulates for belief revision, if we want the epistemic states to satisfy the static rationality criteria. Let $\mathscr{S}$ denote an epistemic state, φ and ψ denote propositional formulas, and let $\circ$ denote a belief-revision operator that is a function from propositional formulas and epistemic states satisfying the static rationality criteria into epistemic states satisfying the static rationality criteria. In the following set of postulates for belief revision (Eloranta 2013, Chapter 5) unbelievable input is rejected:

$$\mathbb{C}(\mathscr{S}) \equiv \mathbb{C}(\mathscr{S} \circ \varphi). \tag{C0}$$

If $\mathbb{C}(\mathscr{S}) \models \neg\varphi$ then $\mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}(\mathscr{S})$. (CR0)

If $\mathbb{C}(\mathscr{S}) \not\models \neg\varphi$ then $\mathbb{B}(\mathscr{S} \circ \varphi) \models \varphi$. (CR1)

If $\mathbb{B}(\mathscr{S}) \not\models \neg\varphi$, then $\mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}(\mathscr{S}) \cup \{\varphi\}$. (CR2)

If $\varphi \models \psi$ and $\mathbb{C}(\mathscr{S}) \not\models \neg\varphi$ then $\mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}((\mathscr{S} \circ \psi) \circ \varphi)$. (CIR1)

If $\varphi \models \neg\psi$ and $\mathbb{C}(\mathscr{S}) \not\models \neg\varphi$ then $\mathbb{B}(\mathscr{S} \circ \varphi) \equiv \mathbb{B}((\mathscr{S} \circ \psi) \circ \varphi)$. (CIR2)

If $\mathbb{B}(\mathscr{S} \circ \psi) \models \varphi$, then $\mathbb{B}((\mathscr{S} \circ \varphi) \circ \psi) \models \varphi$. (CIR3)

If $\mathbb{B}(\mathscr{S} \circ \psi) \not\models \neg\varphi$, then $\mathbb{B}((\mathscr{S} \circ \varphi) \circ \psi) \not\models \neg\varphi$. (CIR4)

To ensure that belief revision does not change convictions, the postulate (C0) has been added. It guarantees that convictions are never given up in revision. Postulates (R1), (IR1), and (IR2) have been made conditional in the corresponding postulates (CR0), (CR1), (CIR1), and (CIR2). Postulates (CR2), (CIR3), and (CIR4) are identical to the corresponding postulates (R2), (IR3), and (IR3).

The original AGM-postulates guide prioritized belief revision, because their success postulate unconditionally states that $\varphi \in \mathbb{B}(\mathscr{S} \circ \varphi)$. The postulates presented here guide non-prioritized belief revision, due to the conditions added to the success postulate (R1).

Postulate (CR0) requires that any attempt to revise beliefs with unbelievable input must not change beliefs. The purpose of this paper is to circumvent such futile attempts in belief-change operations. This is done by taking actions to avoid such revisions altogether via techniques discussed in section 3.

2.6 Changing convictions

We take that convictions can only increase monotonically during the dialogue. Let $\mathscr{S}$ denote an epistemic state, φ and ψ denote propositional formulas that are consistent with $\mathbb{C}(\mathscr{S})$, and let $\oplus$ denote a knowledge-expansion operator that is a function from propositional formulas and epistemic states satisfying the static rationality criteria into epistemic states satisfying the static rationality criteria. Rationality criteria for knowledge expansion (Eloranta 2013, Chapter 5) could then be rephrased as follows:

$$\text{If } \mathbb{C}(\mathscr{S}) \not\models \neg\varphi \text{ then } \mathbb{C}(\mathscr{S} \oplus \varphi) \equiv \mathbb{C}(\mathscr{S}) \cup \{\varphi\}. \tag{C2}$$

$$\text{If } \mathbb{B}(\mathscr{S}) \not\models \neg\varphi \text{ then } \mathbb{B}(\mathscr{S} \oplus \varphi) \equiv \mathbb{B}(\mathscr{S}) \cup \{\varphi\}. \tag{CB2}$$

$$\text{If } \varphi \models \psi \text{ then } \mathbb{B}(\mathscr{S} \oplus \varphi) \equiv \mathbb{B}((\mathscr{S} \oplus \psi) \oplus \varphi). \tag{CIB1}$$

Note that we have assumed that here we only consider input that is consistent with the old convictions. Postulate (C2) is analogous to postulate (R2).

Postulate (C2) together with the static rationality criteria suggests that beliefs may have to be changed when the set of convictions increases. Therefore any knowledge-expansion operator must also satisfy the postulates for belief revision. It is, however, sufficient to have only postulates (CB2) and (CIB1) which are identical to postulates (CR2) and (CIR1). The other postulates are either not applicable or redundant.

3 Two techniques for avoiding unbelievable input

This section proposes techniques for dealing with an unbelievable input. The first solution in section 3.1 is accommodative belief revision (Eloranta et al. 2008), in which the input is first revised to avoid the problem in belief revision. The second solution in section 3.2 is to reject the input but to give the asserter an opportunity to correct the input. In this solution we propose logical interpolation for producing a reason for the rejection (Nykänen et al. 2011).

3.1 Revising unbelievable assertions

Agents are not always able to have a dialogue between them. That would be the case, if instead of the dialogue in Example 1, Bob happened to read a magazine interview of Amy telling how she saw the three-toed woodpecker. Even when communication is not possible, the receiver can still try to extract some information from the input by revising it to fit the convictions as done in accommodative belief revision (Eloranta et al. 2008).

In accommodative belief revision, the input is modified before belief revision takes place. Assuming that the conviction set of an agent can be represented by a propositional

formula, accommodative belief revision extends Equation (1) about the effect of integrity constraints on belief change as follows. Let $\mathscr{S}$ denote an epistemic state, let κ denote a propositional formula representing the set $\mathbb{C}(\mathscr{S})$, and let φ denote a propositional input formula. The accommodative revision of the state $\mathscr{S}$ by the formula φ is defined as

$$\mathscr{S} \otimes \varphi =_{def} \mathscr{S} \circ (\varphi * \kappa), \tag{2}$$

in which $\circ$ denotes a belief-revision operator on epistemic states and $*$ denotes a belief-revision operator on propositional formulas. Note that the definition does not give a single operator but a scheme in which various operators can be applied.

The justification of the method is the following: The modified input is considered as an estimate of the formula that the source of the input would have given, had it had all the knowledge that the receiver has.

If one assumes that the convictions $\mathbb{C}(\mathscr{S})$ in an epistemic state $\mathscr{S}$ have been obtained with a finite sequence of monotonic increase operations starting from a state without any other convictions except tautologies, then by postulate (C2) the resulting convictions can be represented by a propositional formula κ (Eloranta 2013, Chapter 7).

Accommodative revision is restricted only to those epistemic states in which the static constraints (S1) and (S2) are satisfied. Thus the agent believes what it knows and it does not believe contradictions. We can see that accommodative revision fails to satisfy the basic AGM-postulates only when the input is contradictory to the convictions in the state: In those cases the success postulate fails. Thus accommodative revision is a form of non-prioritized belief revision. Accommodative revision preserves the static constraints and is able to guarantee non-contradictory belief sets at all occasions (Eloranta et al. 2008).

3.2 Rejecting unbelievable assertions

Consider now the case where the agents are able to communicate with each other. Then the receiver can inform the asserter that the latest assertion is in conflict with its convictions and hence unbelievable. This allows the asserter in turn to come up with an alternative assertion which avoids this conflict. This task becomes easier, if the receiver not only informs the asserter that there is such a conflict but also explains how it arose. The asserter must in turn come up with an alternative assertion by considering what it would believe instead after learning this explanation. In this way they enter a dialogue where the receiver explains its convictions while the asserter asserts what these explanations would entail in light of its own doxastic conditionals.

Recall the following notion:

Definition 1 (Craig interpolant) Let α and β be two formulas such that $\alpha \models \beta$. An *interpolant* for this entailment is any formula θ such that

(i) $\alpha \models \theta$
(ii) $\mathrm{voc}(\theta) \subseteq \mathrm{voc}(\alpha) \cap \mathrm{voc}(\beta)$, where $\mathrm{voc}(\phi)$ denotes the set of propositional symbols appearing in the formula ϕ, and
(iii) $\theta \models \beta$.

Such formulas always exist in classical propositional logic, by Craig's interpolation theorem (see for instance Boolos and Jeffrey 1989, Part I of Chapter 23). This interpolant can be seen as one kind of *explanation how* or *why* α entails β (Hintikka and Halonen 1999).

Moreover, property (ii) means that this explanation θ is expressed in the common vocabulary of α and β.

In our proposal, the receiver uses interpolation to calculate an explanation for a rejection in a situation where it is given input φ, but its convictions entail $\neg\varphi$. When we apply Definition 1 to the dialogue in Example 1, α is (some part of) Bob's convictions, β is the negation $\neg\varphi$ of Amy's assertion φ, and the interpolant θ explains why α rules out φ. Property (ii) of Definition 1 is now important, because Bob must express his explanation θ in a vocabulary familiar to Amy in order for her to understand it, and the most certain way to guarantee this is to use the same vocabulary as she did in her own assertion φ.

Our general setting is thus as follows: One agent A (here Amy) asserts φ to another agent B (here Bob). The receiving agent B must check whether this assertion φ is consistent with its own convictions $\mathbb{C}(\mathscr{B})$ or not, where $\mathscr{B}$ denotes its epistemic state. If they are not consistent with each other, then agent B must explain this inconsistency to agent A with a corresponding interpolant θ. Otherwise φ is something which agent B can believe.

In section 3.1 the receiving agent B could not communicate with the asserting agent A. Thus agent B had to modify alone the assertion φ it received from agent A to be consistent with its own convictions $\mathbb{C}(\mathscr{B})$. In contrast, communication is possible here, and agent B can therefore shift the task of modifying φ back to its sender, agent A. However, agent B can aid agent A in this task by giving also an explanation why φ was not consistent with $\mathbb{C}(\mathscr{B})$.

Example 3 Consider the dialogue in Example 1. Let the propositional symbol 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", and s stand for "Amy saw a lark". The subdialogue starts when Amy asserts $\varphi = p \wedge q \wedge r$. Let Bob's convictions $\mathbb{C}(\mathscr{B})$ contain a formula $\alpha = (p \vee s) \rightarrow (\neg q \wedge \neg r)$. Then this α entails an interpolant $\theta = p \rightarrow \neg q \wedge \neg r$, which again entails the negation of Amy's assertion, $\beta = \neg(p \wedge q \wedge r)$.

It is natural to view the dialogue in Example 3 as Bob trying to determine whether his convictions $\mathbb{C}(\mathscr{B})$ are consistent with Amy's assertion φ or not. If they are, then he could accept φ. But if they are not, then his reply should explain this inconsistency to her. Definition 1 extends to this setting by taking $\mathbb{C}(\mathscr{B})$ as α and $\neg\varphi$ as β. Hence his reply should be such an interpolant θ for the *refutation proof* R which shows their inconsistency. Recall that such a refutation R is finite even though $\mathbb{C}(\mathscr{B})$ need not be; we only need to assume that the receiver can produce a finite proof that φ is not consistent with its possibly infinite convictions.

Hence we need some way of extracting such a Craig interpolant θ from a given refutation R. Moreover, extracting θ from R should be computationally easy. It is reasonable to expect that the receiver is willing to spend significant computational effort on checking whether it can accept φ or not (that is, on finding one R) because accepting φ will affect its beliefs. In contrast, it may not always be reasonable to expect that the receiver would be willing to spend much more extra effort on explaining to the asserter why it cannot accept φ (that is, on extracting θ from the R found).

We are not tied to any particular way of extracting an interpolant θ from a refutation R. However, the way developed by D'Silva et al. (2010) is particularly apt for our purposes for two reasons: First, their way extracts the interpolant by straightforward structural induction over the refutation R, and therefore the extraction is indeed computationally easy in the sense above. We relegate the details to Appendix A. And second, their way has a labelling parameter which allows us to guide what kind of interpolant it extracts. For instance, the *strong* labelling given as Equation (4) in Appendix A guides it to extract a logically strong

interpolant, whereas the *small* labelling given as Equation (6) in Appendix A guides it to extract an interpolant with few propositional symbols. These labellings allow fine-tuning interpolants according to their current use.

Example 4 If Bob used the labelling mentioned above for extracting interpolants, then he would in fact come up with a terser refutation than the interpolant $\theta = p \rightarrow \neg q \wedge \neg r$ in Example 3: the interpolant would be either $\neg p \vee \neg q$ or $\neg p \vee \neg r$, depending on which of the possible refutations R he discovers. In contrast, coming up with θ would require him to consider more than one such refutation. His interpolant θ namely contains two ways how Amy's assertion conflicts with his convictions: both $p \rightarrow \neg q$ and $p \rightarrow \neg r$.

We shall use also the following stronger notion of interpolation later:

Definition 2 (uniform interpolant) Let U be a finite set of propositional symbols, and let Φ be a propositional formula or theory. Then the *uniform interpolant* of Φ with respect to U is a propositional formula denoted as $\pi_U \Phi$ which suffices as a Craig interpolant for every entailment $\Phi \models \psi$ where $\mathrm{voc}(\Phi) \cap \mathrm{voc}(\psi) \subseteq U$.

Without loss of generality we assume that $\mathrm{voc}(\pi_U \Phi) = U$.

Classical propositional logic possesses these uniform interpolants $\pi_U \Phi$, and they are unique up to logical equivalence by definition. Computing them is reviewed briefly as Appendix B.

Definition 2 is a proof-theoretic formulation of uniform interpolation. An equivalent semantic reformulation is to consider the truth table of Φ as a relation in the database sense; then $\pi_U \Phi$ is its projection to the attributes U, which also explains the notation chosen. That is, $\pi_U \Phi$ has a model w if and only if Φ has a model w' which agrees with w for every propositional symbol $x \in U$. Hence $\pi_U \Phi$ represents the information in Φ about U, and we shall employ uniform interpolation for this purpose. This semantic formulation has also been called projection or marginalization (Kohlas et al. 1999) or variable forgetting (Lang et al. 2003) in the Artificial Intelligence literature, while the database setting was noted by Atserias et al. (2004, Section 3).

Example 5 By this semantic formulation of uniform interpolation, the uniform interpolant $\pi_{\{p,q,r\}} \alpha$ from Example 3 consists of those truth values for p, q and r for which there exists some truth value s such that the formula α is true.

If p is false, then any truth values for q and r suffice, because s can be false too. If p is true, then neither q nor r can be true regardless of s. Hence $\pi_{\{p,q,r\}} \alpha$ is $(\neg p) \vee (p \wedge \neg q \wedge \neg r)$, or any equivalent formula on p, q and r such as $p \rightarrow \neg q \wedge \neg r$. Thus Bob in fact replies with the uniform interpolant in Example 3.

4 Believability restoration protocol for resolving unbelievable assertions

We assume from now on that the agents can communicate with each other as in section 3.2, and propose our protocol for resolving unbelievable assertions. We first define basic concepts in section 4.1, then discuss some actions carried out in the protocol in section 4.2, and finally propose our dialogue protocol in section 4.3.

4.1 Elements of the dialogue

We will consider certain subdialogues, that is, parts of conversations between two agents. One of the agents is the asserter and the other is the receiver; to avoid always repeating these roles, we will refer to the former as 'she' and the latter as 'he', mirroring Amy and Bob in Example 1. The language of their subdialogue is classical propositional logic. A subdialogue begins when the asserter makes an assertion that the receiver cannot accept because it conflicts with his convictions. This *initial assertion* acts as an entry point to a subdialogue where the conflict is dealt with. Henceforth we drop the prefix 'sub-'.

We use the concept of topic to constrain the assertions in a dialogue. By a topic, we mean the atomic formulas in the initial assertion. The utterances in the dialogue remain relevant to the topic in the letter-sharing sense (Makinson 2009, Definition 1.1).

Definition 3 (topic) The *topic* of a dialogue with an initial assertion φ is the set $\mathrm{voc}(\varphi)$.

This definition guarantees that the protocol eventually halts and that the execution of the dialogue does not in itself produce new conflicts with respect to previously unmentioned propositional symbols. However, this strict definition of a topic prevents the asserter from telling about all possible changes in her beliefs. For instance, it may happen that after the protocol the asserter has changed her mind about something that she has previously asserted but is not included in the current topic. These kinds of situations can be handled in the context of the whole conversation. The problem can be solved either by making the asserter tell about the changes or by leaving it to the other agent to detect possible inconsistencies in the asserter's assertions as is done by Snaith and Reed (2012).

Even if the conviction sets of the agents conflict with each other, it is still possible to communicate assertions successfully, if the conviction sets do not disagree on the topic of the assertion.

Definition 4 Two formulas or theories α and β *disagree on a topic* T, if and only if there exists a formula δ such that $\mathrm{voc}(\delta) \subseteq T$ and $\alpha \models \delta$ and $\beta \models \neg\delta$. Such a formula δ is a *witness* of this disagreement between them.

Conversely, when two formulas or theories do *not* disagree on a topic, and one of them entails some formula δ about it, the other one cannot entail its negation $\neg\delta$. However, it can either entail δ as well or stay uncommitted by entailing neither δ nor $\neg\delta$.

Example 6 The two formulas $x \wedge y$ and $\neg y$ are inconsistent with each other as a whole, but they do not disagree on the topic $\{x\}$ since all the witnesses would need y.

Hence two mutually inconsistent formulas or theories do not disagree on a topic that does not permit expressing the corresponding witnesses.

4.2 Rejection calls for reconsideration

The asserter engages the receiver in a dialogue so that she can influence his beliefs about her chosen topic. Moreover, she wants his beliefs to resemble hers. However, his convictions might prevent him from adopting her beliefs outright.

The dialogue proceeds by the asserter asserting some candidate for such a possible belief and the receiver rejecting her assertion based on his own convictions, which in turn gives her more information about what she could assert next, and so on. A central observation behind

our approach is that the asserter can take this information from the receiver into account by revising her beliefs with it. We will make this observation precise in Theorem 1 below.

Since the asserter steers their dialogue with her assertions, let us define the requirements her assertions must meet:

Definition 5 Let T be some topic, and let γ be what the receiver has told the asserter about his own convictions $\mathbb{C}(\mathscr{B})$ regarding T so far during their dialogue. A formula ψ is a *candidate assertion*, if and only if it meets the following five requirements:

(i) It is satisfiable.
(ii) It is about the topic. Formally, $\mathrm{voc}(\psi) \subseteq T$.
(iii) It takes into account all the information in γ. Formally, $\psi \models \gamma$.
(iv) It takes into account all the information in the asserter's own convictions about the topic. Formally, $\psi \models \pi_T \mathbb{C}(\mathscr{A})$.
(v) The asserter considers it to be maximally plausible among all her choices. Formally, $\mathbb{B}(\mathscr{A} \circ_A (\psi \vee \psi')) \models \psi$ for every ψ' satisfying the first four requirements (i)–(iv), where $\circ_A$ denotes her belief revision operator.

Furthermore, such a candidate assertion ψ is also a *precise* candidate assertion, if and only if it meets also the following sixth requirement:

(vi) It is the logically strongest candidate assertion. Formally, $\psi \models \psi'$ for every candidate assertion ψ'.

The first two requirements (i) and (ii) are straightforward, given that the aim is to produce a belief ψ which can be adopted about the topic T.

The third requirement (iii) arises as follows: Suppose to the contrary that the asserter asserts instead some formula $\psi' \vee \psi''$ where $\gamma \models \neg\psi''$. If the receiver admits that this assertion is believable, then the asserter is aware that it must be due to ψ' but not ψ''. Hence including ψ'' in her candidate assertion would be redundant. This requirement drops such redundancies explicitly. (They could have been permitted at the expense of having to qualify each candidate assertion with the current γ implicitly.)

The fourth requirement (iv) arises in turn as follows: Suppose to the contrary that the asserter asserts instead some other formula $\psi' \vee \psi''$ where $\mathbb{C}(\mathscr{A}) \models \neg\psi''$. If the receiver admits that her assertion is believable, then he might be admitting with ψ'' but not ψ'. Hence the asserter is aware that including ψ'' in her assertion is undesirable, because it might hide a fundamental disagreement between their convictions. This requirement excludes such undesirable inclusions explicitly.

The fifth requirement (v) states that after γ has ruled out some otherwise candidate assertions by requirement (iii), the asserter will then choose her next candidate assertion to be as plausible as possible from among these remaining choices. In other words, she is prepared to sacrifice some plausibility in order to keep the dialogue going so that she could influence the receiver's beliefs, but not more than demanded by him. This is what it means for his beliefs to resemble hers given his own convictions.

However, the following example discusses why imprecise candidate assertions may not be enough in every situation.

Example 7 Consider the situation, where the asserter is deliberating between two candidate assertions ψ_1 and ψ_2 where the former is logically strictly stronger than the latter: $\psi_1 \models \psi_2$ but $\psi_2 \not\models \psi_1$. Which one should she assert to the receiver?

If she asserts ψ_2, then the following scenario could take place: The receiver accepts her assertion, because ψ_2 is consistent with his own convictions $\mathbb{C}(\mathscr{B})$. However, it may be the

case that $\mathbb{C}(\mathscr{B})$ is consistent with $\psi_2 \wedge \neg\psi_1$. Thus both agents are willing to believe ψ_2 even though ψ_1 is a witness that their convictions disagree on the topic of their conversation!

The asserter can avoid this particular scenario by asserting ψ_1 instead of ψ_2. Iterating this argument shows that she can avoid all such scenarios only by asserting the logically strongest candidate assertion; that is, by asserting the precise candidate assertion.

Example 7 above leads to two different kinds of agreements: If the receiver admits an imprecise candidate assertion ψ, it means that he can believe *this* particular ψ about the topic. This is already sufficient in many situations. However, the example above shows that his resulting beliefs can still disagree with the asserter's resulting beliefs about their *whole* topic T, and there are other situations where the asserter cannot tolerate such overall disagreements. For those situations we add also the optional sixth requirement (vi).

Let us now characterize these candidate assertions in terms of interpolants.

Theorem 1 *The candidate assertions ψ in Definition 5 are exactly the Craig interpolants of $\mathbb{B}(\mathscr{A} \circ_A \gamma) \models \gamma \wedge \pi_T\mathbb{C}(\mathscr{A})$.*

Proof For the forward direction, we assume that ψ is a Craig interpolant of $\mathbb{B}(\mathscr{A} \circ_A \gamma) \models \gamma \wedge \pi_T\mathbb{C}(\mathscr{A})$ and show that it is also a candidate assertion as follows: The consistency assumption above ensures the consistency of $\mathbb{B}(\mathscr{A} \circ_A \gamma)$ which in turn ensures requirement (i). Since $\mathrm{voc}(\gamma) \subseteq T$ then also $\mathrm{voc}(\gamma \wedge \pi_T\mathbb{C}(\mathscr{A})) \subseteq T$ which in turn ensures requirement (ii). By the assumption we have $\psi \models \gamma \wedge \pi_T\mathbb{C}(\mathscr{A})$ whence $\psi \models \gamma$ and $\psi \models \pi_T\mathbb{C}(\mathscr{A})$; the former is requirement (iii), while the latter is requirement (iv). Consider finally requirement (v) and assume to the contrary that $\mathbb{B}(\mathscr{A} \circ_A (\psi \vee \psi')) \not\models \psi$. It means that ψ' has a model w which is not a model of ψ but the asserter considers w to be at least as plausible as any model of ψ. In contrast, $\mathbb{B}(\mathscr{A} \circ_A \gamma) \models \psi$ means that the models of ψ contain all those models of γ which the asserter considers most plausible. Hence w is not a model of γ, and so ψ' is not a candidate assertion after all, by requirement (iii).

For the other direction, we assume that ψ is a candidate assertion and show that it is also a Craig interpolant of $\mathbb{B}(\mathscr{A} \circ_A \gamma) \models \gamma \wedge \pi_T\mathbb{C}(\mathscr{A})$ as follows: Requirement (ii) gives $\mathrm{voc}(\psi) \subseteq T$, while requirements (iii) and (iv) give $\psi \models \gamma \wedge \pi_T\mathbb{C}(\mathscr{A})$, so it only remains to show that $\mathbb{B}(\mathscr{A} \circ_A \gamma) \models \psi$. Let ψ' be the logically weakest possible choice $\gamma \wedge \pi_T\mathbb{C}(\mathscr{A})$ in requirement (v). Then $\mathbb{B}(\mathscr{A} \circ_A (\psi \vee \psi'))$ equals $\mathbb{B}(\mathscr{A} \circ_A (\gamma \wedge \pi_T\mathbb{C}(\mathscr{A})))$ which in turn equals $\mathbb{B}(\mathscr{A} \circ_A \gamma)$ as needed. □

Theorem 1 shows that candidate assertions exist if and only if the asserter's convictions $\mathbb{C}(\mathscr{A})$ are consistent with the information γ she has received from the receiver. Hence our dialogue protocol must ensure their consistency before making another candidate assertion. It must also handle somehow the case when γ has grown to be inconsistent with $\mathbb{C}(\mathscr{A})$.

Theorem 1 also verifies our central observation given above: The asserter can form her next candidate assertion by first revising her beliefs $\mathbb{B}(\mathscr{A})$ with the information γ she has received from the receiver about his own convictions $\mathbb{C}(\mathscr{B})$ regarding their topic T so far, and then taking into account requirements (iii) and (iv) from Definition 5. Moreover, she can form them using any method of constructing Craig interpolants. In addition, the asserter can also precompute her own convictions $\pi_T\mathbb{C}(\mathscr{A})$ about T beforehand, if they do not change during the dialogue.

Theorem 2 *The precise candidate assertion in Definition 5 is $\pi_T\mathbb{B}(\mathscr{A} \circ_A \gamma)$.*

Proof The claim follows directly by requirement (vi) of Definition 5, Definition 2 and Theorem 1. □

Theorem 2 shows how the asserter can construct the precise candidate assertion with uniform interpolation. However, to do so she must recompute this uniform interpolant as γ changes during the dialogue. Hence the added computational cost of making candidate assertions precise is going from Craig to uniform interpolation during a dialogue.

Definition 5 gave the requirements which the asserter's next assertion in the dialogue must meet. The dialogue itself unfolds as a sequence $\psi_0, \theta_0, \psi_1, \theta_1, \psi_2, \theta_2, \ldots$ where each ψ_i is a candidate assertion by the asserter and θ_i is its rebuttal by the receiver if his convictions preclude adopting it. That is, the asserter forms her next candidate assertion ψ_i so that it satisfies Definition 5 with respect to the topic T she has chosen for the dialogue and the conjunction $\gamma = \theta_0 \wedge \theta_1 \wedge \theta_2 \wedge \ldots \wedge \theta_{i-1}$ of all the previous rebuttals she has received so far. In turn, the receiver forms his next rebuttal θ_i as a Craig interpolant for the entailment $\mathbb{C}(\mathscr{B}) \models \neg\psi_i$ as discussed in section 3.2.

This sequence can end in two outcomes. The positive outcome is when $\mathbb{C}(\mathscr{B}) \not\models \neg\psi_i$, so the receiver has no rebuttal θ_i to make against ψ_i, which is therefore something that he can believe. The negative outcome is when γ becomes inconsistent with $\mathbb{C}(\mathscr{A})$, and hence she can no longer make another candidate assertion ψ_{i+1}. This possible failure is one reason why we are not developing a belief revision operation here. Instead, we are developing an interactive preliminary phase which tries to discover some ψ_i that the agents could believe.

The asserter does not actually believe the rebuttals θ_i the receiver tells her in this dialogue, but merely records them into γ. Hence this strategy could be called *cautious*, since she makes no commitments during the dialogue, but instead waits until she sees its outcome. A more *credulous* strategy would be for her to start believing his rebuttals as soon as she receives them. Then her epistemic state evolves during the dialogue as the sequence $\mathscr{A}_0, \mathscr{A}_1, \mathscr{A}_2, \ldots$ where each $\mathscr{A}_{i+1} = \mathscr{A}_i \circ_A \theta_i$ and she forms her next candidate assertion ψ_{i+1} so that it satisfies Definition 5 with respect to this new epistemic state instead of her original state at the beginning of the dialogue. A still more *trusting* strategy would be to adopt his rebuttals as her convictions instead of beliefs, since they are his convictions after all; that is, her epistemic state would evolve as $\mathscr{A}_{i+1} = \mathscr{A}_i \oplus \theta_i$. However, postulates (CIR1) and (CIB1) of iterated belief revision ensure that all these three strategies lead her to the same candidate assertions at each step. Hence we can concentrate on the cautious strategy without loss of generality. Recovery from a negative outcome would namely be more difficult in the other two strategies, since they have already changed her epistemic state.

4.3 A believability restoration protocol

Let us now put together the protocol for the dialogue between the asserter and the receiver which we sketched at the end of section 4.2. The assertion-rebuttal sequence arises as the co-operation of the asserter's Algorithm 1 and the receiver's Algorithm 2.

In the following, the asserter's part of the protocol, the agent has made some initial assertion ψ to another agent, the receiver, about some topic T, but the receiver has rejected it with an explanation θ why. The asserter conducts their dialogue as follows:

Algorithm 1: The asserter's part of the believability restoration protocol.

1 The asserter has **assert**ed the initial assertion ψ to the receiver who has **reject**ed it with θ;
2 $T = \mathrm{voc}(\psi)$;
3 $\gamma = \top$;
4 *disagreement* = FALSE;
5 **while** (**not** *disagreement*) **and** the receiver **reject**ed with θ
6 $\gamma = \gamma \wedge \theta$;
7 **if** $\mathbb{C}(\mathscr{A})$ is consistent with γ
8 **choose** some candidate assertion ψ according to Definition 5;
9 **assert** ψ to the receiver
10 **else** *disagreement* = TRUE.

In the following, the receiver's part of the protocol, the asserter has offered some candidate assertion ψ to the receiver. Then the receiver compares it against his own convictions as follows:

Algorithm 2: The receiver's part of the believability restoration protocol.

1 **if** $\mathbb{C}(\mathscr{B}) \models \neg\psi$
2 **choose** some Craig interpolant θ for this entailment;
3 **reject** with θ to the asserter
4 **else admit** that this ψ is believable.

The asserter conducts the dialogue by repeatedly offering another candidate assertion based on the conjunction γ of past rebuttals. This continues until an outcome is reached. It also signals as *disagreement* whether this outcome was negative or positive. Algorithm 2 in turn responds to this current candidate assertion ψ with either a rebuttal θ or an admission which results in a positive outcome. However, we treat the asserter's first candidate assertion separately from the rest of the dialogue for two reasons: it sets the topic for this dialogue, and it may come from an enclosing dialogue which contains this one as its subdialogue.

Note that these two algorithms are nondeterministic: in general, the asserter has several choices for her next candidate assertion ψ on line 8 of her Algorithm 1, and the receiver has several choices for its rebuttal θ on line 2 of his Algorithm 2. These choices arise from the various choices for the corresponding interpolant. Each agent makes these choices according to its own strategy, which we do not discuss in this paper. However, we do note that the interpolant construction algorithm described in Appendix A can incorporate some such strategies conveniently as its labelling functions, as already mentioned in Example 4. This strategic component is another reason why we are developing here only a preliminary phase for belief revision, because the outcome of this phase depends on the strategies of the agents. The following example illustrates such strategic choices.

Example 8 Algorithms 1 and 2 treat Example 3 as follows:

1. Amy's initial assertion ψ is $p \wedge q \wedge r$.
2. Bob's convictions $\mathbb{C}(\mathscr{B})$ entail its negation $\neg\psi$, so he rejects it. In this dialogue, he chooses strategically his rejection θ to be the uniform interpolant $p \rightarrow \neg q \wedge \neg r$ of this entailment, as in Example 5.

Algorithm 2 would also permit choosing either $p \to \neg q$ or $p \to \neg r$ as θ instead, as in Example 4. By preferring the uniform interpolant over either of these two Craig interpolants, Bob follows a strategy where he offers Amy as much information as he can.

3. This $\gamma = \top \wedge \theta$ is consistent with Amy's convictions, so she chooses her next candidate assertion ψ' via Algorithm 1 as some Craig interpolant for the entailment $\mathbb{B}(\mathscr{A} \circ_A \gamma) \models \gamma \wedge \pi_{\{p,q,r\}}\mathbb{C}(\mathscr{A})$.
   Now $\pi_{\{p,q,r\}}\mathbb{C}(\mathscr{A}) = (p \vee \neg p) \wedge q \wedge (r \vee \neg r)$ since she is convinced only of q, whereas p and r were merely her beliefs. Hence the entailed formula reduces to $\neg p \wedge q \wedge (r \vee \neg r)$. Thus her next candidate assertion ψ' admits that she did not see a three-toed woodpecker ($\neg p$) but maintains that she did see a bird with a red forehead (q).
   Whether or not she will claim anything about seeing a bird with a red rump (r vs. $\neg r$) depends on whether or not she believes something about it after revising her epistemic state with γ. Moreover, even if she does believe so, she can choose strategically whether or not to include that belief in her Craig interpolant for this entailment, because its entailed formula provides no information about r. Here it is natural to assume that she still continues to believe r after the revision since γ did not challenge it. Thus she would choose strategically her next candidate assertion ψ' to be either $\neg p \wedge q \wedge r$ or $\neg p \wedge q$, depending on whether her strategy is to discuss the red rump issue with Bob or not.
4. Bob can finally admit her ψ' because she no longer claims to have seen a three-toed woodpecker.

5 Properties of the believability restoration protocol

This section is devoted to discussing the properties of the protocol given in section 4.3. We will start by discussing some general principles of cooperative communication in section 5.1. In section 5.2, we will consider the termination of the protocol and the length of the dialogues. In section 5.3, we will study the alternative outcomes of the protocol.

5.1 Satisfying the cooperative principle

We aimed at designing a dialogue protocol in which the agents' assertions satisfy *the Cooperative Principle* presented by Grice (1975, 1978) to govern conversations between cooperative agents. *Grice's Maxims* elaborate the general principle into more specific conversational rules which the participants can be expected to observe:

> *Maxim of Quantity:* (a) Make your contribution as informative as required (for the current purposes of the exchange). (b) Do not make your contribution more informative than is required.
> *Maxim of Quality:* Try to make a contribution which is true. More specifically: (a) Do not say what you believe to be false. (b) Do not say that for which you lack adequate evidence.
> *Maxim of Relevance:* Be relevant.
> *Maxim of Manner:* Be clear.

These maxims rule out the naive extremities in dealing with unbelievable input, that is, either to terminate the dialogue or to reply with everything one knows about the topic.

Various attempts have been made to formalise these maxims (sometimes excluding the Maxim of Manner which seems more resistant to formalisation) either in probabilistic,

information-theoretic or game-theoretic terms (see, e.g., Jäger 2007; Frank and Goodman 2014). In our framework, we can start from Groenendijk (1999) who presents minimal logical conditions that must be satisfied for the maxims to hold. We can then consider whether the receiver follows these minimal conditions when he rejects the candidate assertion ψ with his θ.

According to Groenendijk (1999), Maxim of Quantity requires that the agents make only non-redundant statements, that is, statements that are not entailed by the preceding context. In our framework, this can be taken to mean that one should try to only assert what the other agent does not already believe. (Maxim of Quantity is highly dependent on the purpose of the conversation, and we will discuss more demanding interpretations in due course.) Groenendijk (1999) takes Maxim of Quality to require an agent's statements to be credible, minimally interpreted as requiring that she does not contradict herself. In our context, the receiver's assertions trivially satisfy this because he only asserts his convictions which are assumed to be consistent. We can instead focus on a more literal interpretation of part (a) which states that an agent should not assert something she believes to be false. In addition, Maxim of Relevance is understood by Groenendijk (1999) as a requirement that the statements of the receiver should exclusively address the issues raised by the asserter.

With these interpretations we can show the following:

Theorem 3 *Given that the asserter's candidate assertion ψ satisfies $\mathbb{B}(\mathscr{A}) \not\models \neg\psi$, the receiver's rebuttal θ satisfies*

$$\mathbb{B}(\mathscr{A}) \not\models \theta \qquad (M_{Quan})$$
$$\mathbb{B}(\mathscr{B}) \not\models \neg\theta \qquad (M_{Qual})$$
$$\mathrm{voc}(\theta) \subseteq \mathrm{voc}(\psi). \qquad (M_{Rel})$$

Proof Recall that θ is an interpolant for the entailment $\mathbb{C}(\mathscr{B}) \models \neg\psi$. To see claim ($M_{Quan}$), assume the opposite: $\mathbb{B}(\mathscr{A}) \models \theta$. From requirement (iii) of Definition 1 we get $\theta \models \neg\psi$, leading to $\mathbb{B}(\mathscr{A}) \models \neg\psi$ which contradicts the assumption $\mathbb{B}(\mathscr{A}) \not\models \neg\psi$.

Claim (M_{Qual}) follows from requirement (i) of Definition 1 with requirements (S1) and (S2).

Claim (M_{Rel}) follows directly from requirement (ii) of Definition 1. □

We have not required that the asserter follows the maxims strictly but her assertions satisfy somewhat weaker requirements. The asserter constructs her next candidate assertion by first (possibly tentatively) revising her beliefs with θ and then adhering to the Maxim of Quality by asserting some ψ' which follows from these revised beliefs. However, if her own convictions $\mathbb{C}(\mathscr{A})$ preclude this revision, then she must adhere to this Maxim in some other way instead. In Algorithm 1 this other way was to end the dialogue in *disagreement* without making any other assertion ψ' at all. In section 6 we will consider an alternative way based on narrowing the topic T instead. Moreover, the vocabulary is not tied to the receiver's last assertion but to the topic of the conversation defined by her first assertion. As straightforward consequences of Theorem 1 and Definition 5 we have the following:

Theorem 4 *Let T be the topic of the conversation and γ be what the receiver has told the asserter so far during their dialogue. The asserter's next candidate assertion ψ' satisfies*

$$\mathbb{B}(\mathscr{A} \circ_A \gamma) \not\models \neg\psi' \qquad (M'_{Qual})$$
$$\mathrm{voc}(\psi') \subseteq T. \qquad (M'_{Rel})$$

Proof Claim (M'_{Qual}) follows from Theorem 1 and requirement (i) of Definition 5.
Claim (M'_{Rel}) follows directly from requirement (ii) of Definition 5. □

As mentioned, Maxim of Quantity refers to the purpose of the conversation which is something we have not fixed beforehand. Supposedly the asserter's goal in her first assertion is to inform the receiver of something that she thinks he does not already believe. The subsequent assertions aim to either inform the receiver about a modified piece of information (that typically still contains information that is new to the receiver) or to convey the information that the asserter's epistemic state has changed.

The way how the contributions θ of the receiver follow the other part (b) in the Maxim of Quantity hinges on what he considers the purpose of the current exchange to be. This purpose is at least to reject somehow the assertion ψ from the asserter, if it conflicts with his own convictions $\mathbb{C}(\mathscr{B})$. Algorithm 2 takes the *narrow* view that this is the whole purpose of their exchange, and hence states that the receiver can restrict his attention into just the first rebuttal that comes to mind. Recall that he will have to check whether her assertion ψ is consistent with his convictions $\mathbb{C}(\mathscr{B})$ anyway, and that if they are not, then a corresponding Craig interpolant θ can be extracted from the corresponding proof R of their inconsistency with little extra cost.

However, Example 4 showed that even a brief and natural θ can already contain more information than required for merely rejecting an assertion ψ. Accordingly, an alternative *broader* view of the purpose of the current exchange could be to provide more information for the asserter about $\mathbb{C}(\mathscr{B})$, so that she could form her subsequent assertions ψ' more precisely, since that might end their dialogue sooner. Bob's reply in Example 4 could be considered to still satisfy part (b) in the Maxim of Quantity in this broader view: He anticipates that if he now said only $p \to \neg q$ then Amy would probably continue with something like $p \wedge r$ which he would have to reject later with $p \to \neg r$ anyway, so their dialogue would be shorter if he just stated both these rebuttals at once. Unfortunately, such anticipation might turn out to be unfounded, since the asserter might steer their subsequent dialogue in some unexpected direction instead. Hence it seems prudent for the receiver to refrain from spending resources in such anticipation, since these resources might turn out to be misplaced.

Example 9 The broadest possible view would lead into the following kinds of dialogues:

1. First the asserter makes her initial candidate assertion ψ.
2. If this ψ is consistent with the receiver's convictions $\mathbb{C}(\mathscr{B})$ then he admits ψ to be believable, and the dialogue ends accordingly.
3. Otherwise he continues the dialogue by rejecting it with $\theta = \pi_{\mathrm{voc}(\psi)}\mathbb{C}(\mathscr{B})$; in other words, he tells her *all* his convictions about her chosen topic straight away.
4. If this θ is inconsistent with the asserter's convictions $\mathbb{C}(\mathscr{A})$ then she ends the dialogue in *disagreement*.
5. Otherwise she makes another candidate assertion ψ' based on this θ.
6. He admits that this second candidate assertion ψ' is believable because the way how he constructed his θ ensures that ψ' must be consistent with his convictions.

These are as short as dialogues with rebuttals can be. Unfortunately the receiver must spend more effort in constructing his only rebuttal θ as the uniform instead of just some Craig interpolant, and this might be considered to violate part (b) in the Maxim of Quantity.

If both the asserter and the receiver use uniform interpolation, then their dialogue proceeds as in Example 9 with the addition that its outcome is now completely determined by their epistemic states and topic without any strategic choices, because uniform interpolants are unique up to logical equivalence.

5.2 Dialogue length

Let us consider how the execution of the protocol proceeds. First we show that it terminates eventually.

Theorem 5 (termination) *The protocol depicted as Algorithms 1 and 2 always terminates. Moreover, the asserter makes at most* $2^{|T|}$ *assertions during it.*

Proof We note first that $\mathrm{voc}(\gamma) \subseteq T$ in Algorithm 1: By construction $\gamma = \theta_1 \wedge \theta_2 \wedge \theta_3 \wedge \ldots \wedge \theta_n$, where each θ_i is the rejection the receiver gave to the ith candidate assertion ψ_i made by the asserter in the current dialogue so far. Each θ_i is therefore a Craig interpolant for the entailment $\mathbb{C}(\mathscr{B}) \models \neg\psi_i$, by line 2 of Algorithm 2. Requirement (ii) of Definition 5 ensures that $\mathrm{voc}(\psi_i) \subseteq T$. Hence also $\mathrm{voc}(\theta_i) \subseteq T$ by part (ii) of Definition 1.

After noting this, it suffices to show that γ gets logically strictly stronger at each execution of the **while** loop body: Assume to the contrary that $\gamma \models \theta$. On one hand we have $\psi \models \gamma$ by requirement (iii) of Definition 5. On the other hand we have $\theta \models \neg\psi$ by line 2 of Algorithm 2. Hence we would have $\psi \models \neg\psi$ which violates requirement (i) of Definition 5. □

Now that Theorem 5 has shown that every dialogue does terminate, let us turn to two factors which affect the length of the dialogue: how informative the receiver is in his replies, and how fine epistemic distinctions the asserter draws about their topic. Let us illustrate their effects with three examples about the following "guessing game":

Example 10 Amy and Bob are trying to agree on the value of some n-bit quantity $q \in \mathbb{N}$. Their topic is $T = \{b_0, b_1, b_2, \ldots, b_{n-1}\}$ where b_i denotes bit i of q. Let us denote the assertion "$q = m$" as β_m; that is, the formula β_m is a conjunction of literals, which contains each $b_i \in T$ exactly once, and this b_i occurs positively in β_i exactly when bit i of m is 1. Bob is convinced that β_0. Amy in turn considers β_i more plausible than β_j whenever $i > j$.

Suppose that Bob co-operates only *minimally* with Amy by replying to each of her assertions ψ with just its negation $\neg\psi$. Note that the protocol in section 4.3 permits such bluntness, because it only requires that his refutation θ must be *some* interpolant for the inconsistency of his convictions with her assertion ψ in the sense of Definition 1. Technically, he could use as line 2 of his Algorithm 2 for instance the way to extract interpolants which is described in Appendix A equipped with the corresponding *weak* labelling.

In this case their dialogue progresses as follows: Amy begins with her most plausible candidate assertion "$q = 2^n - 1$" as the formula β_{2^n-1}. Bob responds tersely with $q \neq 2^n - 1$ using its negation $\neg\beta_{2^n-1}$. She then continues with her second candidate assertion β_{2^n-2}, which he again rejects with $\neg\beta_{2^n-2}$. They continue in this way, until she finally asserts β_0, which he admits. They have therefore reached the upper limit $2^{|T|} = 2^n$ of Theorem 5 for conversation length.

Let us then modify the receiver's behaviour in Example 10:

Example 11 Suppose that Amy's epistemic state is still as in Example 10, but now Bob is no longer minimally informative. Instead, he helps her by using the *strong* labelling. Then she starts again their dialogue with the same assertion β_{2^n-1}. Now, however, his reply is always some literal $\neg b_i$ saying "no, bit i must be 0". Such a reply eliminates half of her remaining candidate assertions. For instance, if he replies to β_{2^n-1} with $\neg b_0$, then she knows that q must be even, and so her remaining candidate assertions are $\beta_0, \beta_2, \beta_4, \ldots, \beta_{2^n-2}$. She will then assert the most plausible of them, and that is β_{2^n-2}. In this way, she clears the bits one by one. Their entire dialogue thus consists of only n assertions altogether.

Comparing Examples 10 and 11 reveals that the more information the receiver conveys about his convictions in his replies, the more precise assertions the asserter can come up with, and the fewer assertions she needs. Moreover, he can adjust his level of informativity, by for instance adopting different labellings. This was already noted in Example 9 which considered the *maximal* level of informativity, namely using uniform interpolation instead of Craig interpolation.

Let us then turn from the receiver's behaviour to the asserter's behaviour:

Example 12 Suppose that Bob is again only minimally informative as in the original Example 10. Let Amy now draw coarser epistemic distinctions in her epistemic state than before, so that she considers two values for q equally plausible if they have the same highest 1-bit. That is, she now considers the whole block $B_i = \{\beta_{2^{i-1}}, \beta_{2^{i-1}+1}, \beta_{2^{i-1}+2}, \ldots, \beta_{2^i-1}\}$ of formulas equally plausible, where we take $B_0 = \{\beta_0\}$. She still considers such a block B_i more plausible than another block B_j when $i > j$. Now their conversation ends after only $n+1$ assertions: She begins by asserting $\bigvee B_n$ to which he responds $\neg \bigvee B_n$. She then continues by asserting $\bigvee B_{n-1}$, and so on.

Comparing Examples 10 and 12 reveals that the finer epistemic distinctions the asserter draws between different possible states of affairs in her epistemic state $\mathscr{A}$, the more assertions she can make about the topic T. Requirement (v) of Definition 5 in turn assures that she will indeed make them if she can. In Example 10, Amy drew as fine distinctions as possible by considering none of the 2^n distinct choices $\beta_0, \beta_1, \beta_2, \ldots, \beta_{2^n-1}$ as equally plausible. In Example 12 she grouped them more coarsely into only $n+1$ blocks $B_0, B_1, B_2, \ldots, B_n$ so that she considered all choices in the same block as equally plausible.

These examples let us conclude the following: First, this number of epistemic distinctions drawn by the asserter on the topic T is a sharper upper bound on dialogue length than their maximum number $2^{|T|}$ used in Theorem 5. And second, whether the dialogue ends sooner than this upper bound depends on how rapidly the receiver's replies eliminate these distinctions from further consideration. In particular, these two conclusions show that Example 10 is essentially the only case with this maximum length: the asserter must draw as fine distinctions as possible, while the receiver must eliminate them only one by one. Drawing sharper conclusions about dialogue length would in turn require devising more precise measures for these two factors than we have done here.

5.3 Outcome of the believability restoration protocol

Let us now turn to what the asserter has attained at the end of the dialogue. Let us first justify her decision to end it in disagreement. She has namely then detected that the receiver's refutations to her candidate assertions witness that their convictions disagree on their topic.

Theorem 6 (disagreement) *Algorithm 1 terminates with disagreement if and only if $\mathbb{C}(\mathscr{A})$ and $\mathbb{C}(\mathscr{B})$ disagree on the current topic T in the sense of Definition 4. Moreover, γ is a witness for the disagreement.*

Proof Theorem 5 showed that Algorithm 1 always terminates.

The forward direction of the first claim is shown as follows: Assume that *disagreement* = TRUE at termination. Thus the asserter's convictions $\mathbb{C}(\mathscr{A})$ entail $\neg\gamma$ by line 7 of Algorithm 1. Each value of θ is in turn entailed by the receiver's convictions $\mathbb{C}(\mathscr{B})$ by line 2 of Algorithm 2 and part (i) of Definition 1. Hence their conjunction γ is also entailed by $\mathbb{C}(\mathscr{B})$.

Its other direction is in turn shown as follows: Assume that $\mathbb{C}(\mathscr{A})$ and $\mathbb{C}(\mathscr{B})$ disagree on the current topic T. Then also $\pi_T\mathbb{C}(\mathscr{A})$ and $\mathbb{C}(\mathscr{B})$ are inconsistent with each other, by the witness of that disagreement and Definition 2. Hence the asserter cannot make any candidate assertions which the receiver does not reject, by Theorem 1.

For the second claim, we note here that the forward part of the first claim verified that $\mathbb{C}(\mathscr{B}) \models \gamma$ but $\mathbb{C}(\mathscr{A}) \models \neg\gamma$. The remaining part $\mathrm{voc}(\gamma) \subseteq T$ of Definition 4 was in turn already noted in the proof of Theorem 5. □

By Theorem 6, Algorithm 1 terminates without disagreement if and only if $\mathbb{C}(\mathscr{A})$ and $\mathbb{C}(\mathscr{B})$ do not disagree on the current topic T.

Let us now consider what the asserter has achieved if the dialogue ends in agreement. Her final assertion is also something that the receiver too can believe. Moreover, if her final candidate assertion is precise and both agents choose to believe it, then the receiver believes everything that the asserter believes about the topic of their dialogue. In other words, precise candidate assertions do avoid the problem in Example 7.

Theorem 7 (no disagreement) *Algorithm 1 terminates without disagreement if and only if the final candidate assertion ψ is consistent with $\mathbb{C}(\mathscr{B})$. Moreover, if ψ is precise then $\mathbb{B}(\mathscr{B} \circ_B \psi) \models \psi'$ for every ψ' such that $\mathbb{B}(\mathscr{A} \circ_A \psi) \models \psi'$ and $\mathrm{voc}(\psi') \subseteq T$.*

Proof Theorem 5 shows that Algorithm 1 always terminates.

The first claim holds, because it terminates with *disagreement* = FALSE if and only if the test on line 1 of Algorithm 2 fails.

The second claim can then be shown as follows: By Theorem 2, the precise candidate assertion ψ tells exactly what the asserter would believe about the topic T if she revised her epistemic state with γ. Recall that $\mathrm{voc}(\gamma) \subseteq T$, as already noted in the proof of Theorem 5. Hence revision with ψ instead of γ would yield these same beliefs. Therefore the assumption can be rewritten as $\mathbb{B}(\mathscr{A} \circ_A \gamma) \models \psi'$. Thus $\psi \models \psi'$ by Theorem 2, and $\mathbb{B}(\mathscr{B} \circ_B \psi) \models \psi'$ follows via the success postulate (CR1) for $\circ_B$ and the first claim above. □

Once the two agents have reached an agreement ψ in this way, they can proceed as follows: The asserter can adopt ψ as her own belief, if she has not yet done so, by requirement (iv) of Definition 5. The receiver can in turn adopt ψ as his belief, since it does not conflict with his own convictions, by the first claim of Theorem 7 above.

By Theorem 7, the kind of agreement reached depends only on how the asserter chose her final assertion ψ, but not on her earlier assertions or the receiver's replies to them. (However, her earlier assertions can affect his replies, and hence the meaning of this final ψ.) Naturally, if she does not know when their dialogue will end, then her every assertion must be precise, if she wants to be certain that his resulting beliefs do not disagree with hers on their topic T.

However, Theorem 7 also permits the asserter to adopt a mixed strategy. For instance, she might begin their dialogue with some imprecise candidate assertion ψ_1. If the receiver rejects it, then her subsequent candidate assertions $\psi_2, \psi_3, \psi_4, \ldots, \psi_n$ could be precise. In this way either he agrees with her preferred ψ_1 or he admits the final precise alternative ψ_n to which she had to resort instead.

Finally, if the receiver detects that the dialogue has ended prematurely without reaching any outcome for some reason, then all is not necessarily lost, because he could invoke accommodative belief revision from section 3.1 with the asserter's most recent assertion, thus guessing what she might have come up with if the dialogue could have run its full course. Whether he should resort to such recovery or not depends on the situation.

5.4 On what else can happen during a dialogue

We have tacitly assumed that the asserter and the receiver concentrate only on their dialogue, so that no external factors can change their epistemic states during it. This simplifying assumption is indeed common in for instance argumentation-based dialogues, but it might be beneficial to permit such concurrency from a computer science perspective. We shall therefore consider briefly how this simplifying assumption might be relaxed. Two issues arise: when and what kinds of changes could be allowed?

Regarding the first issue, an agent first decides what to do next based on its current epistemic state and then proceeds to act accordingly in our protocol. Thus the receiver first decides whether he could believe the asserter's most recent assertion ψ or not (on line 1 of his Algorithm 2) and then proceeds to either admit ψ (on line 4) or reject it (on line 3). Similarly, the asserter first decides whether she should continue or end their dialogue and then proceeds to do so (on lines 7–10 of her Algorithm 1). Clearly the epistemic state must not change between this decision and its following action. If that is guaranteed, then Theorems 6 and 7 hold for those epistemic states $\mathscr{A}$ and $\mathscr{B}$ on which the agents based their final actions.

Regarding the second issue, the receiver bases his decisions solely on his convictions $\mathbb{C}(\mathscr{B})$. Hence he can change his other beliefs freely during a dialogue. Moreover, we have assumed that $\mathbb{C}(\mathscr{B})$ changes only monotonically in section 2.6. This assumption in turn ensures that $\mathbb{C}(\mathscr{B}) \models \gamma$ continues to hold during the asserter's Algorithm 1. Thus the receiver can for instance carry out simultaneous dialogues with several asserters at the same time, and Theorems 5–7 still hold for each of these dialogues separately.

In contrast, permitting the asserter to carry out similar simultaneous dialogues with several receivers at the same time would be significantly more complicated. None of the receivers would note that she is having simultaneous dialogues with other receivers, because they do not keep track of her preceding assertions. However, they might give her rebuttals which conflict each other, and she would have to determine somehow whom to believe.

6 Extending the believability restoration protocol to circumvent conflicting convictions

Let us now consider the situation where the protocol in Algorithm 1 leads into disagreement: That is, when the asserter detects that her own convictions $\mathbb{C}(\mathscr{A})$ are inconsistent with the convictions γ stated by the receiver. Then the asserter cannot continue their dialogue directly via Definition 5.

We illustrate such an inconsistency with the following simple example inspired by Fermé and Hansson (1999):

Example 13 ("Dinosaur broke grandma's vase!") When Bob returns home, his daughter Amy tells him that a dinosaur has broken grandmother's vase in the living room. Because Bob is convinced that dinosaurs do not exist, he replies with an irretractable assertion to the contrary: "I cannot believe you, because dinosaurs do not exist." But by dinosaur Amy really meant her toy dinosaur, and because she is convinced that it exists, the dialogue ends in disagreement. In this example the inconsistency arises from an ontological mismatch, not from an epistemic one. Instead of ending the dialogue, the agents could start a suitable subdialogue (for example, ontological negotiation (van Diggelen et al. 2007) or meaning-based argumentation (Laera et al. 2007)) in order to resolve the mismatch. But if such a

protocol is not available or it ends unsuccessfully, Amy can still try to convey some part of her message by narrowing the topic. She could focus only on the vase, its breakdown, the grandmother and the living room, but avoid the dinosaur issue altogether.

Section 6.1 develops this narrowing method and the corresponding modifications to Algorithm 1 on a general level, while section 6.2 focuses on its details.

6.1 Narrowing the topic to avoid disagreements

One way how the asserter can deal with the inconsistency between γ and $\mathbb{C}(\mathscr{A})$ she has now detected is to avoid those parts of their topic T where this inconsistency has manifested itself in her further candidate assertions, and to concentrate instead on those parts of T where she and the receiver might still reach an agreement. In other words, the asserter restricts the topic of her next candidate assertion into some subtopic $T' \subset T$ such that the detected inconsistency of $\mathbb{C}(\mathscr{A})$ with γ has no witnesses in this chosen T':

Example 14 Let Amy be convinced that $x \wedge y$, and let Bob be convinced that $\neg y$, but let him hold no convictions on x. She starts their dialogue on the topic $T = \{x, y\}$ with her conviction about T. He rejects her assertion with his conviction about T. Hence she sees that their convictions disagree on T. However, instead of ending their dialogue in disagreement, she realizes that this disagreement is only about the subtopic $\{y\}$, and thus she might still attain an agreement on the other subtopic $T' = \{x\}$. Accordingly she can continue their dialogue by asserting her conviction x about it, which he admits, as in Example 6.

Example 15 Let Amy assert her conviction $x \wedge (y \leftrightarrow z)$, and let Bob reject it with $\neg(y \leftrightarrow z)$. She can now conclude that it would be fruitless to continue with any subtopic containing both y and z, because she has now seen that his convictions exclude every combination of truth values for y and z which her own convictions permit. However, she can also conclude that he could still admit either y or z separately. Moreover, he did not comment on x in any way. Hence she could continue their dialogue with either of the subtopics $\{x, y\}$ or $\{x, z\}$.

Example 15 shows that the asserter cannot rule out all the propositional symbols in her own assertion or in its rebuttal. Instead, she must consider carefully which symbol combinations she can keep and which she must rule out.

Let us now develop this idea into a protocol for the asserter. Our starting point is her Algorithm 1, into which we shall add the current subtopic $T' \subseteq T$. Her aim is to continue the dialogue with another candidate assertion from Definition 5, except now T is replaced with T' and γ is replaced with the uniform interpolant $\gamma' = \pi_{T'}\gamma$ or what the receiver has told her about T'. She chooses T' so that this γ' is consistent with her convictions $\mathbb{C}(\mathscr{A})$. (Initially this subtopic T' could well be their whole topic T.)

However, we must rule out trivial choices for T' first. For instance, choosing $T' = \emptyset$ is always permitted, because then $\gamma' = \top$. Unfortunately the asserter's next candidate assertion would then also be $\psi = \top$. The receiver would in turn trivially accept her assertion, but he would not have to alter his beliefs in any way. In general, she should refrain from asserting tautologies, because they do not advance her aim to affect his beliefs.

Let us postpone how the asserter could choose such a subtopic T' to section 6.2. Combining the development above with Theorem 1 leads to the following variant of Algorithm 1:

Algorithm 3: The asserter's protocol with subtopics.

1 The asserter has **assert**ed some ψ to the receiver who has **reject**ed it with θ;
2 $T = \mathrm{voc}(\psi)$;
3 $\gamma = \top$;
4 *disagreement* = FALSE;
5 **while** (**not** *disagreement*) **and** the receiver **reject**ed with θ
6 $\gamma = \gamma \wedge \theta$;
7 **if** there is some $T' \subseteq T$ such that $\gamma' = \pi_{T'}\gamma$ is consistent with $\mathbb{C}(\mathscr{A})$ **and** the corresponding precise candidate assertion $\pi_{T'}\mathbb{B}(\mathscr{A} \circ_A \gamma')$ is falsifiable
8 **choose** some falsifiable candidate assertion ψ according to Definition 5 but with T' and γ' instead of T and γ;
9 **assert** ψ to the receiver
10 **else** *disagreement* = TRUE.

Algorithm 3 chooses a subtopic T' anew for each subsequent candidate assertion. One natural possibility would be to stick with the current T' for as long as possible, and choose another only after its corresponding γ' has finally become inconsistent with $\mathbb{C}(\mathscr{A})$. In particular, if the whole original topic T can still be chosen as T', then Algorithm 3 still behaves like Algorithm 1; if this choice is no longer possible, the former may still be able to continue with other smaller choices for T', whereas the latter must end in disagreement. However, the asserter might have other strategic preferences besides falsifiability in choosing her next candidate assertion ψ, and they might warrant changing the subtopic sooner. This nondeterministic choice for T' permits both of these two strategic possibilities.

Example 16 Algorithm 3 handles Example 14 as follows:

1. Amy begins the dialogue by asserting $\psi = x \wedge y$, which Bob rejects with $\neg y$.
2. Bob's rejection is inconsistent with Amy's own convictions, so she realizes that she must now restrict herself to some subtopic T' which avoids y. Her possibilities are $\{x\}$ and $\emptyset$.
3. However, choosing $T' = \emptyset$ would produce $\top$ as its corresponding precise candidate assertion, so her only choice is in fact $T' = \{x\}$.
4. Hence she forms her next assertion as $\psi = x$ which Bob admits.

In this way Amy has found an assertion x which they both can believe by narrowing the topic, even though their dialogue revealed that their convictions disagree on y.

Let us next sketch how replacing Algorithm 1 with Algorithm 3 affects Theorems 5 to 7. We first note that termination is not affected as such:

Theorem 8 *The protocol depicted as Algorithms 3 and 2 always terminates. Moreover, the asserter makes at most* $2^{|T|}$ *assertions during it.*

Proof The argument in the proof of Theorem 5 can be adapted as follows to yield the same bound as before.

We first note that $\mathrm{voc}(\gamma) \subseteq T$ holds also for Algorithm 3 by almost the same argument as before. The only modification is that now $\mathrm{voc}(\psi_i) \subseteq T_i'$ and hence $\mathrm{voc}(\theta) \subseteq T_i'$ for the corresponding ith subtopic $T_i' \subseteq T$.

After noting this, we can proceed almost as before. However, here the contrary assumption $\gamma \models \theta$ entails $\gamma' \models \theta$ by Definition 2 since $\mathrm{voc}(\theta) \subseteq T'$. From it we can then derive the desired contradiction as before but with γ' instead of γ. □

However, Theorem 8 ignores the considerable extra work the asserter spends on line 7 in choosing her current subtopic T'. We shall discuss it in section 6.2.

We note next that Algorithm 3 terminates in disagreement just in case the asserter has deemed it pointless to continue the dialogue:

Theorem 9 *Algorithm 3 terminates with disagreement if and only if all the asserter's remaining candidate assertions are tautologies.*

Proof The test on line 7 fails if and only if for each choice $T' \subseteq T$ either $\pi_{T'}\gamma$ is inconsistent with $\mathbb{C}(\mathscr{A})$ or the asserter's precise candidate assertion about T' is a tautology. Consider each case separately as follows.

In the former case, the asserter cannot make any candidate assertions about T' at all, by requirements (i), (iii) and (iv) of Definition 5.

In the latter case, the asserter can still make them, but they are all tautologies, since they are all entailed by her precise candidate assertion by requirement (vi) of Definition 5. □

We note finally that if Algorithm 3 terminates without disagreement, we again have the same outcome as in Theorem 7, but only for the final subtopic T' instead of the whole original topic T:

Theorem 10 *Algorithm 3 terminates without disagreement if and only if the final candidate assertion ψ is consistent with $\mathbb{C}(\mathscr{B})$. Moreover, if this ψ is a precise candidate assertion with respect to the final subtopic T' then $\mathbb{B}(\mathscr{B} \circ_B \psi) \models \psi'$ for every ψ' such that $\mathbb{B}(\mathscr{A} \circ_A \psi) \models \psi'$ and $\mathrm{voc}(\psi') \subseteq T'$.*

Proof The argument in the proof of Theorem 7 can be adapted as follows.

The first claim follows exactly as before, since neither Algorithm 2 nor the test of the **while** loop have been modified.

The second claim follows almost as before, but with T' and γ' instead of T and γ. The only difference is that now $\mathrm{voc}(\gamma') \subseteq T'$ follows from Definition 2 instead. □

As already mentioned just before Example 16, the asserter might prefer some candidate assertions ψ over others; in other words, she might prefer some subtopics T' over others. Theorem 10 indicates that she should naturally try these T' in their preferred order so that Algorithm 3 might terminate with the corresponding preferred ψ. However, Algorithm 3 considers only epistemic factors, and they alone might not suffice to determine such preferences in general:

Example 17 Consider the conversation in Example 1 and its formalisation in Example 3. Add to Amy's convictions the proposition that the bird she saw was indeed a three-toed woodpecker with a red forehead, that is, the formula $p \wedge q$. Then, Bob's reply $p \rightarrow \neg q \wedge \neg r$ to her original assertion conflicts with this conviction of hers. One way to continue the dialogue could be by dropping either p or q from the present topic, because his convictions allow either one of them to be true and only excludes the possibility that both are true simultaneously. However, since she is convinced that both are true, dropping one of them out of discussion is arbitrary according to her epistemic state. Another way could be by asserting some combination of p and q, like for instance $p \vee q$, but she is now aware that even though such an assertion might result in an agreement in the end, she and Bob would still disagree about their topic.

6.2 Choosing the next subtopic

Let us now consider in more detail how the complicated line 7 of Algorithm 3 could be implemented. We introduce the set F_1 of *forbidden* subtopics: $T'' \in F_1$ if the asserter has already observed that her convictions $\mathbb{C}(\mathscr{A})$ are inconsistent with the corresponding $\gamma'' = \pi_{T''}\gamma$. Then her next chosen subtopic T' must be such that $T' \not\subseteq T''$ for each $T'' \in F_1$, because otherwise choosing T' would encounter again the same inconsistency which caused the addition of T'' into F_1.

The problem of choosing such a T' leads to the following well-known **NP**-complete problem (Garey and Johnson 1979, Problem SP8):

Definition 6 An *instance* of the HITTING SET PROBLEM is some set C of subsets of some finite universe U. A *solution* for this instance is some $S \subseteq U$ such that $S \cap D \neq \emptyset$ for every $D \in C$. This S is a *minimal* solution if none of its proper subsets is also a solution.

In our setting, the whole original topic T forms the universe U, the forbidden subtopics F_1 form C, and the candidates for the next subtopic T' are formed as $T \setminus S$ where S is a solution for F_1. That is, we choose a solution S which contains at least one element from every forbidden subtopic $T'' \in F_1$, and omit them from T' which therefore does not contain any of these $T'' \in F_1$. Moreover, if the asserter prefers some subtopics over others, then these preferences become preferences among these solutions S. For instance, if these preferences are given as a priority for each individual propositional symbol $x \in T$ telling how important it is to discuss this x, then such a priority becomes the penalty for including x into S, and a solution with minimal overall penalty is sought.

These forbidden subtopics F_1 handle the first half on line 7 of Algorithm 3. Its other half concerns the subtopics which are not forbidden, but which would lead into a tautological candidate assertion. We can employ the same design principle for them as well. Thus we introduce another hitting set instance F_2 for them: $T'' \in F_2$ if the asserter has already observed that the corresponding precise candidate assertion $\delta'' = \pi_{T''}\mathbb{B}(\mathscr{A} \circ_A \gamma'')$ is a tautology.

The conjunction on line 7 is the union $F_1 \cup F_2$ of these F_1 and F_2. We keep them separate, because they behave differently as γ grows during the dialogue: Once some subtopic becomes forbidden, then it stays forbidden for the rest of the dialogue; hence F_1 can only grow from its initial value $\emptyset$ during a dialogue. In contrast, this does not hold for the other half of line 7, because when γ grows these precise candidate assertions δ'' change too; hence F_2 must be reset back into its initial value $\emptyset$ when γ grows, so that the asserter reconsiders those subtopics which produced tautologies with the previous value of γ, because they might no longer do so with the new value of γ.

This more precise design for line 7 leads to the following more precise form of Algorithm 3:

Algorithm 4: The asserter's protocol with subtopic selection.

```
1  The asserter has asserted some ψ to the receiver who has rejected it with θ;
2  T = voc(ψ);
3  γ = ⊤;
4  F1 = ∅;
5  disagreement = FALSE;
6  while (not disagreement) and the receiver rejected with θ
7      γ = γ ∧ θ;
8      F2 = ∅;
9      chosen = FALSE;
10     while (not chosen) and the hitting set instance F1 ∪ F2 still has other solutions besides the whole T
11         choose some such solution S;
12         T′ = T \ S;
13         γ′ = π_{T′} γ;
14         if γ′ is inconsistent with ℂ(𝒜)
15             F1 = F1 ∪ {T′}
16         elseif π_{T′} 𝔹(𝒜 ∘_A γ′) is a tautology
17             F2 = F2 ∪ {T′}
18         else chosen = TRUE;
19     if chosen
20         choose some falsifiable candidate assertion ψ according to Definition 5
           but with T′ and γ′ instead of T and γ;
21         assert ψ to the receiver
22     else disagreement = TRUE.
```

The new inner **while** loop in Algorithm 4 terminates, because its hitting set instance $F_1 \cup F_2$ grows at each iteration with the current $T' \notin F_1 \cup F_2$. It terminates without having chosen anything if the only remaining choice for the next subtopic would be the trivial $T' = T \setminus T = \emptyset$ which would result in the tautological assertion $\psi = \top$.

We do not require minimality of the solution S for the hitting set instance $F_1 \cup F_2$ on line 11 of Algorithm 4. This permits using an approximation algorithm for the hitting set problem. However, a larger than minimal S also leads the asserter to choose a narrower subtopic T' than necessary, which in turn leads her into a less detailed assertion ψ than possible. It depends on the overall situation where this subdialogue is carried out whether such lack of detail can be tolerated or not. One natural preference between subtopics would indeed be to prefer more detailed assertions ψ, which corresponds to this minimality of solutions.

Algorithm 4 terminates in fewer rounds if its hitting set instance $F_1 \cup F_2$ consists of smaller sets. In fact, we can optimize its line 15 in this regard as follows: When this line gets executed, the asserter has constructed some proof R of the inconsistency of γ' with $\mathbb{C}(\mathscr{A})$. Since our task is to exclude this inconsistency from further consideration, we can add into F_1 only those propositional symbols which occur in R instead of the whole T'. This optimization can even be strengthened further by adding only those symbols which occur in the Craig interpolant corresponding to R under the *small* labelling given as Equation (6) in Appendix A.

The combinatorial structure of the hitting set instance $F_1 \cup F_2$ yields an estimate on the number of steps taken by Algorithm 4. (Note, however, that many of its steps are already costly by themselves.) It is namely a *Sperner system* on T by the choice of T', and therefore its maximum size is $|F_1 \cup F_2| \leq \binom{|T|}{\lfloor |T|/2 \rfloor} = \mathscr{O}(2^{|T|-1})$ (Bollobás 1986, §3, Theorem 1). This in turn yields two bounds: First, this is an upper bound on the total number of times F_1 can grow during a dialogue. Moreover, the aforementioned optimization for line 15 strives to

make the elements of $F_1 \cup F_2$ small, and so $|F_1 \cup F_2|$ tends to be smaller than this upper bound, by the LYM inequality (Bollobás 1986, §3, Theorem 2). And second, this is also an upper bound on the number of times F_2 can grow between two consecutive assertions.

However, the set F_2 is typically much smaller than this pessimistic upper bound, because $\pi_{T'}\mathbb{B}(\mathscr{A} \circ_A \gamma')$ is a tautology if and only if both γ' and $\pi_{T'}\mathbb{B}(\mathscr{A})$ are tautologies too, by the success postulate (CR2). (Line 16 of Algorithm 4 can be optimized accordingly.) Thus a subtopic T' causes a tautological candidate assertion only if the asserter has no beliefs about it and the receiver has given her no information about it either. Indeed, she typically would not have included such an uninteresting subtopic T' into her overall topic T in the first place.

7 Related work

The approach presented here differs from those proposed in the literature in that it combines belief revision methods and argumentation-based communication methods in a unique way: On the one hand, we use communication to resolve conflicts between convictions and incoming information in order to find out how to revise beliefs. On the other hand, we use (possibly hypothetical) belief revision when formulating new assertions during communication. We also use non-prioritized belief revision in cases in which communication is not feasible. We will discuss related work in the field of belief revision in section 7.1 and in the field of argumentation-based dialogues in section 7.2.

7.1 Belief revision

Our work shares with non-prioritized belief revision the idea that an agent's epistemic state contains some information that it refuses to give up in light of new information. Thus new information is not necessarily prioritized over previously held information. If the incoming information happens to be in conflict with the unrevisable part of the agent's epistemic state and further communication is not possible, an agent must either reject the information or try to learn something from it. For the latter we propose accommodative belief revision. Let us compare accommodative belief revision with some related methods of non-prioritized revision.

Hansson (1997) introduces the term semi-revision for operators that assign no indefeasible priority to new information. Accommodative revision falls under this general characterisation but does not agree with one interpretation Hansson gives to it, namely the idea that revision by a formula might result in a deliberation whether the formula or its negation (or neither) will be accepted. This idea leads to the postulate of negation-neutrality which states that a semi-revision by a formula is equivalent with semi-revision with the negation of that formula. In this method, the decision between accepting a formula or its negation is independent of the input and is therefore not in accordance with accommodative revision that tries to accommodate as much information from the input as possible given the constraints set by the agent's convictions.

Unlike accommodative revision, many approaches to non-prioritized revision make the assumption of relative success, which means that a revision with a formula is either successful or leaves the agent's beliefs unchanged (see, e.g., Hansson et al. 2001). As a special case of accommodative revision, we will get screened revision (Makinson 1997) by defining that $\varphi * \kappa \equiv \kappa$ in Equation (2) whenever φ is unbelievable. Our proposal resembles the proposal by Bellot et al. (1997), which also first revises incoming information with the convictions

and then revises the epistemic state with the result of the first revision, but uses the same fixed distance-based revision operator in both cases. Our proposal lets the agent choose the two components separately, without imposing limitations on the representation of epistemic states. Thus our proposal is a generalization of theirs.

A selective revision operator $\circ$ is defined by the equality $\mathscr{S} \circ \alpha = \mathscr{S} * f(\alpha)$, where $*$ is an AGM revision operator and f a function that intuitively speaking selects the credible part out of the input sentence (Fermé and Hansson 1999). According to the authors, the transformation typically has the property $\vdash \alpha \rightarrow f(\alpha)$. Accommodative revision can be seen as a selective revision operator for which this property does not hold. However, the function $f(\alpha, \mathbb{C}(\mathscr{S}))$ of accommodative revision is not a function of input only, but a function of input and convictions.

There are other approaches in which interaction is used as a preprocessing step before belief revision, but since the possibility of agents having their private convictions is not considered, these approaches fall to the category of prioritized belief revision. These include such merging approaches as mutual belief revision (Jin et al. 2007) and belief negotiation (Booth 2006) in which the beliefs of all the agents are weakened until they no longer contradict each other. Our aim is not that the agents always merge their epistemic states; instead, they have the opportunity to refuse to accept claims that they find unbelievable. Moreover, our setting 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 information about his convictions in case he finds the original assertion unbelievable.

Such a setting is natural in some application areas, for instance, in 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. However, the motivation for asymmetry in our approach does not stem from such application areas but from our wish to model the exchange of information explicitly as communication between agents. Instead of defining a function that merges two epistemic states, we want to make our approach available in situations in which the agents are not willing or able to reveal their full epistemic states but communicate by making assertions concerning some topic to each other. In such dialogues, asymmetry is forced by the idea of turn-taking: at any given point one agent is making an assertion and the other agent is considering whether to accept it. Let us now turn to approaches that share such an idea of communication.

7.2 Argumentation-based dialogues

Certain types of argumentation-based dialogue protocols (Walton and Krabbe 1995; Parsons et al. 2003) can be viewed as preliminary 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 argumentation-based dialogues, agents can in turns perform various linguistic acts. The speaking of these locutions affects their *commitment stores*. Commitment stores (see Hamblin 1970, 257) are public data storages for keeping track of the propositional commitments of the agents. They enable ensuring that their utterances stay consistent during the dialogue.

By asserting a formula φ, or by accepting another agent's assertion of φ, an agent makes a propositional commitment to the truth of φ meaning that the agent is committed to treating φ as a true formula and, if challenged, to provide reasons for φ (Walton and Krabbe

1995). An agent can also reject a formula φ, or, in certain cases, retract an earlier propositional commitment. Typically, the following locutions are available to an agent X at time i with their effects on its commitment store $CS_i(X)$ with respect to a formula φ:

$$\begin{aligned} \mathit{assert}(\varphi)\colon\ & CS_i(X) = CS_{i-1}(X) \cup \{\varphi\} \\ \mathit{accept}(\varphi)\colon\ & CS_i(X) = CS_{i-1}(X) \cup \{\varphi\} \\ \mathit{retract}(\varphi)\colon\ & CS_i(X) = CS_{i-1}(X) \setminus \{\varphi\} \\ \mathit{reject}(\varphi)\colon\ & CS_i(X) = CS_{i-1}(X) \end{aligned}$$

Since we want the agents to communicate both their beliefs and their convictions, we will need, in addition to the standard locutions, two new ones, *irretractable rejection* and *irretractable assertion* which cannot be retracted from the commitment store once stated. Hamblin (1970) calls such statements which are marked as irretractable in commitment stores as *axioms*. Unlike Hamblin, we do not assume that these can be agreed upon (even less pre-agreed upon). Krabbe (2001) states that "[...] each participant's dark-side commitment set may be construed as a constraint upon the dialogue. A dark-side commitment brought to light will, on this stipulation, function as an irretractable principle for the rest of the dialogue." The irretractable locutions combine these thoughts, giving an agent an explicit way to signal that some statement is part of his dark-side commitment set and is therefore irretractable during the rest of the dialogue. Most of the recent work in argumentation seems to have ignored such dark-side commitments and irretractable assertions, but they fit well with our view of agents having both beliefs and convictions. In fact, our protocols can be used to complement existing argumentation protocols in this respect: Our protocols are designed to be applicable for generating conflict-resolution subdialogues within any conversation in which assertions of one agent can conflict with the convictions of another.

Using the irretractable locutions we can outline our subdialogue protocol presented as Algorithm 3 in the style of argumentation-based dialogues (see, e.g., Cogan et al. 2006) as follows, assuming that in the enclosing dialogue, the asserter has just asserted φ.

1. $\begin{cases} \textit{Receiver irretractably rejects } \varphi \textit{ and irretractably asserts } \theta & \text{if } \mathbb{C}(\mathscr{B}) \models \neg\varphi, \\ \text{Return to enclosing dialogue (end of subdialogue)} & \text{otherwise.} \end{cases}$

2. $\begin{cases} \textit{Asserter retracts } \varphi \textit{ and accepts } \theta & \text{if } \mathbb{C}(\mathscr{A}) \not\models \neg\theta, \\ \textit{Asserter retracts } \varphi & \text{else if subtopic available,} \\ \textit{Asserter rejects } \theta \text{ (end of the whole dialogue)} & \text{otherwise.} \end{cases}$

3. *Asserter asserts* ψ (where ψ is a candidate assertion consistent with all irretractable assertions made by the receiver as specified in Def. 5).
4. Go to 1 (substituting ψ for φ).

In step 1, if the receiver finds the assertion φ unbelievable, he irretractably rejects φ with an irretractable assertion θ as an explanation. Otherwise, any unbelievable assertion has been dealt with and the enclosing dialogue can continue. In step 2, if the asserter finds θ unbelievable and there is no suitable subtopic available, the asserter irretractably rejects θ and the conversation ends. Otherwise, the asserter retracts φ and accepts θ if she finds it believable. Then, in step 3, the asserter asserts a new formula ψ consistent with all irretractable assertions made by the receiver, and the subdialogue continues from the beginning. If the conflict-resolution subdialogue ends successfully, the enclosing dialogue can continue in the usual fashion, typically, by the receiver next either challenging or accepting the assertion.

In contrast to typical argumentation-based approaches, the goal in our dialogues is to find out *what* could be believed about the topic when the convictions of the agents are taken into account, not *whether* a particular proposition should be believed or not. The subject of the discussion is not fixed to one particular formula but is allowed to change within the topic. Moreover, our dialogues proceed more in the spirit of inquiry dialogues than persuasion dialogues (see Walton and Krabbe 1995). For example, van Veenen and Prakken (2006) included asking "Why did you refute my assertion?" among the moves in their negotiation protocol as an embedded persuasion game. However, their idea is to bring the grounds for the refutation 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 to both. Indeed, in the presence of a contradictory conviction, an attempt to persuade the other to give it up would turn out to be a futile exercise anyway.

The main difference between our approach and standard approaches to argumentation-based dialogues seems to come down to the use of the epistemic states of the agents in producing the assertions. Even approaches that consider the interplay between belief revision and argumentation, (see, e.g., Parsons and Sklar 2006), make several restricting assumptions: The agents have a common understanding of the degrees of beliefs, the subject of the conversation is fixed to the first assertion, and changes in beliefs take place only after the dialogue. In contrast, in our approach the agents have their private epistemic states and they do not communicate pre-calculated arguments as stored in their argument bases but instead formulate their assertions dynamically based on their beliefs and convictions: The receiver provides grounds for his rejection by calculating an interpolant from his convictions and the asserter formulates a new assertion based on her beliefs (at least hypothetically) revised by information provided by the receiver.

8 Conclusion

We studied situations in which an agent receives unbelievable information, that is, information that contradicts the agent's own convictions. We considered two cases depending on whether the receiver engages in a dialogue with the asserter of this information or not.

The receiver can still learn something from the unbelievable information even without such dialogue, because he can first revise the information with his own convictions to make it believable. The underlying idea of this accommodative revision is that the receiver tries to guess what the asserter would have said if the asserter had shared the same convictions as the receiver.

We also proposed a protocol framework for dialogues between the receiver and the asserter to determine cooperatively what they both could believe about the topic at hand. In such a dialogue, the asserter first tells what she believes about the topic, and then the receiver explains why his own convictions prevent him from believing it, which the asserter in turn takes into account in forming her next assertion. We showed how each agent can use logical interpolation in constructing the messages in such a dialogue; hence their strategies in constructing interpolants are parameters of our framework. Yet another such strategic parameter is how the asserter assesses the receiver's explanations: in ascending order of credulity, she can just assume them tentatively for the duration of the current dialogue, or she can believe them outright, or she can even be convinced by them. However, iterated belief revision assures that her messages are the same in each of these three strategies. A final question is how the asserter should react if she notices that she and the receiver hold conflicting convictions

about the topic. Accordingly we gave two variants for her part of the protocol: in one variant she terminates their dialogue in failure, whereas in the other variant she continues it but with a narrower subtopic which avoids this particular conflict.

Whatever the strategies of the agents are, our protocols ensure the following: First, dialogues always terminate. Second, if a dialogue terminates successfully, then it has produced a statement which both agents can believe, and which moreover the asserter considers most plausible given what she has heard from the receiver during their dialogue. And third, a dialogue may terminate in failure only if the convictions of the agents conflict with each other.

Acknowledgements

We wish to thank the anonymous referees for their diligent reading of our manuscript. Raul Hakli's work has been partially supported by the Academy of Finland.

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.

Albert Atserias, Phokion G. Kolaitis, and Moshe Y. Vardi. Constraint propagation as a proof system. In Mark Wallace, editor, *Principles and Practice of Constraint Programming – CP 2004*, volume 3258 of *Lecture Notes in Computer Science*, pages 77–91. Springer, 2004.

D. Bellot, C. Godefroid, P. Han, J. P. Prost, K. Schlechta, and E. Wurbel. A semantical approach to the concept of screened revision. *Theoria*, 63:24–33, 1997.

Béla Bollobás. *Combinatorics: Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability*. Cambridge University Press, 1986.

George S. Boolos and Richard C. Jeffrey. *Computability and Logic*. Cambridge University Press, 3rd edition, 1989.

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

E. M. Clarke, K. L. McMillan, X. Zhao, M. Fujita, and J. Yang. Spectral transforms for large boolean functions with applications to technology mapping. *Formal Methods in System Design*, 10(2-3):137–148, 1997. Special issue on Multi-Terminal Binary Decision Diagrams.

Eva Cogan, Simon Parsons, and Peter McBurney. New types of inter-agent dialogues. In Simon Parsons, Nicolas Maudet, Pavlos Moraitis, and Iyad Rahwan, editors, *Argumentation in Multi-Agent Systems Second International Workshop, ArgMAS 2005 Utrecht, The Netherlands, July 26, 2005 Revised Selected and Invited Papers*, volume 4049 of *Lecture Notes in Computer Science*, pages 154–168. Springer, 2006.

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

Rina Dechter. *Constraint Processing*. Morgan Kaufmann, 2003.

Vijay D'Silva. Propositional interpolation and abstract interpretation. In A. D. Gordon, editor, *19th European Symposium on Programming (ESOP)*, volume 6012 of *Lecture Notes in Computer Science*, pages 185–204. Springer-Verlag, 2010.

Vijay D'Silva, David Kroening, Mitra Purandare, and Georg Weissenbacher. Interpolant strength. In G. Barthe and M. Hermenegildo, editors, *Verification, Model Checking and Abstract Interpolation (VMCAI)*, volume 5944 of *Lecture Notes in Computer Science*, pages 129–145. Springer-Verlag, 2010.

Satu Eloranta. *Dynamic Aspects of Knowledge Bases*. PhD thesis, Department of Computer Science, University of Helsinki, Finland, June 2013.

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)*, volume 5293 of *Lecture Notes in Artificial Intelligence*, pages 180–191. Springer, 2008.

Eduardo L. Fermé and Sven Ove Hansson. Selective revision. *Studia Logica*, 63(3):331–342, 1999.

Michael C. Frank and Noah D. Goodman. Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75:80–96, 2014.

Peter Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics*, volume 3. Academic Press, 1975. Reprinted as Grice (1989, Chapter 2).

Paul Grice. Further notes on logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics*, volume 9. Academic Press, 1978. Reprinted as Grice (1989, Chapter 3).

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

Jeroen Groenendijk. The logic of interrogation: Classical version. In *Semantics and Linguistic Theory (SALT IX)*, pages 109–126. Cornell University, 1999.

Charles L. Hamblin. *Fallacies*. Methuen & Co, 1970.

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

Sven Ove Hansson. Semi-revision. *Journal of Applied Non-Classical Logics*, 7(1-2):151–175, 1997.

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

Sven Ove Hansson, Eduardo Leopoldo Fermé, John Cantwell, and Marcelo Alejandro Falappa. Credibility limited revision. *The Journal of Symbolic Logic*, 66(4):1581–1596, 2001.

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.

Guoxiang Huang. Constructing Craig interpolation formulas. In Ding-Zhu Du and Ming Li, editors, *First Annual International Conference on Computing and Combinatorics (COCOON '95)*, volume 959 of *Lecture Notes in Computer Science*, pages 181–190. Springer, 1995.

Gerhard Jäger. Game dynamics connects semantics and pragmatics. In Ahti-Veikko Pietarinen, editor, *Game Theory and Linguistic Meaning*, volume 18 of *Current Research in the Semantics/Pragmatics Interface*, pages 103–118. Emerald Group Publishing Limited, Bingley, UK, 2007.

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.

Hirofumi Katsuno and Alberto O. Mendelzon. Propositional knowledgebase revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1992.

Jürg Kohlas, Serafin Moral, and Rolf Haenni. Propositional information systems. *Journal of Logic and Computation*, 9(5):651–681, 1999.

Erik C.W. Krabbe. The problem of retraction in critical discussion. *Synthese*, 127:141–159, 2001.

Jan Krajíček. Interpolation theorems, lower bounds for proof systems, and independence results for bounded arithmetic. *The Journal of Symbolic Logic*, 62(2):457–486, 1997.

Loredana Laera, Ian Blacoe, Valentina Tamma, Terry Payne, Jerôme Euzenat, and Trevor Bench-Capon. Argumentation over ontology correspondences in MAS. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '07)*, pages 1285–1292, New York, NY, USA, 2007. ACM. ISBN 978-81-904262-7-5.

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

David K. Lewis. *Counterfactuals*. Blackwell, 1973.

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

David Makinson. Screened revision. *Theoria*, 63(1-2):14–23, 1997.

Kenneth L. McMillan. Interpolation and SAT-based model checking. In Warren A. Hunt, Jr. and Fabio Somenzi, editors, *Computer Aided Verification (CAV)*, volume 2725 of *Lecture Notes in Computer Science*, pages 1–13. Springer, 2003.

Daniele Mundici. Tautologies with a unique Craig interpolant, uniform vs. nonuniform complexity. *Annals of Pure and Applied Logic*, 27(3):265–273, 1984.

Matti Nykänen, Satu Eloranta, Olli Niinivaara, and Raul Hakli. Cooperative replies to unbelievable assertions: A dialogue protocol based on logical interpolation. In Joaquim Filipe and Ana Fred, editors, *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence (ICAART 2011)*, volume 2 – Agents, pages 245–250. SciTe Press, January 2011.

Simon Parsons and Elizabeth Sklar. How agents alter their beliefs after an argumentation-based dialogue. In Simon Parsons, Nicolas Maudet, Pavlos Moraitis, and Iyad Rahwan, editors, *Second International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2005). Revised Selected and Invited Papers*, volume 4049 of *Lecture Notes in Artificial Intelligence*, pages 297–312. Springer, 2006.

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.

Pavel Pudlák. Lower bounds for resolution and cutting plane proofs and monotone computations. *The Journal for Symbolic Logic*, 62(3):981–998, 1997.

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

Mark Snaith and Chris Reed. Justified argument revision in agent dialogue. In Peter McBurney, Simon Parsons, and Iyad Rahwan, editors, *Proceedings of the 9th International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2012)*, 2012.

Wolfgang Spohn. Ordinal conditional functions: a dynamic theory of epistemic state. In William L. Harper and Brian Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134. Kluwer Academic Publishers, 1988. Reprinted as Spohn (2009, Chapter 1).

Wolfgang Spohn. *Causation, Coherence and Concepts: A Collection of Essays*, volume 256 of *Boston Studies in the Philosophy of Science*. Springer, 2009.

Wolfgang Spohn. *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford University Press, 2012.

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

Jurriaan van Diggelen, Robbert-Jan Beun, Frank Dignum, Rogier M. van Eijk, and John-Jules Meyer. Ontology negotiation: goals, requirements and implementation. *International Journal of Agent-Oriented Software Engineering*, 1(1):63–90, 2007.

Jelle van Veenen and Henry Prakken. A protocol for arguing about rejections in negotiation. In Simon Parsons, Nicolas Maudet, Pavlos Moraitis, and Iyad Rahwan, editors, *Second International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2005). Revised Selected and Invited Papers*, volume 4049 of *Lecture Notes in Artificial Intelligence*, 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.

A On computing Craig interpolants

Recall the well-known propositional resolution inference rule

$$\frac{C \vee x \qquad \neg x \vee D}{C \vee D}\ (\text{Res})$$

where C and D are clauses and x is a propositional symbol called the *pivot*. Recall also that a resolution refutation proof is the derivation of the logical falsity $\bot$ using this rule. For such proofs, the Definition 1 of Craig interpolants given in section 3.2 assumes the form where $\alpha \models \neg\beta$ and $\theta \models \neg\beta$.

D'Silva et al. (2010) have investigated in detail extracting Craig interpolants from such proofs. Their overall algorithmic schema is given as Algorithm 5 below. It proceeds by recursion on the structure of the resolution proof $\mathscr{P}$ which is initially the refutation R. Each recursive call produces both an interpolant for its $\mathscr{P}$ and an updated labelling. These labellings are discussed next.

Algorithm 5 is a schema, because it uses an initial labelling ℓ for the propositional symbol occurrences in α and β, and different labellings yield different interpolant construction algorithms. A natural choice for this labelling is

$$\mathit{symmetric}(x) = \begin{cases} \{\mathtt{a}\} & \text{if this propositional symbol } x \text{ occurs only in } \alpha \text{ but not in } \beta \\ \{\mathtt{b}\} & \text{if this propositional symbol } x \text{ occurs only in } \beta \text{ but not in } \alpha \\ \{\mathtt{a},\mathtt{b}\} & \text{otherwise} \end{cases} \tag{3}$$

which also indicates the intended meanings of these labels: label `a` indicates that x belongs to α, while label `b` indicates that x belongs to β instead. The corresponding interpolant construction algorithm $\textsc{Craig}_{\mathit{symmetric}}$ is so natural that it had already been discovered thrice in the literature (Huang 1995; Krajíček 1997; Pudlák 1997).

D'Silva et al. (2010) provide the novel insight that in fact only the first two cases of Equation (3) are mandatory, whereas we can label freely those propositional symbols which occur in both α and β. We can for instance merge its third case into the second; this gives us the labelling

$$\mathit{strong}(x) = \begin{cases} \{\mathtt{a}\} & \text{if this propositional symbol } x \text{ occurs only in } \alpha \text{ but not in } \beta \\ \{\mathtt{b}\} & \text{otherwise} \end{cases} \tag{4}$$

Algorithm 5: A Craig interpolant extraction algorithm schema.

$\textsc{Craig}_\ell(R$: resolution refutation of $\alpha \wedge \beta)$: formula

1 **return** the first component of $\textsc{Craig}'_\ell(R)$.

$\textsc{Craig}'_\ell(\mathscr{P}$: resolution proof$)$: $\langle$formula, labelling$\rangle$

1 **if** the conclusion Q of $\mathscr{P}$ is a clause of α
2 **return** $\langle Q \lfloor \mathtt{b}, \ell \rangle$ where $Q \lfloor \mathtt{b}$ denotes the clause consisting of those literals of Q whose propositional symbol y receives the label $\ell(y) = \{\mathtt{b}\}$ in this initial labelling ℓ
3 **elseif** Q is a clause of β
4 **return** $\langle \neg(Q \lfloor \mathtt{a}), \ell \rangle$
5 **else** $\mathscr{P}$ ends in some (Res) step with pivot x;
6 $\langle \theta_C, \ell_C \rangle = \textsc{Craig}'_\ell$(subproof for its left antecedent $C \vee x$);
7 $\langle \theta_D, \ell_D \rangle = \textsc{Craig}'_\ell$(subproof for its right antecedent $\neg x \vee D$);
8 $\ell_{C \vee D} =$ the pointwise union of these two labellings ℓ_C and ℓ_D (that is, $\ell_{C \vee D}(z) = \ell_C(z) \cup \ell_D(z)$ for every z);
9 Construct an interpolant from these two interpolants θ_C and θ_D based on the label given to the pivot x by this combined labelling as follows:

$$\theta_{C \vee D} = \begin{cases} \theta_C \vee \theta_D & \text{if } \ell_{C \vee D}(x) = \{\mathtt{a}\} \\ \theta_C \wedge \theta_D & \text{if } \ell_{C \vee D}(x) = \{\mathtt{b}\} \\ (\theta_C \vee x) \wedge (\neg x \vee \theta_D) & \text{if } \ell_{C \vee D}(x) = \{\mathtt{a}, \mathtt{b}\}; \end{cases}$$

10 $\ell'_{C \vee D} = \ell_{C \vee D}$ except that x maps to $\emptyset$;
// This accounts for the disappearance of this x from the resolvent $C \vee D$.
11 **return** $\langle \theta_{C \vee D}, \ell'_{C \vee D} \rangle$.

instead. The corresponding algorithm $\textsc{Craig}_{strong}$ had been discovered earlier by McMillan (2003). Hence this algorithm constructs the interpolant from the pertinent parts of the clauses of α. This preference for α leads into interpolants which are the logically strongest obtainable via this scheme:

$$\textsc{Craig}_{strong}(R) \models \textsc{Craig}_\ell(R) \tag{5}$$

for all refutations R and labellings ℓ satisfying the first two mandatory cases of Equation (3) (D'Silva et al. 2010, Section 4.2). Or conversely, we could define the *weak* labelling which prefers β instead, which in turn yields logically weakest interpolants in this sense.

Instead of logical strength, we may want to optimize the propositional symbols used in the interpolants (D'Silva 2010, Section 5.2). The labelling

$$small(x) = \begin{cases} \{\mathtt{a}\} & \text{if this particular occurrence of } x \text{ is in } \alpha \\ \{\mathtt{b}\} & \text{otherwise} \end{cases} \tag{6}$$

attains this goal: $\mathrm{voc}(\textsc{Craig}_{small}(R)) \subseteq \mathrm{voc}(\textsc{Craig}_\ell(R))$ for all R and ℓ as in Equation (5) above. This *small* labelling is in turn the same one that is used in Craig interpolation proofs for sequent calculi (albeit implicitly, via the so-called shared sequents, as in for instance Troelstra and Schwichtenberg 2000, Chapter 4.4.2) and in the interpolant generation algorithms implicit in them. In contrast to the earlier labellings, it concerns particular occurrences of symbols instead of symbols themselves.

Krajíček (1997) and Pudlák (1997) have developed explicit pathological formulas α' and β' such that the output γ' of $\textsc{Craig}_{symmetric}$ on the corresponding resolution refutation of $\alpha' \wedge \beta'$ is exponentially longer than α' and β'. However, their main motivation has been to study the inherent length of resolution proofs but not the complexity of interpolation in general or $\textsc{Craig}_{symmetric}$ in particular. CRAIG namely constructs its output efficiently with respect to its input, and therefore a long output means that its input refutation must also have been long.

Such pathological formulas for CRAIG do leave open the general question whether or not some other interpolant construction method would be polynomial in the lengths of its input formulas. The general outlook is bleak, since Mundici (1984, Section 3) has shown that if there always exists a Craig interpolant γ of polynomial length with respect to the lengths of the input formulas α and β, then all languages in $\mathbf{NP} \cap \mathbf{coNP}$ have polynomial-size circuits which is considered unlikely. Lang et al. (2003, Proposition 23) have in turn extended this result to literal forgetting which is related to uniform interpolation.

B On computing uniform interpolants

The semantic reformulation of Definition 2 leads into the following brute-force construction of uniform interpolants:

$$\pi_T \Phi = \bigvee \{c \in D\colon\ \Phi \text{ is consistent with } c\}$$

where each conjunction c of literals l_x is built by taking each propositional symbol $x \in T$ exactly once:

$$D = \left\{ \bigwedge_{x \in T} l_x \colon l_x \in \{x, \neg x\} \right\}.$$

Note that this brute force construction extends from formulas into theories, because the finiteness of T ensures the finiteness of D.

This reformulation also shows that uniform interpolant construction can be viewed as a Boolean constraint satisfaction problem, where one must determine those value combinations for the propositional symbols in T which can be extended into full solutions of all the constraints expressed as Φ (for further information on Boolean constraint satisfaction, see Dechter 2003, Chapter 8.4). Such Boolean constraint satisfaction problems can in turn be handled with the resolution rule given in Appendix A (Atserias et al. 2004). This improves the brute force construction above into

1. first convert Φ into Conjunctive Normal Form (CNF) if necessary,
2. then compute the closure of Φ with respect to all the (Res) steps where the pivot is not in T, and
3. finally remove from the result every clause which contains a propositional symbol not in T

Kohlas et al. (1999) build on this idea and go on to develop methods for decomposing the input formula and choosing the order in which these pivots are handled to reduce the overall computational effort.

However, the most promising computational approach in our setting seems to be representing the epistemic state of an agent as a ranking function (Spohn 2012, Definition 5.5) and representing it as a Multi-Terminal (Reduced Ordered) Binary Decision Diagram (Clarke et al. 1997, section 3.2). This data structure namely possesses straightforward algorithms for many of the the required operations, including uniform interpolation viewed as projection. We leave developing this approach to further work.