Reconstructing an Agent's Epistemic State from Observations about its Beliefs and Non-beliefs

RICHARD BOOTH, *Mahasarakham University, Faculty of Informatics, Mahasarakham 44150, Thailand.*
E-mail: richard.b@msu.ac.th

ALEXANDER NITTKA, *University of Leipzig, Department of Computer Science, Leipzig 04103, Germany.*
E-mail: nittka@informatik.uni-leipzig.de

Abstract

We look at the problem in belief revision of trying to make inferences about what an agent believed—or *will* believe—at a given moment, based on an observation of how the agent has responded to some sequence of previous belief revision inputs over time. We adopt a 'reverse engineering' approach to this problem. Assuming a framework for iterated belief revision which is based on sequences, we construct a model of the agent that 'best explains' the observation. Further considerations on this best-explaining model then allow inferences about the agent's epistemic behaviour to be made. We also provide an algorithm which computes this best explanation.

Keywords: Belief revision, non-monotonic reasoning, iterated revision, non-prioritised revision, rational closure, rational explanation, multi-agent systems.

1 Introduction

The problem of belief revision, i.e. of how an agent should modify its (given) initial epistemic state when encountering some new information which possibly contradicts its current beliefs about the world, is by now a well-established research area in AI [1, 11]. Traditionally, the work in this area is done from the *agent's perspective*, being usually pre-occupied with constructing actual revision operators which the agent might use and with rationality postulates which constrain how these operators should behave. We want to change viewpoint and instead cast ourselves in the role of an *observer* of the agent. Imagine the following scenario. We are given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ of a particular agent, hereafter $\mathcal{A}$, covering a certain length of time. The φ_i and θ_i are sentences, the D_i finite sets of sentences. The interpretation of o is that after having received the revision inputs $\varphi_1, \ldots, \varphi_i$, $\mathcal{A}$ believed (at least) θ_i and did not believe any $\delta \in D_i$.[1] In contrast to the traditional work in belief revision, we are *not* given $\mathcal{A}$'s initial state and only a very partial description of the epistemic states after the revisions have taken place. A couple of questions

[1] This article extends work presented in [4] where only information about the beliefs of an agent after receiving a series of revision inputs was considered, i.e. there $\forall i : D_i = \emptyset$ was assumed. The current article generalizes the results also taking information about non-beliefs into account. Notation and definitions have been adapted accordingly.

Vol. 18 No. 5, © The Author, 2008. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org
Published online 15 February 2008 doi:10.1093/logcom/exm091

now suggest themselves:

- What did $\mathcal{A}$ believe immediately *before* it received the first input φ_1?
- What did $\mathcal{A}$ believe after the i^{th} input φ_i—*apart* from θ_i?
- What will $\mathcal{A}$ believe following a *further* revision input φ_{n+1}?

One area in which such questions might arise is in human–machine dialogues [9]. Here, it can be useful for the machine to keep a model of the evolution of a user's beliefs during a dialogue. The φ_i correspond to inputs given to $\mathcal{A}$—the user—by a machine, and the θ_i and D_i are the user's responses or are somehow elicited from the user's responses. For example, if $\mathcal{A}$ responds to φ_i with a question 'Is δ true or not?' then (assuming the user's sincerity) we might infer the user is ignorant about the truth of δ, i.e. δ and $\neg\delta$ are elements of D_i. More generally, the problem of drawing conclusions about the beliefs of other agents is interesting in any multi-agent setting. Successful interaction with others can be supported by understanding them—not only in the present situation but also retrospectively. Parts of the information about the other agent are represented in o which may have been obtained by direct inquiry, observation or through other sources. Our aim in this article is to answer the questions posed above.

Our strategy for dealing with them will be to adopt a 'reverse engineering' approach—constructing a model of the agent. (Similar approaches have already been tried in the context of trying to infer an agent's *goals* from observable *actions*, e.g. [6, 17].) Having no access to the agent's internals, we assume a belief revision framework $\mathcal{A}$ uses for determining its beliefs and for incorporating new information, and construct a model of $\mathcal{A}$ that explains the observation about it. By considering this model, we will then be able to make extra inferences or predictions about $\mathcal{A}$'s epistemic behaviour. Of course, this raises the problem of which belief revision framework to choose. Such a framework will obviously need to support *iterated* revision [3, 7, 19, 23], and preferably also *non-prioritized* revision [15, 22], i.e. revision in which new inputs are allowed to be *rejected*. In this article, we restrict the investigation to one such framework that has been studied in [2, Section 6]. The idea behind it is that an agent's epistemic state $[\rho, \blacktriangle]$ is made up of *two* components: (i) a sequence ρ of sentences representing the sequence of revision inputs the agent has received thus far, and (ii) a single sentence $\blacktriangle$ standing for the agent's set of *core* beliefs, which intuitively are those beliefs of the agent it considers 'untouchable'. The agent's full set of beliefs in the epistemic state $[\rho, \blacktriangle]$ is then determined by a particular calculation on ρ and $\blacktriangle$, while new revision inputs are incorporated by simply appending them to the end of ρ. Note that our choice of this framework does not imply that others are less worthy of investigation. The challenge now becomes to find that *particular* model of this form which *best explains* the observation we have made of $\mathcal{A}$.

The plan of the article is as follows. In Section 2, we describe in more detail the model of epistemic state we will be assuming. This will enable us to pose more precisely the problem we want to solve. We will see that the problem essentially reduces to trying to guess what $\mathcal{A}$'s *initial* epistemic state $[\rho, \blacktriangle]$ (i.e. before it received φ_1) was. In Section 3, inspired by work done on reasoning with *conditional beliefs*, we propose a way of finding the best initial sequence—or *prefix*—$\rho(\blacktriangle)$ for any given *fixed* $\blacktriangle$. Then, in Section 4 we focus on finding the best $\blacktriangle$. This will amount to equating best with logically weakest. The epistemic state $[\rho(\blacktriangle), \blacktriangle]$ obtained by combining our answers will be our proposed best explanation for o, which we will call the *rational explanation*. In Section 5, we present an algorithm which *constructs* the rational explanation for any given o, before giving some examples to show the type of inferences this explanation leads to in Section 6. In Section 7, we discuss two papers with similar topics before concluding and giving some pointers for future research in Section 8. An appendix contains all proofs.

2 Modelling the agent

We assume sentences $\alpha, \beta, \gamma, \varphi, \theta, \delta, \blacktriangle$, etc. are elements of some finitely-generated propositional language L. In our examples, p, q, r, etc. denote distinct propositional variables. The classical logical entailment relation is denoted by $\vdash$, while $\equiv$ denotes classical logical equivalence. Wherever we use a sentence to describe a *belief set* the intention is that it represents all its logical consequences. The operation $\cdot$ denotes concatenation, so $\sigma \cdot \rho$ is the concatenation of two sequences of sentences and $\sigma \cdot \varphi$ is the result of appending the sentence φ to σ.

Before we turn to the representation of the agent $\mathcal{A}$ itself, we want to introduce a function which we will need later. f takes as argument a non-empty sequence $\sigma = (\alpha_m, \ldots, \alpha_1)$ of sentences and returns a sentence.

$$f(\alpha_m, \ldots, \alpha_1) = \begin{cases} \alpha_1 & \text{if } m = 1 \\ \alpha_m \wedge f(\alpha_{m-1}, \ldots, \alpha_1) & \text{if } m > 1 \text{ and } \alpha_m \wedge f(\alpha_{m-1}, \ldots, \alpha_1) \nvdash \bot \\ f(\alpha_{m-1}, \ldots, \alpha_1) & \text{otherwise} \end{cases}$$

$f(\sigma)$ is determined by first taking α_1 and then going backwards through σ, adding each sentence as we go, provided that sentence is consistent with what has been collected so far (cf. the 'linear base-revision operation' of [24] and the 'basic memory operator' of [16]). Note in particular that the calculation does not stop at the first sentence that would cause an inconsistency (cf. 'cut base-revision' of [24]). It is simply left out and the next one is considered. Some useful properties of f are listed in the appendix. The next definition summarizes the revision framework we assume the observed agent $\mathcal{A}$ to employ.

DEFINITION 2.1
The *epistemic state* $[\rho, \blacktriangle]$ of an agent consists of a sequence of sentences ρ and a sentence $\blacktriangle$ which is called *core belief*. The *set of beliefs* of an agent in the epistemic state $[\rho, \blacktriangle]$ is $Bel([\rho, \blacktriangle]) = f(\rho \cdot \blacktriangle)$. The epistemic state resulting from revising $[\rho, \blacktriangle]$ by a sentence λ is $[\rho, \blacktriangle] * \lambda = [\rho \cdot \lambda, \blacktriangle]$.

As indicated in the introduction, this definition follows [2] for representing an agent. Refs [16, 19] also use sequences to represent epistemic states, but without core beliefs. We will refer to the elements of the agent's belief set $Bel([\rho, \blacktriangle])$ as beliefs and to sentences not contained in the belief set as non-beliefs. When calculating its beliefs from its epistemic state $[\rho, \blacktriangle]$, an agent gives highest priority to $\blacktriangle$. After that, it prioritizes more recent information received, ignoring sentences that would cause an inconsistency. Note that $\blacktriangle$ is always believed, and that $Bel([\rho, \blacktriangle])$ is inconsistent if and only if $\blacktriangle$ is inconsistent. Revision is done by appending the input to ρ. This makes the framework a slight variation of linear base-revision [24]. Instead of appending the input to the sequence of sentences as is done there, here the input is inserted in the last but one position. The core belief is always considered more important and could thus also be interpreted as the very last input received.

EXAMPLE 2.2
Consider $\blacktriangle = \neg p$ and $\rho = (q, q \to p, p \wedge r)$. $Bel([\rho, \blacktriangle]) = f(q, q \to p, p \wedge r, \neg p)$. In order to determine $f(q, q \to p, p \wedge r, \neg p)$ we need to know if q is consistent with $f(q \to p, p \wedge r, \neg p)$. As $f(\neg p) = \neg p$ is inconsistent with $p \wedge r$, the latter sentence is ignored and $f(p \wedge r, \neg p) = \neg p$. $q \to p$ is consistent with $\neg p$ and so $f(q \to p, p \wedge q, \neg p) = (q \to p) \wedge \neg p \equiv \neg q \wedge \neg p$. So q is inconsistent with $f(q \to p, p \wedge r, \neg p)$ and we get $Bel([\rho, \blacktriangle]) \equiv \neg q \wedge \neg p$.

Core beliefs are considered in order to allow for non-prioritized revision. A new input λ will not always be believed in the new state. Indeed (when $\blacktriangle$ is consistent) λ will be believed only if it is consistent with $\blacktriangle$. If it contradicts $\blacktriangle$ then it still takes its place in the epistemic state

but it will not be believed, and in fact in this case the agent's belief set will remain unchanged (cf. *screened* revision [22]). Note also that ▲ remains unaffected by a revision input, i.e. $*$ is a *core-invariant* revision operator [2].[2]

As is shown in [2], the above revision method satisfies several natural properties. In particular, it stays largely faithful to the AGM postulates [11] (leaving aside the 'success' postulate, which forces all new inputs to be believed), and satisfies slight, 'non-prioritized' variants of several postulates for iterated revision which have been proposed, including those of [7]. One characteristic property of this method is the following variant of the rule 'Recalcitrance' from [23]:

$$\text{If } \blacktriangle \nvdash (\lambda_2 \rightarrow \neg\lambda_1) \text{ then } Bel([\rho, \blacktriangle] * \lambda_1 * \lambda_2) \vdash \lambda_1$$

This entails if the agent *believes* an input λ_1, then it does so *wholeheartedly*, in that the only way it can be dislodged from the belief set by a succeeding input λ_2 is if that input contradicts it given the core beliefs ▲. We will now turn to the observation we can make of an agent.

DEFINITION 2.3
An *observation* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ is a sequence of triples $(\varphi_i, \theta_i, D_i)$ where φ_i and θ_i are sentences and D_i is a finite sets of sentences, $1 \leq i \leq n$. The set of all possible observations (for all $n \geq 0$) is denoted by O.

The interpretation of an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ is that after the agent $\mathcal{A}$ has received the inputs $\varphi_1, \ldots, \varphi_i$ in an initial epistemic state $[\rho, \blacktriangle]$, it believes at least θ_i and believes no element of D_i. Throughout this article, we make the assumptions that $\mathcal{A}$ received no input between φ_1 and φ_n other than those listed, and that the θ_i and $\delta \in D_i$ are *correct* descriptions of $\mathcal{A}$'s beliefs and non-beliefs after each input. That is, o provides (partial) information about the beliefs and non-beliefs of an agent through a sequence of revisions. After receiving the i-th input φ_i, $\mathcal{A}$'s epistemic state must be $[\rho \cdot (\varphi_1, \ldots, \varphi_i), \blacktriangle]$ and its belief set $Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))$. $[\rho, \blacktriangle]$ is $\mathcal{A}$'s unknown *initial* (i.e. before φ_1) epistemic state. The intuitive reading of an observation is captured formally by

$$\begin{aligned} &\text{for all } i \text{ such that } 1 \leq i \leq n: \\ &Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \vdash \theta_i \quad \text{and} \\ &\forall \delta \in D_i : Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \nvdash \delta. \end{aligned} \tag{1}$$

The following definition formalizes when we consider an epistemic state to explain an observation.

DEFINITION 2.4
Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle \in O$. Then $[\rho, \blacktriangle]$ *explains* o (or is *an explanation for* o) if and only if $\blacktriangle \not\equiv \bot$ and (1) above holds. We say ▲ is an $o-$*acceptable* core if and only if $[\rho, \blacktriangle]$ explains o for *some* ρ.

Note that in our agent model we do not prohibit the core belief ▲ to be inconsistent, in which case $\mathcal{A}$ would believe everything at any point in time. $\blacktriangle \equiv \bot$ satisfies (1) using any sequence ρ if and only if $D_i = \emptyset$ for all i. As mentioned before, in [4] we considered observations without information about the non-beliefs which amounts to precisely this condition. As a consequence, there $\bot$ was considered o-acceptable for any observation. There are technical reasons for eliminating the possibility of $\bot$ being o-acceptable, but there is also an intuitive argument. We believe that it is better to say that we do not have an explanation rather than claiming the agent to be inconsistent.

[2] In fact the model of [2] allows the core itself to be revisable. We do not explore this possibility here.

EXAMPLE 2.5
(i) $[\rho, \blacktriangle] = [(p \to q), r]$ explains $\langle (p,q,\emptyset),(q,r,\emptyset)\rangle$ because $f(p \to q, p, r)$ entails q and $f(p \to q, p, q, r)$ entails r (both are equivalent to $p \wedge q \wedge r$).
(ii) $[(p \to q), \top]$ does not explain $\langle (p,q,\emptyset),(q,r,\emptyset)\rangle$ because $f(p \to q, p, q, \top) \equiv p \wedge q \nvdash r$.
(iii) $[(p \to q), \top]$ does not explain $\langle (p, \top, \{q\})\rangle$ because $f(p \to q, p, \top) \equiv p \wedge q \vdash q$.

If we had some explanation $[\rho, \blacktriangle]$ for o then we would be able to answer the questions in the introduction: following a new input φ_{n+1} $\mathcal{A}$ will believe $f(\rho \cdot (\varphi_1, \ldots, \varphi_n, \varphi_{n+1}, \blacktriangle))$, before receiving the first input $\mathcal{A}$ believes $f(\rho \cdot \blacktriangle)$, and the beliefs after the i-th input are $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))$. However, not all observations have an explanation, $\langle (p, \top, \{p, \neg p\})\rangle$ probably being the simplest example. One property of the assumed belief revision framework is that for every input φ received by an agent either φ or $\neg\varphi$ is believed henceforth, but the observation would require the violation of that property.

Our job now is to choose, from the space of possible explanations for o, the best one. As a guideline, we consider an explanation good if it only makes necessary (or minimal) assumptions about what $\mathcal{A}$ believes. But how do we find this best one? Our strategy is to split the problem into two parts, handling ρ and $\blacktriangle$ separately. First, (i) given a *fixed* o-acceptable core $\blacktriangle$, find a best sequence $\rho(o, \blacktriangle)$ such that $[\rho, \blacktriangle]$ explains o, then, (ii) find a best o-acceptable core $\blacktriangle(o)$. Our best explanation for o will then be $[\rho(o, \blacktriangle(o)), \blacktriangle(o)]$.

3 Finding ρ

Given $o = \langle (\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$, let us assume a fixed core $\blacktriangle$. To find that sequence $\rho(o, \blacktriangle)$ such that $[\rho(o, \blacktriangle), \blacktriangle]$ is the best explanation for o, *given* $\blacktriangle$, we will take inspiration from work done in the area of non-monotonic reasoning on reasoning with *conditional* information.

Let's say a pair (λ, χ) of sentences is a *conditional belief* in the state $[\rho, \blacktriangle]$ if and only if χ would be believed after revising $[\rho, \blacktriangle]$ by λ, i.e. $Bel([\rho, \blacktriangle] * \lambda) \vdash \chi$. In this case, we will write $\lambda \Rightarrow_{[\rho, \blacktriangle]} \chi$.[3] This relation plays an important role, because it turns out $\mathcal{A}$'s beliefs following *any* sequence of revision inputs starting from $[\rho, \blacktriangle]$ is determined *entirely* by the set $\Rightarrow_{[\rho, \blacktriangle]}$ of conditional beliefs in $[\rho, \blacktriangle]$. This is because, using the definitions for calculating the belief set, revision and Proposition A.1 (i) from the appendix, we can show that, for *any* sequence of revision inputs $\varphi_1, \ldots, \varphi_i$, our revision method satisfies

$$Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = Bel([\rho, \blacktriangle] * f(\varphi_1, \ldots, \varphi_i, \blacktriangle)).$$[4]

Thus, as far as their effects on the belief set go, a sequence of revision inputs starting from $[\rho, \blacktriangle]$ can always be reduced to a single input. All this means observation o may be translated into a partial description of the set of conditional beliefs that $\mathcal{A}$ has in its initial epistemic state and the set of conditional beliefs that $\mathcal{A}$ does *not* have in its initial epistemic state. $\mathcal{C}_{\blacktriangle}(o)$ contains the known positive conditionals, $\mathcal{N}_{\blacktriangle}(o)$ the known negative ones:

$$\mathcal{C}_{\blacktriangle}(o) = \{f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \Rightarrow \theta_i \mid i = 1, \ldots, n\},$$

$$\mathcal{N}_{\blacktriangle}(o) = \{f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \Rightarrow \delta \mid i = 1, \ldots, n \wedge \delta \in D_i\}.$$

[3] The relation $\Rightarrow_{[\rho, \blacktriangle]}$ *almost* satisfies all the rules of a rational inference relation [20]. More precisely the modified version does, viz. $\lambda \Rightarrow'_{[\rho, \blacktriangle]} \chi$ if and only if $[\blacktriangle \vdash \neg\lambda$ or $\lambda \Rightarrow_{[\rho, \blacktriangle]} \chi]$.

[4] The key to this result is that $f(f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle) \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$. This is because $f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \vdash \blacktriangle$ and hence $f(f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle) = f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \wedge \blacktriangle \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$.

We want to stress that the set of conditional beliefs $\Rightarrow_{[\rho,\blacktriangle]*\lambda}$ in the state $[\rho,\blacktriangle]*\lambda$ following revision by λ will generally *not* be the same as $\Rightarrow_{[\rho,\blacktriangle]}$. For this reason, we will always consider the conditional beliefs with respect to the unknown initial state $[\rho,\blacktriangle]$. That way, we get much information concerning one relation rather than little information concerning many different relations. As we will deal with conditionals talking about the same state, we will omit the subscript for readability and write $\Rightarrow$ instead of $\Rightarrow_{[\rho,\blacktriangle]}$.

Clearly, if we had access to the *complete* set of $\mathcal{A}$'s conditional beliefs in its initial state, this would give another way to answer the questions of the introduction. Now, the problem of determining which conditional beliefs *follow from* a given set $\mathcal{C}$ of positive conditional beliefs has been well-studied and several solutions have been proposed, e.g. [12, 18]. One particularly elegant and well-motivated solution is to take the *rational closure* of $\mathcal{C}$ [20]. Furthermore, as is shown in, e.g. [10], this construction is amenable to a relatively simple representation as a sequence of sentences! As the original rational closure construction is not capable of dealing with negative information, we will use the generalized version that was proposed in [5] which again gives rise to a sequence of sentences. Our idea is essentially to take $\rho(o,\blacktriangle)$ to be this sequence corresponding to the rational closure of $\mathcal{C}_{\blacktriangle}(o)$ and $\mathcal{N}_{\blacktriangle}(o)$. First let us describe the general construction.

3.1 The rational closure of a set of (positive and negative) conditionals

We will now briefly recall the rational closure construction from [5]. It works with arbitrary sets of positive and negative conditionals, not just those calculated as $\mathcal{C}_{\blacktriangle}(o)$ and $\mathcal{N}_{\blacktriangle}(o)$ are. Given a set of conditionals $\mathcal{C}=\{\lambda_i\Rightarrow\chi_i \mid i=1,\ldots,l\}$ we denote by $\tilde{\mathcal{C}}$ the set of *material counterparts* of all the conditionals in $\mathcal{C}$, i.e. $\tilde{\mathcal{C}}=\{\lambda_i\rightarrow\chi_i \mid i=1,\ldots,l\}$. A conditional $\lambda\Rightarrow\chi$ is *p-exceptional* for a set of sentences U if and only if $U\vdash\neg\lambda$. $\lambda\Rightarrow\chi$ is *n-exceptional* for U if and only if $U\cup\{\lambda\}\vdash\chi$.

Now assume we are given a set $\mathcal{C}$ of positive conditionals and a set $\mathcal{N}$ of negative ones. The rational closure $\rho_R(\mathcal{C},\mathcal{N})$ of $\mathcal{C}$ and $\mathcal{N}$ is determined as follows. We define two decreasing sets of conditionals $\mathcal{C}_0\supseteq\mathcal{C}_1\supseteq\cdots\supseteq\mathcal{C}_m$ and $\mathcal{N}_0\supseteq\mathcal{N}_1\supseteq\cdots\supseteq\mathcal{N}_m$ and a decreasing set of sentences $U_0\supseteq U_1\supseteq\cdots\supseteq U_m$—the U_i will be defined via a least fixpoint construction.

DEFINITION 3.1
Let $\mathcal{C}$ be a set of positive conditionals and $\mathcal{N}$ a set of negative ones. Then the (sequence corresponding to the) *rational closure* $\rho_R(\mathcal{C},\mathcal{N})$ of $\mathcal{C}$ and $\mathcal{N}$ is $\rho_R(\mathcal{C},\mathcal{N})=(\bigwedge U_m,\bigwedge U_{m-1},\ldots,\bigwedge U_0)$ where

1. $\mathcal{C}_0=\mathcal{C}$ and $\mathcal{N}_0=\mathcal{N}$
2. U_i is the smallest set which contains $\tilde{\mathcal{C}}_i$ and which is closed under the following condition
   If $\lambda\Rightarrow\chi$ is in $\mathcal{N}_i$ and $\lambda\Rightarrow\chi$ is n-exceptional for U_i then $\neg\lambda\in U_i$.
3. $\mathcal{C}_{i+1}$ is the set of conditionals in $\mathcal{C}_i$ that are p-exceptional for U_i and
   $\mathcal{N}_{i+1}$ is the set of conditionals in $\mathcal{N}_i$ that are n-exceptional for U_i
4. m is minimal such that $\mathcal{C}_m=\mathcal{C}_{m+1}$ and $\mathcal{N}_m=\mathcal{N}_{m+1}$

This definition contains a reformulation of the rational closure construction from [5]. Step 2 above may be explained as follows. U_i is initialized with $\tilde{\mathcal{C}}_i$. Then we go through all the negative conditionals in $\mathcal{N}_i$. If there is a conditional $\lambda\Rightarrow\chi$ that is n-exceptional for U_i, which means that adding λ to U_i would lead χ to become inferable, its negated antecedent $\neg\lambda$ is added to U_i. The addition of these $\neg\lambda$ may lead *other* negative conditionals in $\mathcal{N}_i$ to become n-exceptional, so we then need to check $\mathcal{N}_i$ for conditionals that are n-exceptional for the set thus obtained. This process stops if no further sentence had to be added.

If $\mathcal{N}=\emptyset$ in the above process then the process simplifies to the one given in, e.g. [4, 10] which handles the case of positive conditionals only. The idea of the construction is that the U_i ensure that positive conditionals $\lambda \Rightarrow \chi$ are satisfied, i.e. χ is believed after revision by λ, while negative conditionals $\lambda \Rightarrow \chi'$ are not, i.e. here χ' is *not* believed after the revision. This is achieved by U_i containing $\lambda \rightarrow \chi$ making χ inferable if λ is added (resp. containing $\neg\lambda$ in case adding λ makes χ' inferable). Different U_i take care of different positive conditionals and if a positive conditional is p-exceptional for U_i we know that it cannot be satisfied using U_i.

Writing α_i for $\bigwedge U_i$, the rational closure of $\mathcal{C}$ and $\mathcal{N}$ is then the relation $\Rightarrow_R$ given by $\lambda \Rightarrow_R \chi$ if and only if either $\alpha_m \vdash \neg\lambda$ or $\left[\alpha_j \wedge \lambda \vdash \chi \text{ where } j \text{ is minimal such that } \alpha_j \nvdash \neg\lambda\right]$. Since $\alpha_0 \vdash \cdots \vdash \alpha_m$ it easy to check that in fact this second disjunct is equivalent to $f(\alpha_m, \ldots, \alpha_0, \lambda) \vdash \chi$. As is shown in [5], $\Rightarrow_R$ 'satisfies' all the conditionals in $\mathcal{C}$ and $\mathcal{N}$, in the sense that $\lambda \Rightarrow_R \chi$ for all positive conditionals $\lambda \Rightarrow \chi$ in $\mathcal{C}$, while $\lambda \not\Rightarrow_R \chi$ for all negative conditionals $\lambda \Rightarrow \chi$ in $\mathcal{N}$.

We now make the following definition:

DEFINITION 3.2
Let $o \in O$ and $\blacktriangle \in L$. We call $\rho_R(\mathcal{C}_\blacktriangle(o), \mathcal{N}_\blacktriangle(o))$ the *rational prefix* of o with respect to $\blacktriangle$, and will denote it by $\rho_R(o, \blacktriangle)$.

EXAMPLE 3.3
(i) Let $o = \langle (p, s, \emptyset), (q, \top, \{s\}), (r, \neg q, \emptyset)\rangle$ and $\blacktriangle = \top$. o says that after receiving p the agent believed s, but after then receiving q does not believe s anymore. Finally, after receiving r it believes $\neg q$. This translates into:

$$\begin{aligned} \mathcal{C}_0 = \mathcal{C}_\blacktriangle(o) &= \{f(p,\top) \Rightarrow s, f(p,q,\top) \Rightarrow \top, f(p,q,r,\top) \Rightarrow \neg q\} \\ &= \{p \Rightarrow s, p \wedge q \Rightarrow \top, p \wedge q \wedge r \Rightarrow \neg q\} \\ \mathcal{N}_0 = \mathcal{N}_\blacktriangle(o) &= \{f(p,q,\top) \Rightarrow s\} \\ &= \{p \wedge q \Rightarrow s\} \end{aligned}$$

So, $\tilde{\mathcal{C}}_0 = \{p \rightarrow s, p \wedge q \rightarrow \top, p \wedge q \wedge r \rightarrow \neg q\}$ and $p \wedge q \Rightarrow s$ is n-exceptional for that set as $\{p \rightarrow s, p \wedge q\} \vdash s$. Hence $U_0 = \{p \rightarrow s, p \wedge q \rightarrow \top, p \wedge q \wedge r \rightarrow \neg q, \neg(p \wedge q)\}$.

Of the positive conditionals only $p \Rightarrow s$ is not p-exceptional for U_0, as $\neg(p \wedge q) \vdash \neg(p \wedge q)$ and $p \wedge q \wedge r \rightarrow \neg q \vdash \neg(p \wedge q \wedge r)$. $p \wedge q \Rightarrow s$ is n-exceptional for U_0 as $\{p \rightarrow s, p \wedge q\} \vdash s$, so $\mathcal{C}_1 = \{p \wedge q \Rightarrow \top, p \wedge q \wedge r \Rightarrow \neg q\}$ and $\mathcal{N}_1 = \{p \wedge q \Rightarrow s\}$.

This time $U_1 = \tilde{\mathcal{C}}_1 = \{p \wedge q \rightarrow \top, p \wedge q \wedge r \rightarrow \neg q\}$ as adding $p \wedge q$ does not make s inferable, anymore. Only $p \wedge q \wedge r \Rightarrow \neg q$ is exceptional for U_1. As indicated above, $p \wedge q \wedge r \Rightarrow \neg q$ is in a sense exceptional for *itself* because $p \wedge q \wedge r$ is inconsistent with $p \wedge q \wedge r \rightarrow \neg q$. So we have $\mathcal{C}_2 = \{p \wedge q \wedge r \Rightarrow \neg q\} = \mathcal{C}_3, \mathcal{N}_2 = \emptyset = \mathcal{N}_3$ and $U_2 = \{p \wedge q \wedge r \rightarrow \neg q\} = U_3$. Here, the calculation stops, as the sets do not change and (omitting conditionals whose material counterparts are tautologies) we get $\rho_R(o, \blacktriangle) = (\bigwedge U_2, \bigwedge U_1, \bigwedge U_0) =$

$$(p \wedge q \wedge r \rightarrow \neg q \ , \ p \wedge q \wedge r \rightarrow \neg q \ , \ (p \rightarrow s) \wedge (p \wedge q \wedge r \rightarrow \neg q) \wedge \neg(p \wedge q)).$$

Using logical equivalences this is the same as $(p \wedge q \rightarrow \neg r \ , \ p \wedge q \rightarrow \neg r \ , \ p \rightarrow (s \wedge \neg q))$.

(ii) For any observation o, if $\blacktriangle \equiv \bot$ then $\rho_R(o, \blacktriangle) = (\top)$. This is because the antecedent $\lambda = f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ of any conditional will be inconsistent, so the negated antecedents of any negative conditional as well as the material counterpart of any positive conditional will be tautologies. Consequently, $\bigwedge U_j$ will always be a tautology. But also any conditional will be exceptional for U_0. This is because $U_0 \vdash \neg\lambda$ for any positive conditional $\lambda \Rightarrow \chi$ (as $\lambda \equiv \bot$), and $U_0 \cup \{\lambda\} \vdash \delta$ for any δ showing that negative conditionals are exceptional, as well. Hence, $\mathcal{C}_1 = \mathcal{C}_0$ and $\mathcal{N}_1 = \mathcal{N}_0$. So, $\rho_R(o, \blacktriangle) = (\bigwedge U_0) = (\top)$.

Now, an interesting thing to note about the rational prefix construction is that it actually goes through *independently* of whether ▲ is o-acceptable. In fact, a useful side-effect of the construction is that it actually *reveals* whether ▲ is o-acceptable. Given we have constructed $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$, all we have to do is to look at sentence α_m and check if it is a tautology:

PROPOSITION 3.4
Let $o \in O$ and $\blacktriangle \in L$, and let $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$ be the rational prefix of o with respect to ▲. Then
(i) if $\blacktriangle \not\equiv \bot$ and $\alpha_m \equiv \top$ then $[\rho_R(o, \blacktriangle), \blacktriangle]$ is an explanation for o.
(ii) if $\blacktriangle \equiv \bot$ or $\alpha_m \not\equiv \top$ then ▲ is not an o-acceptable core.

Thus, this proposition gives us a necessary and sufficient condition for ▲ to be an o-acceptable core. This will be used in the algorithm of Section 5. In Example 3.3 (i), the weakest element $\alpha_2 = p \wedge q \wedge r \rightarrow \neg q$ of $\rho_R(o, \blacktriangle)$ is not a tautology. The above proposition implies that $\top$ is not an o-acceptable core. This can be easily seen as $f(\sigma \cdot (p, q, r, \top)) \vdash p \wedge q \wedge r$ for any sequence σ so $\neg q$ cannot be consistently believed after the last input has been received. In Example 5.4, we will see that an o-acceptable core *does* nevertheless exist.

3.2 *Justification for using the rational prefix*

In the rest of this section, we assume ▲ to be some fixed *o-acceptable* core. As we just saw, $[\rho_R(o, \blacktriangle), \blacktriangle]$ then provides an explanation for o given this ▲. In this section, we want to show in precisely what sense it could be regarded as a *best* explanation given ▲. Let $\Sigma = \{\sigma \mid [\sigma, \blacktriangle] \text{ explains } o\}$.

One way to compare sequences in Σ is by focusing on the *trace* of belief sets they (in combination with ▲) induce through o, i.e. for each $\sigma \in \Sigma$ we can consider the sequence $(Bel_0^\sigma, Bel_1^\sigma, \ldots, Bel_n^\sigma)$, where Bel_i^σ is defined to be the beliefs after the i-th input in o (under the explanation $[\sigma, \blacktriangle]$). In other words $Bel_i^\sigma = f(\sigma \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))$, Bel_0^σ giving the initial belief set.

EXAMPLE 3.5
Let $o = \langle (p, s, \emptyset), (q, \top, \{s\}) \rangle$, $\blacktriangle = p \rightarrow \neg q$ and $\sigma = (p \rightarrow s)$. Then the belief trace $(Bel_0^\sigma, Bel_1^\sigma, Bel_2^\sigma)$ is $(p \rightarrow (s \wedge \neg q), p \wedge \neg q \wedge s, q \wedge \neg p)$.

The idea would then be to define a preference relation $\preceq_1$ over the sequences in Σ (with more preferred sequences corresponding to those 'lower' in the ordering) via some preference relation over their set of associated belief traces. Given any two possible belief traces $(\beta_0, \ldots, \beta_n)$ and $(\gamma_0, \ldots, \gamma_n)$, let us write $(\beta_0, \ldots, \beta_n) \leq_{\text{lex}} (\gamma_0, \ldots, \gamma_n)$ if and only if, for all $i = 0, \ldots, n$, $\big[\beta_j \equiv \gamma_j$ for all $j < i$ implies $\gamma_i \vdash \beta_i\big]$. Note that if two sequences are equivalent up to $i-1$ and neither $\beta_i \vdash \gamma_i$ nor $\gamma_i \vdash \beta_i$ then the two are simply incomparable with respect to $\leq_{\text{lex}}$. Then we define, for any $\rho, \sigma \in \Sigma$:

$$\rho \preceq_1 \sigma \text{ iff } (Bel_0^\rho, \ldots, Bel_n^\rho) \leq_{\text{lex}} (Bel_0^\sigma, \ldots, Bel_n^\sigma).$$

$\preceq_1$ is a pre-order, i.e. a reflexive and transitive relation, on Σ. Thus, given two sequences in Σ, we prefer that one which leads to $\mathcal{A}$ having fewer (i.e. weaker) beliefs before any of the inputs φ_i were received. If the two sequences lead to equivalent beliefs at this initial stage, then we prefer that which leads to $\mathcal{A}$ having fewer beliefs after φ_1 was received. If they lead to equivalent beliefs also after this stage, then we prefer that which leads to $\mathcal{A}$ having fewer beliefs after φ_2 was received, and so on. Thus, under this ordering, we prefer sequences which induce $\mathcal{A}$ to have *fewer* beliefs, *earlier* in o. The next result shows $\rho_R(o, \blacktriangle)$ is one among several best elements in Σ under this ordering.[5]

[5]There are always many different sequences that induce the same belief trace given a sequence of revision inputs. Consider $[(p \rightarrow q), \top]$ and $[(p \wedge \neg q \wedge r, p \rightarrow q), \top]$ both of which explain $\langle (p, q, \emptyset) \rangle$ as their belief traces are $(p \rightarrow q, p \wedge q)$. The first state

THEOREM 3.6
$\rho_R(o, \blacktriangle) \preceq_1 \sigma$ for all $\sigma \in \Sigma$.

That Theorem 3.6 holds is thanks essentially to the minimization which is already at play in the rational closure construction. Every rational inference relation may be specified by an ordering of 'naturalness', or ' plausibility' on the set of propositional worlds [20]. Then, as is shown in [5, 20], the rational closure of a given set of positive ($\mathcal{C}$) and negative ($\mathcal{N}$) conditionals corresponds to that ordering which assumes worlds to be as natural as $\mathcal{C}$ and $\mathcal{N}$ will allow. As a consequence of this, e.g. the rational closure $\Rightarrow_R$ enjoys the property that $\top \Rightarrow_R \lambda$ only if $\top \Rightarrow \lambda$ *for all* rational inference relations satisfying $\mathcal{C}$ and $\mathcal{N}$. And in Theorem 3.6, this particular property is responsible for the fact that $Bel_0^{\sigma} \vdash Bel_0^{\rho_R}$ for all $\sigma \in \Sigma$.

Another way to compare sequences is to look at their consequences for predicting what will happen at the next step after o.

$$\rho \preceq_2 \sigma \text{ iff } Bel([\sigma, \blacktriangle] * \varphi_1 * \cdots * \varphi_n * \lambda) \vdash Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_n * \lambda) \text{ for all } \lambda$$

Thus, according to *this* preference criterion we prefer ρ to σ if it always leads to fewer beliefs being predicted after the next revision input. Recall that in the assumed belief revision framework the beliefs after a sequence of revisions can be obtained via a single revision. That is, this criterion is also about *any sequence* of further revisions. It turns out $\rho_R(o, \blacktriangle)$ is a most preferred element under $\preceq_2$ *amongst all minimal elements under* $\preceq_1$.

THEOREM 3.7
For all $\sigma \in \Sigma$, if $\sigma \preceq_1 \rho_R(o, \blacktriangle)$ then $\rho_R(o, \blacktriangle) \preceq_2 \sigma$.

Thus if we take a lexicographic combination of $\preceq_1$ and $\preceq_2$ (with $\preceq_1$ being considered as more important), $\rho_R(o, \blacktriangle)$ emerges overall as a best, most preferred, member of Σ. Consequently, these theorems point out a single best explanation given an observation o and an o-acceptable core. This explanation is not unique but any other optimal one will yield exactly the same belief trace. Having provided a method for finding the best explanation $[\rho, \blacktriangle]$ *given* $\blacktriangle$, we now turn our attention to finding the best $\blacktriangle$ itself.

4 Minimizing $\blacktriangle$

As argued earlier, core beliefs are needed, but at the same time we try to minimize the assumptions about the agent's beliefs. This includes minimizing $\blacktriangle$. The first idea would be to simply take the disjunction of all possible o-acceptable cores, i.e. to take $\blacktriangle_\vee(o)$, defined by

$$\blacktriangle_\vee(o) \equiv \bigvee \{\blacktriangle \mid \blacktriangle \text{ is } o\text{-acceptable}\}.$$

First note that $\blacktriangle_\vee(o)$ is inconsistent if and only if there is no o-acceptable core as $\bigvee \emptyset \equiv \bot$. So from $\blacktriangle_\vee(o)$ being consistent we can read off that there is an explanation. But is $\blacktriangle_\vee(o)$ itself o-acceptable in this case? Thankfully the answer is yes, a result which follows (in our finite setting) from the following proposition which says that the family of o-acceptable cores is closed under disjunctions.

---

corresponds to the rational explanation of the observation and it will yield a weaker belief than the other one when considering a further input $p \wedge \neg q$.

PROPOSITION 4.1
If $\blacktriangle_1$ and $\blacktriangle_2$ are o-acceptable then so is $\blacktriangle_1 \vee \blacktriangle_2$.

So as a corollary $\blacktriangle_\vee(o)$ does indeed satisfy:

(Acceptability) If an o-acceptable core exists then $\blacktriangle(o)$ is o-acceptable

What other properties does $\blacktriangle_\vee(o)$ satisfy?

(Consistency) If $\blacktriangle(o) \not\equiv \bot$ then there is an o-acceptable core.

If we are returned a consistent sentence, we want to be guaranteed that an explanation does indeed exist. Acceptability and Consistency would appear to be absolute rock-bottom properties which we would expect of *any* method for finding a good o-acceptable core, as they express that $\blacktriangle(o)$ yields an o-acceptable core if and only if such a core exists. However for $\blacktriangle_\vee$ we can say more. Given two observations o and o', recall that $o \cdot o'$ denotes the concatenation of o and o'. We shall use $o \sqsubseteq_{\text{right}} o'$ to denote that o' *right extends* o., i.e. $o' = o \cdot o''$ for some (possibly empty) $o'' \in O$, and $o \sqsubseteq_{\text{left}} o'$ to denote o' *left extends* o, i.e. $o' = o'' \cdot o$ for some (possibly empty) $o'' \in O$.

PROPOSITION 4.2
Suppose $o \sqsubseteq_{\text{right}} o'$ or $o \sqsubseteq_{\text{left}} o'$. Then every o'-acceptable core is an o-acceptable core.

As a result of this we see $\blacktriangle_\vee$ satisfies the following two properties, which say extending the observation into the future or past leads only to a logically stronger core being returned.

(Right Monotony) If $o \sqsubseteq_{\text{right}} o'$ then $\blacktriangle(o') \vdash \blacktriangle(o)$
(Left Monotony) If $o \sqsubseteq_{\text{left}} o'$ then $\blacktriangle(o') \vdash \blacktriangle(o)$.

Right- and Left Monotony provide ways of expressing that $\blacktriangle(o)$ leads only to *safe* conclusions that something is a core belief of $\mathcal{A}$—conclusions that cannot be 'defeated' by additional information about $\mathcal{A}$ that might come along in the form of observations prior to, or after o.

We should point out, though, that it is *not* the case that by inserting any observation *anywhere* in o, $\blacktriangle_\vee$ will always lead to a logically stronger core. Consider $o_1 = \langle (p,p,\emptyset),(q,\neg p,\emptyset)\rangle$ and $o_2 = \langle (p,p,\emptyset),(\neg p,\neg p,\emptyset),(q,\neg p,\emptyset)\rangle$, i.e. $(\neg p,\neg p,\emptyset)$ was inserted in the middle of o_1. $\blacktriangle_\vee(o_1) \equiv q \rightarrow \neg p$ whereas $\blacktriangle_\vee(o_2) \equiv \top$. So although o_2 extends o_1 in a sense, the corresponding $\blacktriangle_\vee$ is actually weaker. Looking at o_1, assuming as we do that $\mathcal{A}$ received *no* inputs between p and q, the *only* way to explain the end belief in $\neg p$ is to ascribe core belief $q \rightarrow \neg p$ to $\mathcal{A}$ (cf. the 'Recalcitrance' rule in Section 2). However, looking at o_2, the information that $\mathcal{A}$ received (and believed) the intermediate input $\neg p$ is enough to 'explain away' this end belief without recourse to core beliefs. Our assumption that $\mathcal{A}$ received no other inputs between φ_1 and φ_n during an observation $o = \langle (\varphi_1,\theta_1,D_1),\ldots,(\varphi_n,\theta_n,D_n)\rangle$ is rather strong. It amounts to saying that, during o, we kept our eye on $\mathcal{A}$ the whole time. The above example shows that relaxing this assumption gives us an extra degree of freedom with which to explain o, via the inference of intermediate inputs (the first steps of an investigation can be found in [25]).

It turns out the above four properties are enough to actually *characterise* $\blacktriangle_\vee$. In fact, given the first two, just *one* of Right- and Left Monotony is sufficient for this task:

PROPOSITION 4.3
Let $\blacktriangle : O \rightarrow L$ be any function which returns a sentence given any $o \in O$. Then the following are equivalent:
(i) $\blacktriangle$ satisfies Acceptability, Consistency and Right Monotony.
(ii) $\blacktriangle$ satisfies Acceptability, Consistency and Left Monotony.
(iii) $\blacktriangle(o) \equiv \blacktriangle_\vee(o)$ for all $o \in O$.

Note that as a corollary to this proposition we get the surprising result that, in the presence of Acceptability and Consistency, Right- and Left Monotony are in fact *equivalent*. Combining the findings of the last two sections, we are now ready to announce our candidate for the best explanation for o. By analogy with 'rational closure', we make the following definition:

DEFINITION 4.4
Let $o \in O$ be an observation for which an o-acceptable core exists. Then we call $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ the *rational explanation* for o.

In Section 6, we will give some examples of what we can infer about $\mathcal{A}$ under the rational explanation. But how might we find it in practice? The next section gives an algorithm for just that.

5 Constructing the rational explanation

The idea behind the algorithm is as follows. Given an observation o, we start with the weakest possible core $\blacktriangle_0 = \top$ and construct the rational prefix $(\alpha_m, \ldots, \alpha_0) = \rho_0$ of o with respect to $\blacktriangle_0$. We then check whether α_m is a tautology. If it is then we know by Proposition 3.4 that $[\rho_0, \blacktriangle_0]$ is an explanation for o and so we stop and return this as output. If it is not then Proposition 3.4 tells us $\blacktriangle_0$ cannot be o-acceptable. In this case, we modify $\blacktriangle_0$ by *conjoining* α_m to it, i.e. by setting $\blacktriangle_1 = \blacktriangle_0 \wedge \alpha_m$. Constructing the rational prefix of o with respect to the new core then leads to a *different* prefix, which can be dealt with the same way.

Algorithm 1: calculation of the rational explanation

Input: observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$
Output: the rational explanation for o

$\blacktriangle \Leftarrow \top$
repeat
  $\rho \Leftarrow \rho_R(o, \blacktriangle)$ /* $\rho = (\alpha_m, \ldots, \alpha_0)$ */
  $\blacktriangle \Leftarrow \blacktriangle \wedge \alpha_m$
until $\alpha_m \equiv \top$
return $[\rho, \blacktriangle]$ if $\blacktriangle \not\equiv \bot$, "no explanation" otherwise

Before showing that the output of this algorithm matches the rational explanation, we need to be sure it always terminates. This is a consequence of the following:

LEMMA 5.1
Let $\blacktriangle$ and α_m be as after the calculation of $\rho_R(o, \blacktriangle)$. If $\alpha_m \not\equiv \top$ then $\blacktriangle \not\equiv \blacktriangle \wedge \alpha_m$.

This result assures us that if the termination condition of the algorithm does not hold, the new core will be *strictly* logically stronger than the previous one. Thus the cores generated by the algorithm become progressively strictly stronger. In our setting, in which we assumed a *finite* propositional language, this means, in the worst case, the process will continue until $\blacktriangle \equiv \bot$ in which case the rational prefix will be $(\top)$, as shown in Example 3.3 (ii), and the calculation terminates.

Now, to show the output matches the rational explanation in case an explanation exists, consider the sequence $[\rho_0, \blacktriangle_0], \ldots, [\rho_k, \blacktriangle_k]$ of epistemic states generated by the algorithm. We need to show $\blacktriangle_k \equiv \blacktriangle_\vee(o)$. The direction $\blacktriangle_k \vdash \blacktriangle_\vee(o)$ follows from the fact that $[\rho_k, \blacktriangle_k]$ is an explanation for o and so $\blacktriangle_k$ is an o-acceptable core. The converse $\blacktriangle_\vee(o) \vdash \blacktriangle_k$ is proved by showing inductively that

$\blacktriangle_\vee(o) \vdash \blacktriangle_i$ for each $i = 0, \ldots, k$: the case $i = 0$ clearly holds since $\blacktriangle_0 \equiv \top$. The inductive step uses the following property:

LEMMA 5.2
Let $0 < i \leq k$ and suppose $\rho_{i-1} = (\alpha_m, \ldots, \alpha_0)$. Then, for *any* o-acceptable core $\blacktriangle'$, if $\blacktriangle' \vdash \blacktriangle_{i-1}$ then $\blacktriangle' \vdash \alpha_m$.

This enables us to prove that, given $\blacktriangle_\vee(o) \vdash \blacktriangle_{i-1}$, we must also have $\blacktriangle_\vee(o) \vdash \blacktriangle_i$. Thus $\blacktriangle_\vee(o) \vdash \blacktriangle_k$ as required. This lemma does not depend on o having an explanation, i.e. if $\blacktriangle_k \equiv \bot$ we still have $\blacktriangle_\vee(o) \vdash \blacktriangle_k$. Hence, if the algorithm tells us there is no explanation for o, then indeed there is no o-acceptable core. Since obviously ρ_k is the rational prefix of o with respect to $\blacktriangle_k$ by construction, we have:

PROPOSITION 5.3
Given an observation o for which an o-acceptable core exists, the algorithm outputs the rational explanation for o.

EXAMPLE 5.4
Let $o = \langle (p, s, \emptyset), (q, \top, \{s\}), (r, \neg q, \emptyset) \rangle$ as in Example 3.3 (i). Starting with $\blacktriangle = \top$, we get $\rho_R(o, \blacktriangle) = (p \wedge q \wedge r \rightarrow \neg q, p \wedge q \wedge r \rightarrow \neg q, (p \rightarrow s) \wedge (p \wedge q \wedge r \rightarrow \neg q) \wedge \neg(p \wedge q))$ as shown in that example. $\alpha_m = \alpha_2 = p \wedge q \wedge r \rightarrow \neg q \equiv \neg p \vee \neg q \vee \neg r$ is not a tautology, so we need to modify the core to $\blacktriangle = \neg p \vee \neg q \vee \neg r$. Consequently, we get the following sets of conditionals for the second iteration.

$$\begin{aligned}
\mathcal{C}_0 = \mathcal{C}_\blacktriangle(o) &= \{f(p, \neg p \vee \neg q \vee \neg r) \Rightarrow s, f(p, q, \neg p \vee \neg q \vee \neg r) \Rightarrow \top, \\
&\quad f(p, q, r, \neg p \vee \neg q \vee \neg r) \Rightarrow \neg q\} \\
&= \{p \wedge (\neg q \vee \neg r) \Rightarrow s, p \wedge q \wedge \neg r \Rightarrow \top, \neg p \wedge q \wedge r \Rightarrow \neg q\} \\
\mathcal{N}_0 = \mathcal{N}_\blacktriangle(o) &= \{f(p, q, \neg p \vee \neg q \vee \neg r) \Rightarrow s\} \\
&= \{p \wedge q \wedge \neg r \Rightarrow s\}
\end{aligned}$$

$p \wedge q \wedge \neg r \Rightarrow s$ is n-exceptional for $\tilde{\mathcal{C}_0}$ as $\{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r\} \vdash s$, so $U_0 = \{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r \rightarrow \top, \neg p \wedge q \wedge r \rightarrow \neg q, \neg(p \wedge q \wedge \neg r)\}$. Only $p \wedge (\neg q \vee \neg r) \Rightarrow s$ is not p-exceptional for U_0, so we get

$$\begin{aligned}
\mathcal{C}_1 &= \{p \wedge q \wedge \neg r \Rightarrow \top, \neg p \wedge q \wedge r \Rightarrow \neg q\} \\
\mathcal{N}_1 &= \{p \wedge q \wedge \neg r \Rightarrow s\}
\end{aligned}$$

Note that this time $p \wedge q \wedge \neg r \Rightarrow s$ is not n-exceptional for $\tilde{\mathcal{C}_1}$ so $U_1 = \{p \wedge q \wedge \neg r \rightarrow \top, \neg p \wedge q \wedge r \rightarrow \neg q\}$. Only $\neg p \wedge q \wedge r \Rightarrow \neg q$ is p-exceptional for U_1 (in fact again the material counterpart of this conditional is inconsistent with its own antecedent) so we get

$$\begin{aligned}
\mathcal{C}_2 &= \mathcal{C}_3 = \{\neg p \wedge q \wedge r \Rightarrow \neg q\} \\
\mathcal{N}_2 &= \mathcal{N}_3 = \emptyset \\
U_2 &= U_3 = \{\neg p \wedge q \wedge r \rightarrow \neg q\}
\end{aligned}$$

Again $\alpha_2 \equiv p \vee \neg q \vee \neg r \not\equiv \top$, so the core belief has to be adapted once more. Conjoining the old one with α_2 leads to a core that is equivalent to $\blacktriangle = \neg q \vee \neg r$, so this time the conditionals look as follows

$$\begin{aligned}
\mathcal{C}_0 = \mathcal{C}_\blacktriangle(o) &= \{f(p, \neg q \vee \neg r) \Rightarrow s, f(p, q, \neg q \vee \neg r) \Rightarrow \top, \\
&\quad f(p, q, r, \neg q \vee \neg r) \Rightarrow \neg q\} \\
&= \{p \wedge (\neg q \vee \neg r) \Rightarrow s, p \wedge q \wedge \neg r \Rightarrow \top, p \wedge \neg q \wedge r \Rightarrow \neg q\} \\
\mathcal{N}_0 = \mathcal{N}_\blacktriangle(o) &= \{f(p, q, \neg q \vee \neg r) \Rightarrow s\} \\
&= \{p \wedge q \wedge \neg r \Rightarrow s\}
\end{aligned}$$

So, $\tilde{\mathcal{C}}_0 = \{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r \rightarrow \top, p \wedge \neg q \wedge r \rightarrow \neg q\}$, the last two implications being tautologies. $p \wedge q \wedge \neg r \Rightarrow s$ is n-exceptional for that set as $\{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r\} \vdash s$. Hence $U_0 = \{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r \rightarrow \top, p \wedge \neg q \wedge r \rightarrow \neg q, \neg(p \wedge q \wedge \neg r)\}$. $p \wedge q \wedge \neg r \Rightarrow \top$ is p-exceptional for U_0 as we had to add the negated antecedent of the only negative conditional. However, the other positive conditionals are not p-exceptional for U_0 so we get $\mathcal{C}_1 = \{p \wedge q \wedge \neg r \Rightarrow \top\}$, $\mathcal{N}_1 = \{p \wedge q \wedge \neg r \Rightarrow s\}$. $\tilde{\mathcal{C}}_1$ amounts to $\{\top\}$ and none of the remaining two conditionals is exceptional for that set. Hence we get $U_1 = \{\top\}$ and $\mathcal{C}_2 = \mathcal{C}_3 = \emptyset, \mathcal{N}_2 = \mathcal{N}_3 = \emptyset$ and $U_2 = U_3 = \emptyset$. As a consequence, we get $\rho_R(o, \blacktriangle) = (\top, \top, (p \wedge (\neg q \vee \neg r) \rightarrow s) \wedge \neg(p \wedge q \wedge \neg r))$. As this means $\alpha_2 = \top$, no further run is necessary and the rational explanation for o is $[(\top, \top, (p \wedge (\neg q \vee \neg r) \rightarrow s) \wedge \neg(p \wedge q \wedge \neg r)), \neg q \vee \neg r]$.

$((\neg q \vee \neg r) \wedge (p \rightarrow s \wedge \neg q), p \wedge \neg q \wedge s, p \wedge q \wedge \neg r, p \wedge \neg q \wedge r \wedge s)$ is the belief trace according to this explanation. The core belief must entail $\neg q \vee \neg r$ as otherwise due to 'Recalcitrance' $\mathcal{A}$ would have to believe q after receiving r but the observation tells us the opposite. Further, nothing indicates that $\blacktriangle$ would have to be stronger than that. Initially, the agent also believes $p \rightarrow s \wedge \neg q$. The belief in $p \rightarrow s$ is clear as after hearing p the agent believes s and p alone does not entail s. The belief in $p \rightarrow \neg q$ is more subtle and is best explained when looking at the beliefs after the first revision input was received. The beliefs in p and s are clear, but why should $\mathcal{A}$ commit to $\neg q$? If it did not then q would be consistent with the current beliefs and revision by q would turn out to be an expansion of the belief set. However, o tells us that $\mathcal{A}$ ceases to believe s, so it cannot be an expansion by q—this also explains the belief in $p \rightarrow \neg q$ in the initial state. The belief in r, $\neg q$ and p in the final state are quite intuitive; r has just been received, the observation requires $\neg q$ to be believed and there is no reason why $\mathcal{A}$ should suddenly reject p. s is believed again, as the apparent reason not to believe s, namely q, is gone.

Algorithm 1 is computationally costly. Calculating $\mathcal{C}_{\blacktriangle}(o)$ and $\mathcal{N}_{\blacktriangle}(o)$ from a given observation o and a core belief $\blacktriangle$ requires at most a polynomial number of satisfiability tests. Calculating the rational prefix from these sets of conditionals can also be done using a polynomial number of satisfiability tests. However, the main source of complexity is the repeat loop which is needed to refine the core belief until the weakest o-acceptable one is found. An exponential number of iterations may be required for that. The following observation is an example where the algorithm goes through an exponential number of different core beliefs. $o = \langle (p_1, \top, \emptyset), \ldots, (p_n, \top, \emptyset), (p_{n+1}, \theta, \emptyset) \rangle$ where $\theta = \bigwedge_{1 \leq i \leq n+1} \neg p_i$ and all p_j, $1 \leq j \leq n+1$, are distinct propositional variables. This indicates that the algorithm itself may not be suitable for giving complexity results of the general problem. Both deciding if $[\rho, \blacktriangle]$ explains an observation o and whether $\blacktriangle$ is o-acceptable are Δ_2^P-complete problems. As indicated above we need at most a polynomial number of NP-oracle calls and the entailment problem of linear base revision, which is Δ_2^P-complete [24], can be reduced to these decision problems. Deciding whether a given observation o has an explanation is in Σ_2^P, deciding whether a given core belief $\blacktriangle$ is the weakest possible one ($\blacktriangle \equiv \blacktriangle_{\vee}(o)$) is in Π_2^P (hardness of these problems is an open question). The proof of these complexity results involves guessing the right core belief and some further machinery is needed to show that this is indeed possible. Full details can be found in [26].

6 More examples

In this section, we want to give a few more simple examples to illustrate the rational explanation.

For $o = \langle (p, q, \emptyset) \rangle$, the rational explanation is $[(\top, p \rightarrow q), \top]$. So we infer $\mathcal{A}$'s initial belief set is $p \rightarrow q$. Indeed to explain $\mathcal{A}$'s belief in q following receipt of p it is clear $\mathcal{A}$ *must* initially believe *at least* $p \rightarrow q$ since p itself does not entail q. It seems fair to say we are not justified in ascribing to $\mathcal{A}$ any initial beliefs beyond this. After $\mathcal{A}$ receives p we assume $\mathcal{A}$ *believes* this input—we have

no reason to expect otherwise—and so has belief set $p \wedge q$. If $\mathcal{A}$ is given a *further* input $\neg(p \wedge q)$ we predict $\mathcal{A}$ will also believe this input, but will hold on to its belief in p. The reason being we assume $\mathcal{A}$, having only just been told p, now has stronger reasons to believe p than q. If, instead, $\mathcal{A}$ is given further input $\neg p$ we predict its belief set will be just $\neg p$, i.e. we do *not* assume $\mathcal{A}$'s belief in q persists. Essentially the rational explanation assumes the prior input p must have been *responsible* for $\mathcal{A}$'s prior belief in q. And with this input now being 'overruled' by the succeeding input, $\mathcal{A}$ can no longer draw any conclusions about the truth of q.

Another illustrative example is $o = \langle (p, \neg p, \emptyset) \rangle$, for which the rational explanation is $[(\top), \neg p]$. Indeed $\neg p$ must be a core belief, as that is the only possibility for p to be rejected. And if p was not rejected, the agent could not consistently believe $\neg p$.

As mentioned before, some observations do not have an explanation. We gave one observation that would have forced the violation of a basic property of the belief revision framework. $o = \langle (p, \neg p, \emptyset), (\neg p, p, \emptyset) \rangle$ is another example but here the problem lies elsewhere. o says that upon receiving p its negation is to be believed. Consequently, p must be blocked which is achieved by a core belief entailing $\neg p$. But then $\neg p$ must also be blocked as well, so the core belief must entail p as well. Hence, the core would have to be inconsistent and we do not consider this an explanation.

7 Related work

Before concluding, we want to discuss the relation of our work to two other papers. In 'Belief reconstruction in cooperative dialogues' [9], Herzig *et al.* deal with the question of determining agents' beliefs through a sequence of speech acts. The new beliefs should depend on the old ones and the input received and old beliefs should persist if possible. The key point of the motivation is that an input should not always be accepted. In particular, it should be be rejected if the speaker is incompetent with respect to the content of the utterance.

Having a setting of a man–machine dialogue in mind, the authors present a multi-modal framework as well as functions and axioms for modelling notions of subject, scope and competence. One such axiom expresses that if an agent is competent on the topic of a formula φ, which does not contain modal operators, and it also believes that formula to hold then φ does indeed hold. An axiom for preservation expresses that if the scope of a speech act does not touch the topic of some formula then that formula remains to be true after the speech act is carried out in case it was true before. This additional machinery is used to put restrictions on models. Together with the laws governing the revision process this allows to calculate the beliefs after a speech act has been performed.

The paper has a traditional first person perspective of the agent—what should it believe upon receiving some new information and progressing the beliefs given the initial state. The assumed revision framework is more sophisticated than the one we use. However, the paper does not deal with reasoning about the other agent retrospectively, what prior beliefs it may have held. Competence etc. are fixed and given for all parties involved. In analogy to the motivation of our work, it would be interesting to actually infer information about the competence of an agent, static laws (beliefs that cannot be changed by revision), former beliefs etc. given a dialogue and information about the evolution of the beliefs of agents involved in it. Consequently, the title suggests a connection that turns out to be superficial.

A paper that deals with completing information about beliefs over time is [8] in which Dupin de Saint-Cyr and Lang present *belief extrapolation* operators. The starting point is a scenario $\langle \theta_1, \ldots, \theta_N \rangle$ representing that θ_i holds at time point i. Such a scenario is a partial description of how the world evolved. Assuming that fluents (literals) tend not to change, i.e. that the world is inertial, the operator

tries to identify preferred trajectories of models $\langle m_1, \ldots, m_N \rangle$ such that $m_i \models \theta_i$. The authors present several strategies for minimization change: counting all changes of fluents, counting changes *per* fluent, penalties for changes, temporal considerations like changes occurring as late as possible, etc. Each strategy gives rise to a preference relation among trajectories and, choosing one preference relation, the result of the extrapolation is what is true at each point according to all preferred trajectories explaining a scenario. A preference relation is called inertial if all static trajectories, i.e. those where all models are identical and thus no change occurs, are equally preferred and preferred to any non-static trajectory. The rationale behind choosing an inertial preference relation is that as long as we can assume the world not to have changed, we should do so. The authors present properties and connections between the different preference relations and the extrapolation operators they give rise to and position extrapolation with respect to belief revision and update.

There are some essential differences between belief extrapolation and our approach. Once more, the work in [8] is focused on a first person perspective and describes what an agent *should* believe at each point in time rather than reasoning about what the agent *does* believe. The second very important difference is that both approaches minimize very different things. For the sake of illustrating this, let us for now forget that we assumed the world to be static and only the agent's beliefs about it change. Given the scenario $\langle p, q, r \rangle$, an extrapolation operator based on an inertial preference relation will conclude that $p \wedge q \wedge r$ held at every point in time. This is because the conjunction of all the sentences is consistent and it can thus be assumed that nothing changed at all. We will now look at three potential translations from a scenario $\langle \theta_1, \ldots, \theta_N \rangle$ to observations in our setting: (i) $\langle (\theta_1, \top, \emptyset), \ldots, (\theta_N, \top, \emptyset) \rangle$, (ii) $\langle (\theta_1, \theta_1, \emptyset), \ldots, (\theta_N, \theta_N, \emptyset) \rangle$ and (iii) $\langle (\top, \theta_1, \emptyset), \ldots, (\top, \theta_N, \emptyset) \rangle$. The rational explanation for any observation of the form (i) and (ii) will be $[(), \top]$. This is because the material counterparts for all positive conditionals will be tautologies. The belief trace for the example scenario will thus be $(\top, p, p \wedge q, p \wedge q \wedge r)$. So with respect to these translations we do not conclude that $p \wedge q \wedge r$ is believed at every point in time. Our approach tries to minimize the *beliefs* we assign to the agent and not the changes. As the agent *may* consider $\neg p$ more plausible than p etc. we cannot conclude that it believed p before being informed about it. In this sense belief extrapolation is credulous, using the inertia assumption in order to come up with strong beliefs. The rational explanation is a sceptical approach. Although the world considered may be static the agent's information about it may be highly unreliable and hence the agent's *beliefs* may change often and dramatically.

For the given scenario the third translation yields the same conclusion as the belief extrapolation operator. However, the inputs $\top$ in fact *force* us to conclude that the agent's belief set did not change at all and every belief must have been already present in the initial state. (iii) will fail whenever $\bigwedge \theta_i$ is inconsistent, i.e. in all interesting cases. The resulting observation will not have an explanation at all.

A third essential difference between our work and [8] is that (in its original form) the extrapolation operator does not incorporate information *that* a change occurred or *why* it may have occurred. Given a scenario, *a priori* any fluent may have changed at any time and the operator tries to minimize these changes according to the preference criterion. This is because the only information available to the extrapolation operator is the scenario which only contains a partial description of the world at every point but no information about what happened, or if anything happened at all. For our work, we assume to be provided with richer information. The revision inputs φ_i can be considered to be a possible cause for θ_i to be believed. This input, which has indeed been received, may have triggered the change in mind. But conversely, nothing but the recorded inputs may have triggered a change. The observation $\langle (p, p \wedge \neg q, \emptyset), (p, p \wedge q, \emptyset) \rangle$ does not have an explanation. However, the scenario $\langle p \wedge \neg q, p \wedge q \rangle$ does—in principle any fluent may change at any point in time.

The authors of [8] indicate how explicit information about change can be incorporated into their approach. They suggest mixed scenarios where each sentence is labelled indicating whether it denotes

an actual description of the state of the world or an *expected change* caused by an update. Then not all possible trajectories are considered but only those which fit the expected changes recorded in the mixed scenario. These trajectories are then compared with respect to the *unexpected* changes. This still does not cause the two approaches to collapse into the same method. The rational explanation of an observation is about finding an initial epistemic state that makes all changes expected ones.

Finally, there is a large body of work on explanations in a number fields, one being causal reasoning (see e.g. [14]). It often looks for atomic events as explanations for some observed fact. However, in our observations the events that have occurred are in fact given. We may try to find out *if* they were the cause for $\mathcal{A}$ believing what is recorded (given the initial state and the recorded inputs up to that point). In other words, would the agent have believed the same when receiving a tautology? But this is not really at the core of our work. In analogy to [14], we try to identify (and construct) the causal structure and the context which we then call an explanation for o.

8 Conclusion

To conclude, in this article we made an attempt at reconstructing an agent's initial epistemic state in order to explain a given observation of the agent. We did so by assuming a simple yet powerful model for epistemic states allowing for iterated non-prioritized revision. The algorithm we provided constructs a best (in terms of beliefs ascribed to the agent) explanation based on the rational closure of conditional beliefs. The main contribution here is the calculation of a suitable set of conditionals capturing the observation. Due to the nature of the belief revision framework this is non-trivial as they heavily depend on the core belief. Note that in our work the explanation is not the event that triggered an change—these events are in fact known. We are after the relation linking revision inputs with beliefs, not for a single step but a sequence. That is, we gave a method for reasoning about information of the form 'Revision by φ_1 leads to beliefs θ_1 and non-beliefs D_1 and an epistemic state in which further revision by φ_2 leads to …' Having calculated an initial epistemic state, the questions posed in the introduction are answered by simply progressing the revision inputs starting in that state, yielding conclusions about prior as well as future beliefs. This should be applicable to problems requiring the modelling of agents' beliefs, e.g. in the area of user modelling [27]. We want to remark that conclusions based on the rational explanation have to be used with care and point the reader to [26] for a detailed discussion. It would be of interest to see whether similar results can be obtained when assuming $\mathcal{A}$ to employ other belief revision frameworks (e.g. [21]) and how they relate to the results presented in this article. In analogy to [8], one could also investigate whether it is possible to construct explanations that follow other minimization strategies.

Acknowledgements

Thanks are due to the anonymous reviewers for some helpful comments, and also to the organisers and audience of the 2005 Dagstuhl seminar on "Belief Change in Rational Agents" for providing an environment for some stimulating discussion on the paper's topic.

References

[1] C. Alchourron, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, **50**, 510–530, 1985.

[2] R. Booth. On the logic of iterated non-prioritised revision. In *Conditionals, Information and Inference – Selected papers from the Workshop on Conditionals, Information and Inference, 2002*, pp. 86–107. Springer's LNAI 3301, 2005.

[3] R. Booth and T. Meyer. Admissible and restrained revision. *Journal of Artificial Intelligence Research* , **26**, 127–151, 2006.
[4] R. Booth and A. Nittka. Reconstructing an agent's epistemic state from observations. In *Proceedings of IJCAI-05*, pp. 394–399, Professional Book Center, Denver, CO, 2005.
[5] R. Booth and J. B. Paris. A note on the rational closure of knowledge bases with both positive and negative knowledge. *Journal of Logic, Language and Information*, **7**, 165–190, 1998.
[6] R. I. Brafman and M. Tennenholtz. Modeling agents as qualitative decision makers. *Artificial Intelligence*, **94**, 217–268, 1997.
[7] A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, **89**, 1–29, 1997.
[8] F. Dupin de Saint-Cyr and J. Lang. Belief extrapolation (or how to reason about observations and unpredicted change). In *Proceedings of KR'02*, pp. 497–508, Morgan Kaufmann, San Francisco, CA, 2002.
[9] L. Fariñas del Cerro, A. Herzig, D. Longin, and O. Rifi. Belief reconstruction in cooperative dialogues. In *Proceedings of AIMSA'98*, pp. 254–266. Vol. 1480 of *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg, 1998.
[10] M. Freund. On the revision of preferences and rational inference processes. *Artificial Intelligence*, **152**, 105–137, 2004.
[11] P. Gärdenfors. *Knowledge in Flux*. MIT Press, Cambridge, MA, 1988.
[12] H. Geffner and J. Pearl. Conditional entailment: Bridging two approaches to default entailment. *Artificial Intelligence*, **53**, 209–244, 1992.
[13] A. Grove. Two modelings for theory change. *Journal of Philosophical Logic*, **17**, 157–170, 1988.
[14] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach – Part ii: Explanations. In *Proceedings of IJCAI'01*, pp. 27–34, Morgan Kaufmann, San Francisco, CA, 2001.
[15] S. O. Hansson, E. Fermé, J. Cantwell, and M. Falappa. Credibility-limited revision. *Journal of Symbolic Logic*, **66**, 1581–1596, 2001.
[16] S. Konieczny and R. Pino Pérez. A framework for iterated revision. *Journal of Applied Non-Classical Logics*, **10**, 339–367, 2000.
[17] J. Lang. A preference-based interpretation of other agents' actions. In *Proceedings of KR'04*, pp. 644–653, AAAI Press, Menlo Park, CA, 2004.
[18] D. Lehmann. Another perspective on default reasoning. *Annals of Mathematics and Artificial Intelligence*, **15**, 61–82, 1995.
[19] D. Lehmann. Belief revision, revised. In *Proceedings of IJCAI'95*, pp. 1534–1540, Morgan Kaufmann, San Francisco, CA, 1995.
[20] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, **55**, 1–60, 1992.
[21] D. Lehmann, M. Magidor, and K. Schlechta. Distance semantics for belief revision. *Journal of Symbolic Logic*, **66**, 295–317, 2001.
[22] D. Makinson. Screened revision. *Theoria*, **63**, 14–23, 1997.
[23] A. Nayak, M. Pagnucco, and P. Peppas. Dynamic belief revision operators. *Artificial Intelligence*,**146**, 193–228, 2003.
[24] B. Nebel. Base revision operations and schemes: Semantics, representation and complexity. In *Proceedings of ECAI'94*, pp. 342–345, John Wiley and Sons, Chichester, 1994.
[25] A. Nittka. Reasoning about an agent based on its revision history with missing inputs. In *Proceedings of JELIA06*, M. Fisher and W. van der Hoek, eds, LNAI. Springer-Verlag, Berlin/Heidelberg, 2006.

[26] A. Nittka. *A Method for Reasoning About Other Agents' Beliefs from Observations*. PhD thesis, Universität Leipzig, 2008.
[27] W. Pohl. Logic-based representation and reasoning for user modeling shell systems. *User Modeling and User-Adapted Interaction*, **9**, 217–283, 1999.
[28] H. Rott. *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning*. Oxford University Press, Oxford, 2001.

Appendix: Proofs

A.1 Properties of f

When setting up our model of an agent's epistemic state in Section 2, we assumed the agent's belief set may be calculated using the function f. We begin this appendix by giving some important properties of this function which will be useful in our proofs. Throughout this section σ, ρ, etc. will denote sequences of sentences and $\alpha, \theta, \varphi, \lambda$, etc. will denote sentences. First we have the following easy properties.

PROPOSITION A.1
(i) $f(\alpha_m, \ldots, \alpha_1) = f(\alpha_m, \ldots, f(\alpha_i, \ldots, \alpha_1))$ for any $1 \leq i \leq m$.
(ii) If $\alpha_i \equiv \beta_i$ for $i = 1, 2$ then $f(\alpha_2, \alpha_1) \equiv f(\beta_2, \beta_1)$.
(iii) $f(\sigma \cdot \alpha) \vdash \alpha$.
(iv) If $f(\sigma \cdot \alpha) \equiv \bot$ then $\alpha \equiv \bot$.
(v) $f(\alpha, \beta) \equiv f(\alpha \wedge \beta, \beta)$.
(vi) If $\beta \vdash \alpha$ then $f(\alpha, \beta) = \beta$.

The above properties are all easy to prove, and will be used freely in what follows. (i) says f is right-associative, which follows immediately from its definition. (ii) is a syntax-irrelevance property. In combination with (i) it entails that when applying f to a sequence we can freely replace any sentence in that sequence by a logically equivalent one without changing the logical content of the result. The next property is a little more involved.

PROPOSITION A.2
If $f(\rho_1 \cdot \rho_2) \nvdash \neg\theta$ then $f(\rho_1 \cdot \theta \cdot \rho_2) \equiv f(\rho_1 \cdot \rho_2) \wedge \theta$.

PROOF. We know $f(\rho_1 \cdot \theta \cdot \rho_2) = f(\rho_1 \cdot f(\theta, f(\rho_2)))$ and $f(\rho_1 \cdot \rho_2) = f(\rho_1 \cdot f(\rho_2))$. Hence these two collect the same sentences from ρ_2. From $f(\rho_1 \cdot \rho_2) \nvdash \neg\theta$ it follows that $f(\rho_2) \nvdash \neg\theta$ and hence $f(\theta, f(\rho_2)) = \theta \wedge f(\rho_2)$.

So, if we can show that $f(\rho_1 \cdot \theta \wedge f(\rho_2))$ and $f(\rho_1 \cdot f(\rho_2))$ also collect the same sentences from ρ_1 the proposition immediately follows. This is because the order of the elements in a conjunction does not matter.

The argument that the same sentences from ρ_1 are chosen is an inductive one. Assume both have collected the same elements from some suffix of ρ_1 their conjunction being denoted by χ so far. The next sentence to be considered is ψ. Assume $f(\rho_1 \cdot f(\rho_2))$ rejects ψ, i.e. $f(\rho_2) \wedge \chi \vdash \neg\psi$. This implies $\theta \wedge f(\rho_2) \wedge \chi \vdash \neg\psi$ and hence $f(\rho_1 \cdot \theta \wedge f(\rho_2))$ also rejects ψ. However, if $f(\rho_1 \cdot f(\rho_2))$ accepts ψ, from $f(\rho_1 \cdot \rho_2) \nvdash \neg\theta$ we know $f(\rho_2) \wedge \chi \wedge \psi \nvdash \neg\theta$ and hence $\theta \wedge f(\rho_2) \wedge \chi \nvdash \neg\psi$. And so, $f(\rho_1 \cdot \theta \wedge f(\rho_2))$ also accepts ψ. ■

Proposition A.2 is a powerful property. If σ is any sequence such that $f(\sigma) \nvdash \neg\theta$, then it means that *wherever* θ is inserted in σ, f will return the same sentence modulo logical equivalence when

applied to this expanded sequence, namely $f(\sigma)\wedge\theta$. This fact will prove useful, for example, when proving Proposition 4.1. In particular we get the following:

PROPOSITION A.3
(vii) If $f(\sigma)\nvdash\neg\theta$ then $f(\sigma\cdot\theta)\equiv f(\sigma)\wedge\theta$.
(viii) If $f(\sigma\cdot\varphi)\nvdash\neg\theta$ then $f(\sigma\cdot(\varphi\wedge\theta))\equiv f(\sigma\cdot\varphi)\wedge\theta$.

PROOF. *(vii)* follows immediately as a special case of Proposition A.2. For (viii) first note that from (vii) we get if $f(\sigma\cdot\varphi)\nvdash\neg\theta$ then $f(\sigma\cdot\varphi\cdot\theta)\equiv f(\sigma\cdot\varphi)\wedge\theta$. But the left-hand side here is just $f(\sigma\cdot f(\varphi,\theta))$, and since $f(\sigma\cdot\varphi)\vdash\varphi$ we know $\varphi\nvdash\neg\theta$. Hence $f(\varphi,\theta)=\varphi\wedge\theta$ so the left-hand side is $f(\sigma\cdot(\varphi\wedge\theta))$, giving the desired result. ■

If we look at properties (ii)–(iv) and (vii) and (viii) above, we can recognize a close correspondence between them and the list of AGM postulates for belief revision. More precisely, from a fixed sequence σ, if we set $K=Cn(f(\sigma))$ (where Cn is the operation of closure under logical entailment) and define a revision operator $*$ for K by setting $K*\alpha=Cn(f(\sigma\cdot\alpha))$ for all $\alpha\in L$, then $*$ forms an AGM revision operator for K. Given the correspondence between AGM revision and the theory of *rational inference* relations (see, e.g. [28]), we can also formulate this in another way: If we denote by $\Rightarrow_\sigma$ the binary relation over L defined by $\alpha\Rightarrow_\sigma\beta$ iff $f(\sigma\cdot\alpha)\vdash\beta$, then $\Rightarrow_\sigma$ forms a rational inference relation which is consistency-preserving (i.e. $\alpha\Rightarrow_\sigma\bot$ implies $\alpha\equiv\bot$) [5, 10, 20]. This latter viewpoint will be especially useful in proving the results from Section 3. This link between f and AGM revision means we can in principle focus for the most part on a particular subset of the set of all sequences. If $\alpha_1\vdash\alpha_2\cdots\vdash\alpha_m$ then we shall say the sequence $(\alpha_m,\ldots,\alpha_1)$ is a *logical chain*.

PROPOSITION A.4
For any sequence σ there exists a logical chain $\rho=(\alpha_m,\ldots,\alpha_1)$ such that $\alpha_m\equiv\top$ and, for all $\lambda\in L$, $f(\sigma\cdot\lambda)\equiv f(\rho\cdot\lambda)$.

PROOF. From established results in AGM theory, we know any revision operator satisfying the AGM postulates can be represented by a logical chain. To be precise, for any belief set K and AGM revision operator $*$ for K, we know there always exists a logical chain $(\alpha_m,\ldots,\alpha_1)$, with $Cn(\alpha_1)=K$ and $\alpha_m\equiv\top$, such that for all $\theta\in L$,

$$K*\theta=\begin{cases}Cn(\theta\wedge\alpha_{r(\theta)}) & \text{if } \theta\not\equiv\bot\\ Cn(\bot) & \text{otherwise.}\end{cases}$$

where $r(\theta)\stackrel{\text{def}}{=}\min\{i\,|\,\theta\wedge\alpha_i \text{ is consistent}\}$. The α_i basically correspond to the spheres in Grove's famous 'systems-of-spheres' representation of AGM revision [13]. Given all this, and given the correspondence between f and AGM revision described above, we know that for our given sequence σ, there must exist a logical chain $(\alpha_m,\ldots,\alpha_1$ such that $\alpha_1\equiv f(\sigma)$ and $\alpha_m\equiv\top$, and such that for all λ,

$$f(\sigma\cdot\lambda)\equiv\begin{cases}\lambda\wedge\alpha_{r(\lambda)} & \text{if } \lambda\not\equiv\bot\\ \bot & \text{otherwise.}\end{cases}$$

Let $\rho=(\alpha_m,\ldots,\alpha_1)$ be the logical chain. Then it can be checked that, for all $\lambda\in L$, $f(\rho\cdot\lambda)\equiv f(\sigma\cdot\lambda)$ as required. ■

Given our definition of $Bel([\rho,\blacktriangle])$ in terms of f an immediate corollary of this result is that for any epistemic state $[\sigma,\blacktriangle]$, there is a logical chain ρ with weakest element $\top$ such that, for all $\lambda\in L$, $Bel([\sigma,\blacktriangle]*\lambda)\equiv Bel([\rho,\blacktriangle]*\lambda)$.

The next property will be used in the proofs of Propositions 4.1 and 4.3.

LEMMA A.5
(i) If $f(\rho)\vdash\varphi$ then $f(\varphi\to\theta\cdot\rho)\equiv f(\theta\cdot\rho)$
(ii) If $f(\rho)\vdash\neg\varphi$ then $f(\varphi\to\theta\cdot\rho)\equiv f(\rho)$

PROOF. (i). We have $f(\varphi\to\theta\cdot\rho)\equiv f(\varphi\to\theta\cdot f(\rho))\equiv f(f(\rho)\wedge(\varphi\to\theta)\cdot f(\rho))$. If $f(\rho)\vdash\varphi$ then $f(\rho)\wedge(\varphi\to\theta)\equiv f(\rho)\wedge\theta$ so $f(f(\rho)\wedge(\varphi\to\theta)\cdot f(\rho))\equiv f(f(\rho)\wedge\theta\cdot f(\rho))\equiv f(\theta\cdot f(\rho))\equiv f(\theta\cdot\rho)$ as required.
(ii). If $f(\rho)\vdash\neg\varphi$ then $f(\varphi\to\theta\cdot f(\rho))=f(\rho)$ by Proposition A.1 (vi). ■

Finally we have the following property, which will be relied upon when proving Theorems 3.6 and 3.7.

LEMMA A.6
Either $f(\sigma\cdot\rho\cdot\blacktriangle)\vdash\neg f(\sigma\cdot\blacktriangle)$ or $f(\sigma\cdot\rho\cdot\blacktriangle)\vdash f(\sigma\cdot\blacktriangle)$

PROOF. $f(\sigma\cdot\rho\cdot\blacktriangle)\equiv f(\sigma\cdot\blacktriangle\cdot\rho\cdot\blacktriangle)=f(\sigma\cdot\blacktriangle\cdot f(\rho\cdot\blacktriangle))$. The first equivalence is due to Proposition A.2. $\blacktriangle$ will be entailed, so adding it somewhere in the sequence will have no impact. Now, either $f(\sigma\cdot\blacktriangle)\wedge f(\rho\cdot\blacktriangle)$ is consistent or it is not. In the first case by Proposition A.3 (vii) we get $f(\sigma\cdot\blacktriangle\cdot f(\rho\cdot\blacktriangle))\equiv f(\sigma\cdot\blacktriangle)\wedge f(\rho\cdot\blacktriangle)\vdash f(\sigma\cdot\blacktriangle)$, while in the second case we get $f(\sigma\cdot\blacktriangle\cdot f(\rho\cdot\blacktriangle))\vdash f(\rho\cdot\blacktriangle)\vdash\neg f(\sigma\cdot\blacktriangle)$. ■

A.2 Proofs from Section 3

Recall (from just before Definition 3.2) that, given we have constructed the sequence $\rho_R(\mathcal{C},\mathcal{N})=(\alpha_m,\ldots,\alpha_0)$ corresponding to the rational closure of $\mathcal{C}$ and $\mathcal{N}$ according to Definition 3.1, the rational closure itself is the binary relation $\Rightarrow_R$ given by

$$\lambda\Rightarrow_R\chi \text{ iff } \alpha_m\vdash\neg\lambda \text{ or } f(\alpha_m,\ldots,\alpha_0,\lambda)\vdash\chi.$$

The proof of our next proposition will make use of the following property of $\Rightarrow_R$:

LEMMA A.7 ([5, 20])
If $\theta\Rightarrow_R\bot$ then $\theta\Rightarrow\bot$ for *every* rational inference relation $\Rightarrow$ satisfying $\mathcal{C}$ and $\mathcal{N}$.

PROPOSITION 3.4
Let $o\in O$ and $\blacktriangle\in L$, and let $\rho_R(o,\blacktriangle)=(\alpha_m,\ldots,\alpha_0)$ be the rational prefix of o with respect to $\blacktriangle$. Then
(i) if $\blacktriangle\not\equiv\bot$ and $\alpha_m\equiv\top$ then $[\rho_R(o,\blacktriangle),\blacktriangle]$ is an explanation for o.
(ii) if $\blacktriangle\equiv\bot$ or $\alpha_m\not\equiv\top$ then $\blacktriangle$ is not an o-acceptable core.

PROOF. Suppose $o=\langle(\varphi_1,\theta_1,D_1),\ldots,(\varphi_n,\theta_n,D_n)\rangle$. For each $i=1,\ldots,n$ we will let ι_i denote the sequence of first i inputs $(\phi_1,\ldots,\phi_i)$.
(i) Suppose $\blacktriangle\not\equiv\bot$ and $\alpha_m\equiv\top$. By the definition of explanation we need to show two things: (a) $\blacktriangle\not\equiv\bot$, and (b) for all $i=1,\ldots,n$, $f(\rho_R(o,\blacktriangle)\cdot f(\iota_i\cdot\blacktriangle))\vdash\theta_i$ and $f(\rho_R(o,\blacktriangle)\cdot f(\iota_i\cdot\blacktriangle))\nvdash\delta$ for all $\delta\in D_i$. Condition (a) we are already given, so it remains to prove (b). Since the rational closure relation $\Rightarrow_R$ satisfies all the positive conditionals ($\mathcal{C}$) and none of the negative ones ($\mathcal{N}$) we know, for each i, $f(\iota_i\cdot\blacktriangle)\Rightarrow_R\theta_i$ and $f(\iota_i\cdot\blacktriangle)\not\Rightarrow_R\delta$ for all $\delta\in D_i$. Now, for *any* sentence χ we have $f(\iota_i\cdot\blacktriangle)\Rightarrow_R\chi$ iff either $\alpha_m\vdash\neg f(\iota_i\cdot\blacktriangle)$ or $f(\rho_R(o,\blacktriangle)\cdot f(\iota_i\cdot\blacktriangle))\vdash\chi$. But since $\alpha_m\equiv\top$ and $\blacktriangle\not\equiv\bot$, the first disjunct here cannot hold. Hence $f(\iota_i\cdot\blacktriangle)\Rightarrow_R\chi$ iff $f(\rho_R(o,\blacktriangle)\cdot f(\iota_i\cdot\blacktriangle))\vdash\chi$. In particular, $f(\rho_R(o,\blacktriangle)\cdot f(\iota_i\cdot\blacktriangle))\vdash\theta_i$ and $f(\rho_R(o,\blacktriangle)\cdot f(\iota_i\cdot\blacktriangle))\nvdash\delta$ for all $\delta\in D_i$ as required.
(ii) First, if $\blacktriangle\equiv\bot$ then $\blacktriangle$ cannot be o-acceptable by definition of explanation. So assume instead $\alpha_m\not\equiv\top$. If $\blacktriangle$ was o-acceptable then there would be some ρ such that $[\rho,\blacktriangle]$ explained o. Let $\Rightarrow_\rho$ be the consistency-preserving rational inference relation corresponding to ρ (see the paragraph just before

Proposition A.4). Since $[\rho, \blacktriangle]$ explains o, $\Rightarrow_\rho$ satisfies all the positive conditionals ($\mathcal{C}$) and none of the negative ones ($\mathcal{N}$). Since $\alpha_m \not\equiv \top$, i.e. $\neg\alpha_m$ is consistent, and $\Rightarrow_\rho$ is consistency-preserving we have $\neg\alpha_m \not\Rightarrow_\rho \bot$. Hence, by Lemma A.7, $\neg\alpha_m \not\Rightarrow_R \bot$ – contradiction. Hence also in this case $\blacktriangle$ cannot be o-acceptable. ■

THEOREM 3.6
$\rho_R(o, \blacktriangle) \preceq_1 \sigma$ *for all* $\sigma \in \Sigma$.

PROOF. By Proposition A.4 we can restrict our attention to logical chains. To ease notation, let us denote $\rho_R(o, \blacktriangle)$ in this proof by just $\rho_R = (\alpha_m, \ldots, \alpha_0)$, $\sigma = (\beta_l, \ldots, \beta_0)$ such that $[\sigma, \blacktriangle]$ explains o. Both sequences are logical chains and $\alpha_m \equiv \top \equiv \beta_l$. Again, we abbreviate $(\varphi_1, \ldots, \varphi_i)$ by ι_i.

Also, let us introduce the following notation: Given any $\lambda \in L$, $rank_{\rho_R}(\lambda) \stackrel{\text{def}}{=} \min\{k \mid \lambda \wedge \alpha_k \text{ is consistent}\}$ ($\stackrel{\text{def}}{=} \infty$ if no such k exists. Though note that since $\alpha_m \equiv \top$ this can happen only if $\lambda \equiv \bot$). Analogously, $rank_\sigma(\lambda) \stackrel{\text{def}}{=} \min\{k \mid \lambda \wedge \beta_k \text{ is consistent}\}$ ($\stackrel{\text{def}}{=} \infty$ if no such k exists). Recall that as ρ_R is a logical chain we have for any λ such that $\lambda \not\equiv \bot$, $f(\rho_R \cdot \lambda) \equiv \lambda \wedge \alpha_s$ where $s = rank_{\rho_R}(\lambda)$ (and analogously for σ).

To show $\rho_R \preceq_1 \sigma$ we must prove that for any $i \in \{0, \ldots, n\}$, if $Bel_j^{\rho_R} \equiv Bel_j^\sigma$ for all $j < i$ then $Bel_i^\sigma \vdash Bel_i^{\rho_R}$. So fix $i \in \{0, \ldots, n\}$ and assume $Bel_j^{\rho_R} \equiv Bel_j^\sigma$ for all $j < i$. We have $Bel_i^{\rho_R} = f(\rho_R \cdot \iota_i \cdot \blacktriangle) = f(\rho_R \cdot f(\iota_i \cdot \blacktriangle))$. As $\blacktriangle$ — and consequently $f(\iota_i \cdot \blacktriangle)$ — is consistent, we have $f(\rho_R \cdot f(\iota_i \cdot \blacktriangle)) \equiv f(\iota_i \cdot \blacktriangle) \wedge \alpha_s$, where $s = rank_{\rho_R}(f(\iota_i \cdot \blacktriangle))$. What does α_s look like? Well, by construction of ρ_R, we know

$$\alpha_s \equiv \bigwedge_{k \in I} \big(f(\iota_k \cdot \blacktriangle) \to \theta_k\big) \wedge \bigwedge_{k \in J} \big(\neg f(\iota_k \cdot \blacktriangle)\big),$$

where $I = \{k \mid 1 \le k \le n \text{ and } rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) \ge s\}$ and $J = \{k \in I \mid \bigwedge_{i' \in I} \big(f(\iota_{i'} \cdot \blacktriangle) \to \theta_{i'}\big) \wedge \bigwedge_{k' \prec_e k} \big(\neg f(\iota_{k'} \cdot \blacktriangle)\big) \wedge f(\iota_k \cdot \blacktriangle) \vdash \psi$ for some negative conditional $f(\iota_k \cdot \blacktriangle) \Rightarrow \psi\}$. $\prec_e$ is a total order on the indices in J, indicating in which order the corresponding negative conditionals become exceptional in the least fixpoint calculation of the rational prefix construction.

Thus we have obtained

$$Bel_i^{\rho_R} \equiv f(\iota_i \cdot \blacktriangle) \wedge \bigwedge_{k \in I} \big(f(\iota_k \cdot \blacktriangle) \to \theta_k\big) \wedge \bigwedge_{k \in J} \big(\neg f(\iota_k \cdot \blacktriangle)\big)$$

Hence to show $Bel_i^\sigma \vdash Bel_i^{\rho_R}$ we need:

(a) $Bel_i^\sigma \vdash f(\iota_i \cdot \blacktriangle)$.
(b) $Bel_i^\sigma \vdash \big(f(\iota_k \cdot \blacktriangle) \to \theta_k\big)$ for all $k \in I$, equivalently
$Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) \vdash \theta_k$, for all $k \in I$
(c) $Bel_i^\sigma \vdash \neg f(\iota_k \cdot \blacktriangle)$, for all $k \in J$.

We show each of these in turn.
(a). $Bel_i^\sigma = f(\sigma \cdot \iota_i \cdot \blacktriangle) = f(\sigma \cdot f(\iota_i \cdot \blacktriangle)) \vdash f(\iota_i \cdot \blacktriangle)$.
(b). Let $k \in I$. $Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) \equiv \beta_t \wedge f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle)$, $t = rank_\sigma(f(\iota_i \cdot \blacktriangle))$. If this is inconsistent, we trivially get the desired conclusion. So suppose $Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle)$ is consistent, which implies $f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle)$ is consistent. We have to consider two cases $i \le k$ and $i > k$.
Case $i \le k$: Lemma A.6 tells us $f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$. Hence $\beta_t \wedge f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle) \equiv \beta_t \wedge f(\iota_k \cdot \blacktriangle)$. We claim that $rank_\sigma(f(\iota_k \cdot \blacktriangle)) = t$ which implies $\beta_t \wedge f(\iota_k \cdot \blacktriangle) \equiv Bel_k^\sigma \vdash \theta_k$—yielding the desired result.

To prove our claim assume $rank_\sigma(f(\iota_k \cdot \blacktriangle)) = u < t$. Then $\beta_u \wedge f(\iota_k \cdot \blacktriangle) \nvdash \bot$, as $f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$ we know $\beta_u \wedge f(\iota_i \cdot \blacktriangle) \nvdash \bot$ contradicting that $f(\iota_i \cdot \blacktriangle)$ is inconsistent with all β_u where $u < t$ (definition of *rank*). $u > t$ is not possible as $\beta_t \wedge f(\iota_k \cdot \blacktriangle)$ is consistent and *rank* yields the smallest possible index. This proves our claim.
Case $i > k$: Lemma A.6 tells us $f(\iota_i \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$. By Proposition A.3 $Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) = f(\sigma \cdot f(\iota_i \cdot \blacktriangle)) \wedge f(\iota_k \cdot \blacktriangle)$ is equivalent to $f(\sigma \cdot (f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle)))$. If we could show $f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle)$ is consistent then Proposition A.2 would yield

$$\begin{aligned} Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) &\equiv f(\sigma \cdot (f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle))) \\ &\equiv f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle) \text{by Proposition A.3} \\ &\vdash f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) = Bel_k^\sigma, \end{aligned}$$

and so, since $Bel_k^\sigma \vdash \theta_k$ as $[\sigma, \blacktriangle]$ explains o, we would get the required $Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) \vdash \theta_k$. To show $f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle)$ is indeed consistent, first note that, by the assumption $Bel_k^{\rho_R} \equiv Bel_k^\sigma$ for all $k < i$, we have

$$\begin{aligned} f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) &\equiv f(\rho_R \cdot f(\iota_k \cdot \blacktriangle)) \\ &\equiv f(\iota_k \cdot \blacktriangle) \wedge \alpha_x \end{aligned}$$

where $x = rank_{\rho_R}(f(\iota_k \cdot \blacktriangle))$. Hence

$$\begin{aligned} f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle) &\equiv f(\iota_k \cdot \blacktriangle) \wedge \alpha_x \wedge f(\iota_i \cdot \blacktriangle) \\ &\equiv \alpha_x \wedge f(\iota_i \cdot \blacktriangle). \end{aligned}$$

Now since $k \in I$ we know $x \geq s$, hence $\alpha_s \vdash \alpha_x$. We know already $\alpha_s \wedge f(\iota_i \cdot \blacktriangle)$ is consistent. Thus it follows that $\alpha_x \wedge f(\iota_i \cdot \blacktriangle)$ is consistent and so $f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle)$ is consistent as required.
(c). We know by construction of the rational prefix that we can order the elements of J using a total order $\prec_e$. For a $k \in J$ there is a negative conditional $f(\iota_k \cdot \blacktriangle) \Rightarrow \psi$ such that

$$\bigwedge_{j \in I} \big(f(\iota_j \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \Big(\bigwedge_{k' \prec_e k} \neg f(\iota_{k'} \cdot \blacktriangle)\Big) \wedge f(\iota_k \cdot \blacktriangle) \vdash \psi.$$

We will prove $Bel_i^\sigma \vdash \neg f(\iota_k \cdot \blacktriangle)$ iteratively ordering the k according to $\prec_e$. So assume $Bel_i^\sigma \vdash \neg f(\iota_k \cdot \blacktriangle)$ for all $k' \prec_e k$.

Hence $Bel_i^\sigma \vdash \bigwedge_{j \in I} \big(f(\iota_j \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \bigwedge_{k' \prec_e k} \big(\neg f(\iota_k \cdot \blacktriangle)\big)$. Now assume $Bel_i^\sigma \nvdash \neg f(\iota_k \cdot \blacktriangle)$ and so $f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle)$ is consistent.

Recall, $Bel_i^{\rho_R} = f(\iota_i \cdot \blacktriangle) \wedge \alpha_s$. As $Bel_i^{\rho_R}$ is consistent we know $Bel_i^{\rho_R} \nvdash \neg f(\iota_i \cdot \blacktriangle)$. This implies $i \notin J$ (cf. the structure of α_s — if $i \in J$ then $\alpha_s \vdash \neg f(\iota_i \cdot \blacktriangle)$). Hence, we only need to consider the cases $i > k$ and $i < k$.
Case $i > k$: $rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) = u > s$ as $k \in J$ (since $\alpha_s \vdash \neg f(\iota_k \cdot \blacktriangle)$ for $k \in J$). Lemma A.6 again tells us $f(\iota_i \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$ implying $rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) = u \leq s$. This is because $rank_{\rho_R}(f(\iota_i \cdot \blacktriangle)) = s$ and hence α_s must be consistent with $f(\iota_k \cdot \blacktriangle)$. So this case is impossible, as well.
Case $i < k$: Lemma A.6 tells us $f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$. $Bel_i^\sigma = \beta_t \wedge f(\iota_i \cdot \blacktriangle)$ with $t = rank_\sigma(f(\iota_i \cdot \blacktriangle))$.

We claim $Bel_k^\sigma = \beta_t \wedge f(\iota_k \cdot \blacktriangle)$. Note for all $u < t, \beta_u \wedge f(\iota_i \cdot \blacktriangle)$ is inconsistent, hence for all $u < t$, $\beta_u \wedge f(\iota_k \cdot \blacktriangle)$ is inconsistent ($f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$). Further $Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) \nvdash \bot$, implying that $\beta_t \wedge f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle) \nvdash \bot$, hence $\beta_t \wedge f(\iota_k \cdot \blacktriangle) \nvdash \bot$. Consequently $rank_\sigma(f(\iota_k \cdot \blacktriangle)) = t$, proving the claim.

So $Bel_k^\sigma = \beta_t \wedge f(\iota_k \cdot \blacktriangle)$. This implies $Bel_k^\sigma \vdash Bel_i^\sigma$ ($f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$ and $Bel_i^\sigma = \beta_t \wedge f(\iota_i \cdot \blacktriangle)$).

So we know $Bel_k^\sigma \vdash \bigwedge_{j \in I} \big(f(\iota_j \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \bigwedge_{k' \prec_e k} \big(\neg f(\iota_k \cdot \blacktriangle)\big)$ and also $Bel_k^\sigma \vdash f(\iota_k \cdot \blacktriangle)$. From the definition of J, we know that there exists a negative conditional $f(\iota_k \cdot \blacktriangle) \Rightarrow \psi$ such that

$$\bigwedge_{j \in I} \big(f(\iota_j \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \bigwedge_{k' \prec_e k} \big(\neg f(\iota_k \cdot \blacktriangle)\big) \wedge f(\iota_k \cdot \blacktriangle) \vdash \psi.$$

So $Bel_k^\sigma \vdash \psi$. Hence σ cannot be a solution — contradiction — and consequently $Bel_i^\sigma \vdash \neg f(\iota_k \cdot \blacktriangle)$. ■

THEOREM 3.7
For all $\sigma \in \Sigma$, if $\sigma \preceq_1 \rho_R(o, \blacktriangle)$ then $\rho_R(o, \blacktriangle) \preceq_2 \sigma$.

PROOF. This is proved along exactly similar lines to the last part of Theorem 3.6. Again it suffices to restrict the argument to logical chains. We again use $\rho_R = (\alpha_m, \ldots, \alpha_0)$ to denote $\rho_R(o, \blacktriangle)$. Let $[\sigma, \blacktriangle]$ be an explanation for o and suppose $\sigma \preceq_1 \rho_R$. We already know by Theorem 3.6 that also $\rho_R \preceq_1 \sigma$. Taking these two inequalities together means we must have $Bel_i^\sigma \equiv Bel_i^{\rho_R}$ for *all* $i = 0, \ldots, n$. Now to show $\rho_R \preceq_2 \sigma$ choose any $\lambda \in L$. We must show $Bel([\sigma, \blacktriangle] * \varphi_1 * \cdots * \varphi_n * \lambda) \vdash Bel([\rho_R, \blacktriangle] * \varphi_1 * \cdots * \varphi_n * \lambda)$, i.e. $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle)$

We know $f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle) = f(\rho_R \cdot f(\iota \cdot \lambda \cdot \blacktriangle)) = \alpha_s \wedge f(\iota \cdot \lambda \cdot \blacktriangle)$, where $s = rank_{\rho_R}(f(\iota \cdot \lambda \cdot \blacktriangle))$ and

$$\alpha_s \equiv \bigwedge_{k \in I} \big(f(\iota_k \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \bigwedge_{k \in J} \big(\neg f(\iota_k \cdot \blacktriangle)\big),$$

where $I = \{k \mid 1 \leq k \leq n$ and $rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) \geq s\}$ and $J = \{k \in I \mid \bigwedge_{k \in I} \big(f(\iota_k \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \bigwedge_{k' \prec_e k} \big(\neg f(\iota_{k'} \cdot \blacktriangle)\big)$ $\wedge f(\iota_k \cdot \blacktriangle) \vdash \psi$ for some negative conditional $f(\iota_k \cdot \blacktriangle) \Rightarrow \psi\}$. So, again we have to prove

(a) $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash f(\iota \cdot \lambda \cdot \blacktriangle)$

(b) $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash \big(f(\iota_k \cdot \blacktriangle) \rightarrow \theta_k\big)$ for all $k \in I$, equivalently
$f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle) \vdash \theta_k$, for all $k \in I$

(c) $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash \neg f(\iota_k \cdot \blacktriangle)$, for all $k \in J$.

(a) and (b) are exactly as in the proof for Theorem 3.6. Note that for (b) only the case analogous to $i > k$ is possible. In order to show (c) take an arbitrary $k \in J$. If $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash \neg f(\iota_k \cdot \blacktriangle)$ we are done. So assume $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \nvdash \neg f(\iota_k \cdot \blacktriangle)$. Hence $f(\iota \cdot \lambda \cdot \blacktriangle)$ is consistent with $f(\iota_k \cdot \blacktriangle)$. Lemma A.6 now tells us that $f(\iota \cdot \lambda \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$. Consequently $f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$, but we already know $f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash \neg f(\iota_k \cdot \blacktriangle)$ as $k \in J$. So we get a contradiction as $f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle)$ must be is consistent ($\blacktriangle$ is). So it is impossible that $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \nvdash \neg f(\iota_k \cdot \blacktriangle)$. This concludes the proof. ■

A.3 Proofs from Section 4

PROPOSITION 4.1
If $\blacktriangle_1$ and $\blacktriangle_2$ are o-acceptable then so is $\blacktriangle_1 \vee \blacktriangle_2$.

PROOF. Assume $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and let $\blacktriangle_1$ and $\blacktriangle_2$ be two o-acceptable cores. Then there are $\rho_1 = (\beta_{11}, \ldots, \beta_{1m_1})$ and $\rho_2 = (\beta_{21}, \ldots, \beta_{2m_2})$ such that $[\rho_1, \blacktriangle_1]$ and $[\rho_2, \blacktriangle_2]$ explain o.

It suffices to show there is a ρ such that $[\rho, \blacktriangle_1 \vee \blacktriangle_2]$ explains o. We will show that

$$\rho = (\neg\blacktriangle_1 \rightarrow \beta_{21}, \ldots, \neg\blacktriangle_1 \rightarrow \beta_{2m_2}, \blacktriangle_1 \rightarrow \beta_{11}, \ldots, \blacktriangle_1 \rightarrow \beta_{1m_1}, \blacktriangle_1)$$

is such a sequence. In fact, we will show that $Bel([\rho, \blacktriangle_1 \vee \blacktriangle_2] * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_1, \blacktriangle_1] * \varphi_1 * \ldots * \varphi_i)$ or $Bel([\rho, \blacktriangle_1 \vee \blacktriangle_2] * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_2, \blacktriangle_2] * \varphi_1 * \ldots * \varphi_i)$ for all $1 \leq i \leq n$. Then the proposition immediately follows as $[\rho_1, \blacktriangle_1]$ and $[\rho_2, \blacktriangle_2]$ are explanations for o. Fixing an i we show $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)) \equiv f(\rho_1 \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_1))$ or $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)) \equiv f(\rho_2 \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_2))$.

We claim that $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1)$ or $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \vdash \neg\blacktriangle_1$ and $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle_2)$. The $\blacktriangle_1$ in the very front is the last element of ρ. So before considering the implications in ρ which are constructed from the sentences in ρ_1 and ρ_2 we have collected a sentence which is equivalent to one that has been collected in the original cases.

Lemma A.5 then tells us how to treat the implications in ρ with respect to the original sentences in ρ_1 and ρ_2. In the first case where $\blacktriangle_1$ is entailed all $\blacktriangle_1 \rightarrow \beta_{1j}$ are treated exactly like the β_{1j} from ρ_1 and the $\neg\blacktriangle_1 \rightarrow \beta_{2k}$ from ρ_2 can be ignored. As a consequence, we get the same beliefs as for the epistemic state $[\rho_1, \blacktriangle_1]$. In the second case, $\neg\blacktriangle_1$ is entailed and hence all $\blacktriangle_1 \rightarrow \beta_{1j}$ from ρ_1 can be ignored and all $\neg\blacktriangle_1 \rightarrow \beta_{2k}$ are treated exactly like the β_{2k} from ρ_2 and hence we get the same beliefs as for the epistemic state $[\rho_2, \blacktriangle_2]$

For proving the claim, let us first assume $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \vdash \blacktriangle_1$ then we must have $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \nvdash \neg\blacktriangle_1$. By Proposition A.2 $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1, \blacktriangle_1 \vee \blacktriangle_2)$ which is equivalent to $f(\varphi_1, \ldots, \varphi_i, f(\blacktriangle_1, \blacktriangle_1 \vee \blacktriangle_2))$ and as $\blacktriangle_1 \vdash \blacktriangle_1 \vee \blacktriangle_2$ and $\blacktriangle_1$ is consistent we have $f(\blacktriangle_1, \blacktriangle_1 \vee \blacktriangle_2) \equiv \blacktriangle_1$ and so this is equivalent to $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1)$ as claimed.

Now assume $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \nvdash \blacktriangle_1$. Consequently $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \vdash \neg\blacktriangle_1$ and $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)$. If we can show $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)$ and $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_2)$ collect the same elements from $(\varphi_1, \ldots, \varphi_i)$ we are done (let χ denote the conjunction of elements collected from that sequence, then $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_2) = \chi \wedge \blacktriangle_2$ and $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) = \chi \wedge (\blacktriangle_1 \vee \blacktriangle_2)$ but as this entails $\neg\blacktriangle_1$ it is equivalent to $\chi \wedge \blacktriangle_2$).

To see that the two indeed collect the same elements from $(\varphi_1, \ldots, \varphi_i)$ assume they have collected the same elements φ_i down to φ_{j+1} their conjunction being denoted by χ. Now they are considering φ_j. If $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_2)$ accepts φ_j, i.e. $\chi \wedge \blacktriangle_2 \wedge \varphi_j$ is consistent then $\chi \wedge (\blacktriangle_1 \vee \blacktriangle_2) \wedge \varphi_j$ is consistent and hence $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)$ accepts φ_j as well. If $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_2)$ rejects φ_j then $\chi \wedge \blacktriangle_2 \vdash \neg\varphi_j$ so $\chi \wedge \varphi_j \vdash \neg\blacktriangle_2$ and as a consequence $\chi \wedge \varphi_j \wedge (\blacktriangle_1 \vee \blacktriangle_2) \vdash \blacktriangle_1$. Hence, $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)$ must reject φ_j as well, since $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \vdash \neg\blacktriangle_1$. ■

PROPOSITION 4.2
Suppose $o \sqsubseteq_{\text{right}} o'$ or $o \sqsubseteq_{\text{left}} o'$. Then every o'-acceptable core is an o-acceptable core.

PROOF. Suppose $o \sqsubseteq_{\text{right}} o'$. Then it is easy to check that any explanation $[\rho, \blacktriangle]$ for o' is an explanation for o. Hence any o'-acceptable core is automatically o-acceptable. Suppose $o \sqsubseteq_{\text{left}} o'$, say $o' = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and $o = \langle(\varphi_j, \theta_j, D_j), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ for some $1 \leq j \leq n$, and let $\blacktriangle$ be an o'-acceptable core. Then $[\rho, \blacktriangle]$ explains o' for some ρ. In this case it is easy to check $[\rho \cdot (\varphi_1, \ldots, \varphi_{j-1}), \blacktriangle]$ then provide an explanation for o. So $\blacktriangle$ is also o-acceptable. ■

PROPOSITION 4.3
Let $\blacktriangle : O \rightarrow L$ be any function which returns a sentence given any $o \in O$. Then the following are equivalent:
(i) $\blacktriangle$ satisfies Acceptability, Consistency and Right Monotony.
(ii) $\blacktriangle$ satisfies Acceptability, Consistency and Left Monotony.
(iii) $\blacktriangle(o) \equiv \blacktriangle_{\vee}(o)$ for all $o \in O$.

PROOF. Propositions 4.1 and 4.2 yield (iii) → (i) and (iii) → (ii). It remains to show (i) → (iii) and (ii) → (iii).
(i)→ (iii):
Let ▲ satisfy Acceptability, Consistency and Right Monotony. First, for *any* $o \in O$ we know already $\blacktriangle(o) \vdash \blacktriangle_\vee(o)$. If $\blacktriangle(o) \equiv \bot$ then this is clear, while if $\blacktriangle(o) \not\equiv \bot$ then there must exist at least one o-acceptable core by Consistency, and so $\blacktriangle(o)$ is itself o-acceptable by Acceptability. Hence in this case $\blacktriangle(o) \vdash \blacktriangle_\vee(o)$ by definition of $\blacktriangle_\vee$. Now, to show (iii) assume for contradiction there is some $o \in O$ such that $\blacktriangle(o) = \psi$ and $\blacktriangle_\vee(o) = \phi \not\equiv \psi$. We just proved $\psi \vdash \phi$, so we must have $\phi \nvdash \psi$.

Now consider $o' = o \cdot (\neg\psi, \neg\psi, \emptyset)$. Since $\phi \nvdash \psi$ we know $\phi \not\equiv \bot$ so ϕ is o-acceptable since we know $\blacktriangle_\vee$ satisfies Acceptability and Consistency. Hence there is some ρ such that $[\rho, \phi]$ explains o. It can be checked that also $[\rho, \phi]$ explains o' (since $\phi \nvdash \psi$, core ϕ does not prevent $\neg\psi$ from being introduced into the belief set upon receiving it, which is the only condition needed to satisfy the additional observation $(\neg\psi, \neg\psi, \emptyset)$). Hence, there is an o'-acceptable core, and so $\blacktriangle(o')$ is an o'-acceptable core by Acceptability. In particular (assuming the full sequence of revision inputs in o is $(\varphi_1, \ldots, \varphi_n)$), this means $f(\rho' \cdot (\varphi_1, \ldots, \varphi_n) \cdot \neg\psi \cdot \blacktriangle(o')) \vdash \neg\psi$ for some prefix ρ'. But since $o \sqsubseteq_{\text{right}} o'$ we know $\blacktriangle(o') \vdash \psi$ by Right Monotony. Hence also

$$f(\rho' \cdot (\varphi_1, \ldots, \varphi_n) \cdot \neg\psi \cdot \blacktriangle(o')) \vdash \blacktriangle(o') \vdash \psi.$$

so $f(\rho' \cdot (\varphi_1, \ldots, \varphi_n) \cdot \neg\psi \cdot \blacktriangle(o')) \equiv \bot$. But this can only happen if $\blacktriangle(o') \equiv \bot$, which contradicts $\blacktriangle(o')$ being o'-acceptable. Hence there can be no o such that $\blacktriangle(o) \not\equiv \blacktriangle_\vee(o)$ as required.
(ii) → (iii):
Now let ▲ satisfy Acceptability, Consistency and Left Monotony. To show (iii), again assume for contradiction there is some $o \in O$ such that $\blacktriangle(o) = \psi$ and $\blacktriangle_\vee(o) = \phi \not\equiv \psi$. As above this means we have $\psi \vdash \phi$ and $\phi \nvdash \psi$, but we further have $\phi \nvdash \neg\psi$ (otherwise $\psi \vdash \neg\psi$, but ψ is consistent).

Consider $o' = \langle(\neg\psi, \neg\psi, \emptyset), (\psi, \psi, \emptyset)\rangle \cdot o$. Since $o \sqsubseteq_{\text{left}} o'$ we know $\blacktriangle(o') \vdash \psi$ by Left Monotony. But this means $\blacktriangle(o')$ cannot be o'-acceptable, since if it were it would be $\langle(\neg\psi, \neg\psi, \emptyset)\rangle$-acceptable by Proposition 4.2 (since $\langle(\neg\psi, \neg\psi, \emptyset)\rangle \sqsubseteq_{\text{right}} o'$), but this is not possible. Since ▲ satisfies Acceptability, this in turn implies *no* o'-acceptable core can exist. But we will now show that in fact ϕ is an o'-acceptable core, giving the required contradiction.

As this proof is constructive, we need to look into the observation we assumed to exist—$o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and thus $o' = \langle(\neg\psi, \neg\psi, \emptyset), (\psi, \psi, \emptyset), (\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. Note that both ϕ and ψ are o-acceptable. For ϕ this follows since we assume $\phi \nvdash \psi$ and so ϕ is consistent, which means ϕ must be o-acceptable from the fact that $\blacktriangle_\vee$ satisfies Acceptability and Consistency. The o-acceptability of ψ follows in turn using the assumption ▲ satisfies Acceptability. Hence there exist sequences $\rho_1 = (\beta_{11}, \ldots, \beta_{1m_1})$ and $\rho_2 = (\beta_{21}, \ldots, \beta_{2m_2})$ such that $[\rho_1, \phi]$ and $[\rho_2, \psi]$ explain o. We will show that there is a sequence ρ such that $[\rho, \phi]$ explains o', namely

$$\rho = (\psi \to \beta_{21}, \ldots, \psi \to \beta_{2m_2}, \neg\psi \to \beta_{11}, \ldots, \neg\psi \to \beta_{1m_1}).$$

Note that ϕ explains the prefix $\langle(\neg\psi, \neg\psi, \emptyset), (\psi, \psi, \emptyset)\rangle$ of o' using *any* sequence. This is because that observation only requires $\neg\psi$ and ψ to be believed upon receiving them, but this is guaranteed as ϕ is consistent with both. ($\phi \nvdash \psi$ by assumption, while if $\phi \vdash \neg\psi$ then ψ is inconsistent since we established at the start that $\psi \vdash \phi$, but this contradicts ψ being o-acceptable.) We will show that for all the remaining inputs $\varphi_1, \ldots, \varphi_n$, $Bel([\rho, \phi] * \neg\psi * \psi * \varphi_1 * \ldots * \varphi_i) \equiv Bel([\rho_1, \phi] * \varphi_1 * \ldots * \varphi_i)$ or $Bel([\rho, \phi] * \neg\psi * \psi * \varphi_1 * \ldots * \varphi_i) \equiv Bel([\rho_2, \psi] * \varphi_1 * \ldots * \varphi_i)$ which then yields that $[\rho, \phi]$ indeed explains o'. The argument is basically identical to that in the proof for Proposition 4.1.

$f(\neg\psi,\psi,\varphi_1,\ldots,\varphi_i,\phi)\vdash\psi$ or $f(\neg\psi,\psi,\varphi_1,\ldots,\varphi_i,\phi)\vdash\neg\psi$. This is because ψ is an element of the sequence and hence is collected yielding the first case or rejected yielding the second. For the first case Proposition A.2 tells us $f(\neg\psi,\psi,\varphi_1,\ldots,\varphi_i,\phi)\equiv f(\neg\psi,\varphi_1,\ldots,\varphi_i,\psi,\phi)$. But as $\psi\vdash\phi$, which implies $f(\psi,\phi)\equiv\psi$, and $f(\neg\psi,\varphi_1,\ldots,\varphi_i,\psi,\phi)\equiv f(\neg\psi,\varphi_1,\ldots,\varphi_i,f(\psi,\phi))$ we get $f(\neg\psi,\psi,\varphi_1,\ldots,\varphi_i,\phi)\equiv f(\neg\psi,\varphi_1,\ldots,\varphi_i,\psi)$. This definitely entails ψ, so the $\neg\psi$ in the beginning is irrelevant. Hence in this case $f(\neg\psi,\psi,\varphi_1,\ldots,\varphi_i,\phi)\equiv f(\varphi_1,\ldots,\varphi_i,\psi)$.

In other words, before ρ is processed a sentence has been constructed that is equivalent to that which has been collected before processing ρ_2. Lemma A.5 yields that the $\neg\psi\to\beta_{1j}$ can now be ignored when processing ρ and that the $\psi\to\beta_{2k}$ are treated exactly like the β_{2k} in ρ_2. Hence, in this case $Bel([\rho,\phi]*\neg\psi*\psi*\varphi_1*\ldots*\varphi_i)\equiv Bel([\rho_2,\psi]*\varphi_1*\ldots*\varphi_i)$ as claimed.

In the second case ($f(\neg\psi,\psi,\varphi_1,\ldots,\varphi_i,\phi)\vdash\neg\psi$), we know $f(\varphi_1,\ldots,\varphi_i,\phi)\vdash\neg\psi$ as otherwise the next sentence ψ would have to be accepted. But this means $f(\neg\psi,\psi,\varphi_1,\ldots,\varphi_i,\phi)\equiv f(\varphi_1,\ldots,\varphi_i,\phi)$, i.e. before ρ is processed a sentence has been constructed that is equivalent to that which has been collected before processing ρ_1. Lemma A.5 now yields that the $\psi\to\beta_{2j}$ can now be ignored when processing ρ and that the $\neg\psi\to\beta_{1k}$ are treated exactly like the β_{1k} in ρ_1 and hence, $Bel([\rho,\phi]*\neg\psi*\psi*\varphi_1*\ldots*\varphi_i)\equiv Bel([\rho_1,\phi]*\varphi_1*\ldots*\varphi_i)$. ■

A.4 Proofs from Section 5

LEMMA 5.1
Let ▲ and α_m be as after the calculation of $\rho_R(o,\blacktriangle)$. If $\alpha_m\not\equiv\top$ then $\blacktriangle\not\equiv\blacktriangle\wedge\alpha_m$.

PROOF. From Example 3.3 (i) we know if $\alpha_m\not\equiv\top$ then we must have $\blacktriangle\not\equiv\bot$. We have $\alpha_m\equiv\bigwedge U_m$ with $U_m=\tilde{\mathcal{C}}_m\cup\{\neg\lambda\mid\lambda\Rightarrow\chi\in\mathcal{N}_m\}$ and all the conditionals in $\mathcal{C}_m$ and $\mathcal{N}_m$ are exceptional for U_m otherwise the rational prefix construction would not have terminated. This means $U_m\vdash\neg\lambda$ for any conditional $\lambda\Rightarrow\chi\in\mathcal{C}_m\cup\mathcal{N}_m$. Such a conditional must exist as otherwise $U_m=\emptyset$ and $\alpha_m=\top$. Recall that $\lambda=f(\varphi_1,\ldots,\varphi_i,\blacktriangle)$ for some i, so $\lambda\vdash\blacktriangle$. Now assume $\blacktriangle\equiv\blacktriangle\wedge\alpha_m$, i.e. $\blacktriangle\vdash\alpha_m$. Then $\lambda\vdash\neg\lambda$ and hence λ is inconsistent, but this is possible only if $\blacktriangle\equiv\bot$—contradiction. Hence, $\blacktriangle\not\equiv\blacktriangle\wedge\alpha_m$. ■

LEMMA 5.2
Let $0<i\leq k$ and suppose $\rho_{i-1}=(\alpha_m,\ldots,\alpha_0)$. Then, for *any* o-acceptable core ▲′, if $\blacktriangle'\vdash\blacktriangle_{i-1}$ then $\blacktriangle'\vdash\alpha_m$.

PROOF. We will show that if $\blacktriangle'\nvdash\alpha_m$ then the rational prefix construction using that core will not be successful, i.e. $\bigwedge U'_{m'}\not\equiv\top$ and hence by Proposition 3.4 (ii) ▲′ is not o-acceptable — contradiction. We will abbreviate $\blacktriangle_{i-1}$ by ▲. The proposition is trivial for o-acceptable ▲, as then $\alpha_m\equiv\top$. Further $\blacktriangle\not\equiv\bot$ as that would contradict $\blacktriangle'\vdash\blacktriangle$ and ▲′ being o-acceptable.

So let ▲ be consistent and not o-acceptable. Then $\alpha_m=\bigwedge U_m\not\equiv\top$ which implies $\mathcal{C}_m\neq\emptyset$ and possibly $\mathcal{N}_m\neq\emptyset$. Let $I_{\mathcal{C}}=\{i\mid f(\iota_i\cdot\blacktriangle)\Rightarrow\theta_i\in\mathcal{C}_m\}$ and $I_{\mathcal{N}}=\{i\mid f(\iota_i\cdot\blacktriangle)\Rightarrow\delta\in\mathcal{N}_m\}$ be the index set of the ultimately exceptional positive and negative conditionals and let $\prec_e$ be the order on $I_{\mathcal{N}}$ in which the corresponding conditional became exceptional in the least fixpoint construction of U_m. So we know from the conditionals being exceptional for U_m that $U_m\equiv\bigwedge_{i\in I_{\mathcal{C}}}(f(\iota_i\cdot\blacktriangle)\to\theta_i)\wedge\bigwedge_{j\in I_{\mathcal{N}}}\neg f(\iota_j\cdot\blacktriangle)\vdash\bigwedge_{i\in I_{\mathcal{C}}}\neg f(\iota_i\cdot\blacktriangle)$ and $\forall j\in I_{\mathcal{N}}\exists\delta\in D_j\colon\bigwedge_{i\in I_{\mathcal{C}}}(f(\iota_i\cdot\blacktriangle)\to\theta_i)\wedge\bigwedge_{k\in I_{\mathcal{N}}\wedge k\prec_e j}\neg f(\iota_k\cdot\blacktriangle)\wedge f(\iota_j\cdot\blacktriangle)\vdash\delta$. We will use these entailments later in the proof and will refer back to them with (*).

First note $\alpha_m\equiv\bigwedge_{i\in I_{\mathcal{C}}}\neg f(\iota_i\cdot\blacktriangle)$. This is because $\mathcal{N}_m\subseteq\mathcal{C}_m$; if a negative conditional $f(\iota_i\cdot\blacktriangle)\Rightarrow\delta$ with index i is n-exceptional for U_m then the corresponding positive conditional $(f(\iota_i\cdot\blacktriangle)\Rightarrow\theta_i)$ must be

p-exceptional as $\neg f(\iota_i \cdot \blacktriangle)$ is added to U_m. We will now show that if $\blacktriangle' \nvdash \bigwedge_{i \in I_\mathcal{C}} \neg f(\iota_i \cdot \blacktriangle) \equiv \alpha_m$ (implying $\blacktriangle'$ to be consistent) then $\mathcal{C}'_{m'} \neq \emptyset$ and hence $\bigwedge U'_{m'} \not\equiv \top$ as claimed.

So let $\blacktriangle' \vdash \blacktriangle$, $\blacktriangle' \nvdash \bigwedge_{i \in I_\mathcal{C}} \neg f(\iota_i \cdot \blacktriangle)$ and $J = \{j \mid j \in I_\mathcal{C} \wedge \blacktriangle' \nvdash \neg f(\iota_j \cdot \blacktriangle)\}$, i.e. J is the set of indexes of ultimately exceptional conditionals whose antecedents remain consistent with the new core belief $\blacktriangle'$. We claim that

$$\bigwedge_{i \in J} (f(\iota_i \cdot \blacktriangle') \rightarrow \theta_i) \wedge \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle') \vdash \bigwedge_{i \in J} \neg f(\iota_i \cdot \blacktriangle')$$

and

$$\forall j \in J \cap I_\mathcal{N} \exists \delta \in D_j : \bigwedge_{i \in J} (f(\iota_i \cdot \blacktriangle') \rightarrow \theta_i) \wedge \bigwedge_{k \in J \cap I_\mathcal{N} \wedge k \prec_e j} \neg f(\iota_k \cdot \blacktriangle') \wedge f(\iota_j \cdot \blacktriangle') \vdash \delta$$

This means a conditional with an index $j \in J$ will be ultimately exceptional when using $\blacktriangle'$ as the core belief yielding $\mathcal{C}'_{m'} \neq \emptyset$. We know $\blacktriangle' \vdash \blacktriangle$, so $\blacktriangle' \wedge \blacktriangle \equiv \blacktriangle'$ and for $j \in J$ we have $f(\iota_j \cdot \blacktriangle') \nvdash \neg \blacktriangle'$, so using Proposition A.3 we get $f(\iota_j \cdot \blacktriangle') \equiv f(\iota_j \cdot \blacktriangle' \wedge \blacktriangle) \equiv f(\iota_j \cdot \blacktriangle) \wedge \blacktriangle'$. The corresponding equivalence does not hold for $f(\iota_j \cdot \blacktriangle')$, $j \notin J$. Here we have $f(\iota_j \cdot \blacktriangle) \wedge \blacktriangle' \vdash \bot$ so that $\bigwedge_{j \notin J} (f(\iota_j \cdot \blacktriangle) \wedge \blacktriangle' \rightarrow \theta_j)$ is a tautology. We start by proving $\bigwedge_{i \in J} (f(\iota_i \cdot \blacktriangle') \rightarrow \theta_i) \wedge \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle') \vdash \bigwedge_{i \in J} \neg f(\iota_i \cdot \blacktriangle')$.

$$\begin{aligned}
&\bigwedge_{i \in J} (f(\iota_i \cdot \blacktriangle') \rightarrow \theta_i) && \wedge && \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle') \\
\equiv &\bigwedge_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \wedge \blacktriangle' \rightarrow \theta_i) && \wedge && \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle') \\
\equiv &\blacktriangle' \rightarrow \bigwedge_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) && \wedge && \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle') \\
\vdash &\blacktriangle' \rightarrow \bigwedge_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) && \wedge && \blacktriangle' \rightarrow \left(\blacktriangle' \wedge \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle') \right)
\end{aligned}$$

As $\blacktriangle' \vdash \neg f(\iota_i \cdot \blacktriangle)$ for $i \in I_\mathcal{N} \setminus J$ and $f(\iota_j \cdot \blacktriangle') = f(\iota_j \cdot \blacktriangle)$ for $j \in J$

$$\begin{aligned}
\vdash &\blacktriangle' \rightarrow \bigwedge_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) && \wedge && \blacktriangle' \rightarrow \left(\bigwedge_{k \in I_\mathcal{N} \setminus J} \neg f(\iota_k \cdot \blacktriangle) \wedge \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle) \right) \\
\equiv &\blacktriangle' \rightarrow \bigwedge_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) && \wedge && \blacktriangle' \rightarrow \bigwedge_{j \in I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle) \\
\equiv &\blacktriangle' \rightarrow \left(\bigwedge_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \wedge \bigwedge_{j \in I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle) \right)
\end{aligned}$$

using (*) and $J \subseteq I_\mathcal{C}$

$$\begin{aligned}
&\vdash \blacktriangle' \rightarrow \left(\bigwedge_{i \in J} \neg f(\iota_i \cdot \blacktriangle) \right) \\
&\equiv \bigwedge_{i \in J} \neg \left(f(\iota_i \cdot \blacktriangle) \wedge \blacktriangle' \right) \\
&\equiv \bigwedge_{i \in J} \neg f(\iota_i \cdot \blacktriangle')
\end{aligned}$$

$\forall j \in J \cap I_\mathcal{N} \exists \delta \in D_j : \bigwedge_{i \in J} (f(\iota_i \cdot \blacktriangle') \rightarrow \theta_i) \wedge \bigwedge_{k \in J \cap I_\mathcal{N} \wedge k \prec_e j} \neg f(\iota_k \cdot \blacktriangle') \wedge f(\iota_j \cdot \blacktriangle') \vdash \delta$ is proved as follows. Let $j \in J \cap I_\mathcal{N}$ then

$$\bigwedge_{i \in J} (f(\iota_i \cdot \blacktriangle') \rightarrow \theta_i) \qquad \wedge \qquad \bigwedge_{k \in J \cap I_\mathcal{N} \wedge k \prec_e j} \neg f(\iota_k \cdot \blacktriangle') \wedge f(\iota_j \cdot \blacktriangle')$$

$$\equiv \bigwedge_{i\in I_{\mathcal{C}}} (f(\iota_i\cdot\blacktriangle)\wedge\blacktriangle'\to\theta_i) \quad\wedge \bigwedge_{k\in J\cap I_{\mathcal{N}}\wedge k\prec_e j} \neg f(\iota_k\cdot\blacktriangle')\wedge\ f(\iota_j\cdot\blacktriangle')$$

$$\equiv \blacktriangle'\to \bigwedge_{i\in I_{\mathcal{C}}} (f(\iota_i\cdot\blacktriangle)\to\theta_i) \wedge \bigwedge_{k\in J\cap I_{\mathcal{N}}\wedge k\prec_e j} \neg f(\iota_k\cdot\blacktriangle')\wedge\ f(\iota_j\cdot\blacktriangle')$$

as $f(\iota_i\cdot\blacktriangle')\equiv f(\iota_i\cdot\blacktriangle)\wedge\blacktriangle'$ for $i\in J$

$$\equiv [\blacktriangle'\to \bigwedge_{i\in I_{\mathcal{C}}} (f(\iota_i\cdot\blacktriangle)\to\theta_i)\wedge \bigwedge_{k\in J\cap I_{\mathcal{N}}\wedge k\prec_e j} \neg f(\iota_k\cdot\blacktriangle)\ \wedge\ f(\iota_j\cdot\blacktriangle)]\wedge\blacktriangle'$$

$$\vdash \bigwedge_{i\in I_{\mathcal{C}}} (f(\iota_i\cdot\blacktriangle)\to\theta_i) \quad\wedge \bigwedge_{k\in J\cap I_{\mathcal{N}}\wedge k\prec_e j} \neg f(\iota_k\cdot\blacktriangle)\ \wedge\ f(\iota_j\cdot\blacktriangle)\wedge\blacktriangle'$$

as $\blacktriangle'\vdash\neg f(\iota_i\cdot\blacktriangle)$ for $i\in I_{\mathcal{N}}\setminus J$

$$\vdash \bigwedge_{i\in I_{\mathcal{C}}} (f(\iota_i\cdot\blacktriangle)\to\theta_i) \quad\wedge \bigwedge_{k\in J\cap I_{\mathcal{N}}\wedge k\prec_e j} \neg f(\iota_k\cdot\blacktriangle)\ \wedge\ f(\iota_j\cdot\blacktriangle)\wedge \bigwedge_{l\in I_{\mathcal{N}}\setminus J} \neg f(\iota_l\cdot\blacktriangle)$$

$$\vdash \bigwedge_{i\in I_{\mathcal{C}}} (f(\iota_i\cdot\blacktriangle)\to\theta_i) \quad\wedge \bigwedge_{k\in I_{\mathcal{N}}\wedge k\prec_e j} \neg f(\iota_k\cdot\blacktriangle) \quad\wedge\ f(\iota_j\cdot\blacktriangle)$$

Using (*) we know that there is a $\delta\in D_j$ such that the above sentence entails δ which concludes the proof. ■

Received December 2007