Beyond the Rational Explanation

Alexander Nittka[1], Richard Booth[2]

[1] Department of Computer Science, University of Leipzig
Augustusplatz 10/11, 04109 Leipzig, Germany
nittka@informatik.uni-leipzig.de
[2] Department of Computer Science, Macquarie University
Sydney NSW 2109 Australia
ribooth@gmail.com

Abstract. In this paper, we generalise recent work on reconstructing an agent's epistemic state from observations about what it believed after receiving a series of revision inputs. We do so by also allowing to take into account information about what the agent did *not* believe after each revision step.

Keywords. belief revision, iterated revision, non-prioritised revision, non-monotonic reasoning, rational closure, rational explanation

1 Introduction

The problem of belief revision, i.e., of how an agent should modify its beliefs about the world given some new information which possibly contradicts its current beliefs, is by now a well-established research area in AI [7]. 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. In [2] we proposed to change the viewpoint and to cast ourselves in the role of an *observer* of the agent.

The scenario there is as follows. We are given some sequence $(\phi_1, \ldots, \phi_n)$ of revision inputs which a particular agent, hereafter $\mathcal{A}$, has received over a certain length of time and we are also given a sequence $(\theta_1, \ldots, \theta_n)$ with the interpretation that following the i^{th} input ϕ_i, $\mathcal{A}$ believed (at least) θ_i. 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 are *correct* (but possibly *partial*) descriptions of $\mathcal{A}$'s beliefs after each input. We are interested in trying to guess, on the basis of these two sequences, what the agent believed before receiving the first input ϕ_1, after each input ϕ_i $(i \leq n)$ and after a possible further input ϕ_{n+1}.

In this paper, we generalise this scenario. In addition to the inputs ϕ_i the agent received and information θ_i about what the agent believed afterwards, we will allow information about what the agent did *not* believe after receiving ϕ_i. It should be clear that it is not always possible to express this using a single

Dagstuhl Seminar Proceedings 05321
Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics
http://drops.dagstuhl.de/opus/volltexte/2005/332

sentence, if we think of believing neither p nor $\neg p$, for example. Hence, this additional information is encoded as a set of sentences rather than a sentence.

So, an observation $o = \langle(\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n), (D_1, \ldots, D_n)\rangle$ will contain a third sequence $(D_1, \ldots, D_n)$, each D_i being a (finite) set of sentences. The interpretation of an observation is that after receiving the first i inputs $\phi_1, \ldots, \phi_i$, $\mathcal{A}$ believes θ_i but does *not* believe any $\delta \in D_i$. The observations considered in [2] basically were special cases where $D_i = \emptyset$ for all i. In fact, most of the results presented in this paper subsume the corresponding results in [2].

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. We restrict the investigation to the special case of a framework for iterated non-prioritised revision, i.e., revision in which the new input is not necessarily always believed after revision, that has been studied in [1]. The idea behind it is that an agent's epistemic state 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 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. The 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 $o = \langle(\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n), (D_1, \ldots, D_n)\rangle$ we have made of $\mathcal{A}$.

We assume sentences $\phi_i, \theta_i, \lambda, \delta, \blacktriangle$, etc. are elements of some finitely-generated propositional language L. In our examples, p, q, r, s denote distinct propositional variables. The classical logical entailment relation between sentences 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 $\cdot$ on sequences denotes sequence concatenation. Let $O_n^{\pm}$ be the set of all observations $o = \langle(\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n), (D_1, \ldots, D_n)\rangle$ of length n. We denote by $O^{\pm} = \bigcup O_i^{\pm}$ the set of all possible observations. We denote by O the subset of $O^{\pm}$ where $D_i = \emptyset$ for all i, i.e. those observations that contain no information about what is not believed after a revision step. Note that this is the set considered in [2].

In Section 2 we will explain the model of the agent which we assume throughout the paper. The problem will be formally defined and solved in Section 3. We conclude with some discussion in Section 4.

2 Modelling the Agent

As indicated in the introduction, we follow [1] by assuming that, at any given moment in time, an agent's epistemic state is represented by a pair $[\rho, \blacktriangle]$. ([9,11]

also use sequences to represent epistemic states, but without core beliefs). In order to fully specify the agent's epistemic processes, we also need to formally specify *(i)* how the agent determines its set of beliefs $Bel([\rho, \blacktriangle])$ in any given state $[\rho, \blacktriangle]$, and *(ii)* how it incorporates new revision inputs into its epistemic state. Turning first to *(i)*, we can describe $Bel([\rho, \blacktriangle])$ neatly with the help of a function f, which takes as argument a non-empty sequence $\sigma = (\alpha_m, \ldots, \alpha_1)$ of sentences, and returns a sentence. Basically, $f(\sigma)$ 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 [15] and the "basic memory operator" of [9].)

Definition 1. *Let $\sigma = (\alpha_m, \ldots, \alpha_1)$ be a sequence of sentences.*
If $m = 1$ then $f(\sigma) = \alpha_1$.
If $m > 1$ then $f(\sigma) = \begin{cases} \alpha_m \wedge f(\alpha_{m-1}, \ldots, \alpha_1) & \text{if } \alpha_m \wedge f(\alpha_{m-1}, \ldots, \alpha_1) \nvdash \bot \\ f(\alpha_{m-1}, \ldots, \alpha_1) & \text{otherwise} \end{cases}$
The belief set associated to the state $[\rho, \blacktriangle]$ is defined by $Bel([\rho, \blacktriangle]) = f(\rho \cdot \blacktriangle)$.

Hence, when calculating $\mathcal{A}$'s beliefs from the sentences appearing in its epistemic state, an agent gives highest priority to $\blacktriangle$. After that, it prioritises more recent information received. Note that $\blacktriangle$ is always believed, and that $Bel([\rho, \blacktriangle])$ is inconsistent if and only if $\blacktriangle$ is inconsistent.[1]

Example 1. Consider $\blacktriangle = \neg p$ and $\rho = (q, q \rightarrow p)$. $Bel([\rho, \blacktriangle]) = f(q, q \rightarrow p, \neg p)$. In order to determine $f(q, q \rightarrow p, \neg p)$ we need to know if q is consistent with $f(q \rightarrow p, \neg p)$. As $f(\neg p) = \neg p$ and $q \rightarrow p$ is consistent with $\neg p$, we have that $f(q \rightarrow p, \neg p) = (q \rightarrow p) \wedge \neg p \equiv \neg q \wedge \neg p$. So q is inconsistent with $f(q \rightarrow p, \neg p)$. Hence, $f(q, q \rightarrow p, \neg p) = f(q \rightarrow p, \neg p)$ and $Bel([\rho, \blacktriangle]) = f(q \rightarrow p, \neg p) \equiv \neg q \wedge \neg p$.

An agent incorporates a new revision input λ into its epistemic state $[\rho, \blacktriangle]$ by simply appending λ to ρ, i.e., the agent's *revision function* $*$ is specified by

Definition 2. *For every $\lambda \in L$, $[\rho, \blacktriangle] * \lambda = [\rho \cdot \lambda, \blacktriangle]$.*

Given this, we see that a new input λ will not always be believed in the new state. Indeed (when $\blacktriangle$ is consistent) it will be so only if it is consistent with $\blacktriangle$. If it contradicts $\blacktriangle$, the agent's belief set will remain unchanged (c.f.*screened* revision [13]) although the input is incorporated into $\mathcal{A}$'s epistemic state. Note also that $\blacktriangle$ remains unaffected by a revision input, i.e., $*$ is a *core-invariant* revision operator [1].[2] Core beliefs are needed to ensure that revision inputs can be rejected. If they were not allowed, which corresponds to demanding $\blacktriangle = \top$ in the above definitions, any consistent revision input would belong to the agent's beliefs.

As is shown in [1], the above revision method satisfies several natural properties. In particular, it stays largely faithful to the AGM postulates [7] (leaving

[1] In [1], the core beliefs were always assumed to be consistent. This is a small but important difference to framework assumed here.

[2] In fact, the model of [1] allows the core itself to be revisable. We do not explore this possibility here.

aside the "success" postulate, which forces all new inputs to be believed), and satisfies slight, "non-prioritised" variants of several postulates for iterated revision which have been proposed, including the well-known ones of Darwiche and Pearl [5]. One characteristic property of this method is the following variant of the rule "Recalcitrance" from [14]:

$$\text{If } \blacktriangle \not\vdash (\lambda_2 \rightarrow \neg\lambda_1) \text{ then } Bel([\rho, \blacktriangle] * \lambda_1 * \lambda_2) \vdash \lambda_1$$

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 $\blacktriangle$.

If not explicitly stated otherwise, from now on we assume $\mathcal{A}$'s epistemic state is always of the form $[\rho, \blacktriangle]$, and that $\mathcal{A}$ determines its belief set and incorporates new inputs into its epistemic state as described above.

3 Explaining an observation

Suppose we make the observation $o = \langle(\phi_1, \dots, \phi_n), (\theta_1, \dots, \theta_n), (D_1, \dots, D_n)\rangle$ about $\mathcal{A}$. Consequently, after receiving the i^{th} input ϕ_i, $\mathcal{A}$'s epistemic state must be $[\rho \cdot (\phi_1, \dots, \phi_i), \blacktriangle]$ and its belief set $f(\rho \cdot (\phi_1, \dots, \phi_i) \cdot \blacktriangle)$, where $[\rho, \blacktriangle]$ is $\mathcal{A}$'s unknown *initial* (i.e., before receiving ϕ_1) epistemic state. Observation o now amounts to the following:

$$\begin{array}{rl} f(\rho \cdot (\phi_1, \dots, \phi_i) \cdot \blacktriangle) \vdash \theta_i & \text{and} \\ \forall \delta \in D_i : f(\rho \cdot (\phi_1, \dots, \phi_i) \cdot \blacktriangle) \not\vdash \delta & i = 1, \dots, n \end{array} \tag{1}$$

We make the following definitions:

Definition 3. *Let* $o = \langle(\phi_1, \dots, \phi_n), (\theta_1, \dots, \theta_n), (D_1, \dots, D_n)\rangle \in O^{\pm}$. *Then* $[\rho, \blacktriangle]$ explains o *(or is* an explanation for o*) iff (1) above holds. We say* $\blacktriangle$ *is an* o-acceptable core *iff* $[\rho, \blacktriangle]$ *explains* o *for* some ρ.

Example 2. (i) The state $[\rho, \blacktriangle] = [(p \rightarrow q), r]$ explains $\langle(p, q), (q, r), (\emptyset, \emptyset)\rangle$. As required $f(p \rightarrow q, p, r) \equiv p \wedge q \wedge r \vdash q$ and $f(p \rightarrow q, p, q, r) \equiv p \wedge q \wedge r \vdash r$.
(ii) $[(p \rightarrow q), \top]$ does not explain the observation $\langle(p, q), (q, r), (\{p\}, \emptyset)\rangle$. We have $f(p \rightarrow q, p\top) \equiv p \wedge q$ which entails p. But according to this observation p is not supposed to be believed after receiving p. Furthermore, $f(p \rightarrow q, p, q, \top) \equiv p \wedge q$ which does not entail r.

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 $f(\rho \cdot (\phi_1, \dots, \phi_n, \phi_{n+1}) \cdot \blacktriangle)$, before receiving the first input $\mathcal{A}$ believes $f(\rho \cdot \blacktriangle)$, and the beliefs after the i^{th} input are $f(\rho \cdot (\phi_1, \dots, \phi_i) \cdot \blacktriangle)$. An obvious question to ask is: are explanations guaranteed to exist for any given observation? The next result goes some way to answering that question.

Proposition 1. *Let* $o = \langle(\phi_1, \dots, \phi_n), (\theta_1, \dots, \theta_n), (D_1, \dots, D_n)\rangle$.
Then $\bot$ *is* o*-acceptable if and only if* $D_i = \emptyset$ *for all* i.

For the setting in [2] where $D_i = \emptyset$ for all i this means that there always exists *some* explanation $[\rho, \blacktriangle]$ for any such o, since the contradiction $\bot$ is an o-acceptable core using *any* ρ. But this would be a most unsatisfactory explanation, since it means we just infer $\mathcal{A}$ believes everything at every step. In the more general case, however, as soon as we know of a sentence that is not believed after a revision step, we cannot guarantee the exisitence of an explanation. Of course, this is because $\bot$ entails anything. Nevertheless, an explanation $[\rho, \blacktriangle]$ might still exist and we can try to find it.

Example 3. A simple example for an observation that does not have an explanation is $\langle (p), (\top), (\{p, \neg p\}) \rangle$. It tells us that after receiving p, $\mathcal{A}$ was agnostic about p. In particular it says p was not believed after receiving it. This means that *any* o-acceptable core $\blacktriangle$ for this o must be such that $\blacktriangle \vdash \neg p$. This in turn means that *any* explanation $[\rho, \blacktriangle]$ for this o yields $f(\rho \cdot p \cdot \blacktriangle) \vdash \neg p$, i.e., $\neg p$ is believed after receiving p. However, the observation tells us $\neg p$ is not supposed to be believed either.

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 $[\rho(o, \blacktriangle(o)), \blacktriangle(o)]$.

3.1 Finding ρ

Given $o = \langle (\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n), (D_1, \ldots, D_n) \rangle$, 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 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 $[\rho, \blacktriangle]$ iff χ would be believed after revising $[\rho, \blacktriangle]$ by λ, i.e., $Bel([\rho, \blacktriangle] * \lambda) \vdash \chi$. In this case we will write $\lambda \Rightarrow_{[\rho, \blacktriangle]} \chi$. The relation $\Rightarrow_{[\rho, \blacktriangle]}$ *almost* satisfies all the rules of a rational inference relation [12]. More precisely the modified version does, viz., $\lambda \Rightarrow'_{[\rho, \blacktriangle]} \chi$ iff $[\blacktriangle \vdash \neg\lambda$ or $\lambda \Rightarrow_{[\rho, \blacktriangle]} \chi]$. 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, for *any* sequence of revision inputs $\phi_1, \ldots, \phi_m$, our revision method satisfies

$$Bel([\rho, \blacktriangle] * \phi_1 * \cdots * \phi_m) \equiv Bel([\rho, \blacktriangle] * f(\phi_1, \ldots, \phi_m, \blacktriangle)).$$

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. (But note 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 $\Rightarrow_{[\rho, \blacktriangle]}$.)

All this means observation o – not yet taking the D_i into account – may be translated into a partial description of the set of conditional beliefs that $\mathcal{A}$ has in its initial epistemic state:

$$\mathcal{C}_{\blacktriangle}(o) = \{f(\phi_1, \dots, \phi_i, \blacktriangle) \Rightarrow \theta_i \mid i = 1, \dots, n\}.$$

However, the observation contains more information. From the D_i we can also extract a set of conditionals we do *not* want to be part of the agent's initial epistemic state. We call them *negative* conditionals. If one of these conditionals did hold in the initial state, this state would not be an explanation for o.

$$\mathcal{N}_{\blacktriangle}(o) = \{f(\phi_1, \dots, \phi_i, \blacktriangle) \Rightarrow \delta \mid i = 1, \dots, n \wedge \delta \in D_i\}.$$

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 such (positive conditional) beliefs has been well-studied and several solutions have been proposed, e.g., [8,10]. One particularly elegant and well-motivated solution is to take the *rational closure* of $\mathcal{C}$ [12]. Furthermore, as is shown in, e.g., [6], this construction is amenable to a relatively simple representation as a sequence of sentences! Our idea is essentially to take $\rho(o, \blacktriangle)$ to be this sequence corresponding to the rational closure of $\mathcal{C}_{\blacktriangle}(o)$. For the case $o \in O$, this is what was done in [2]. However, allowing non-empty D_i complicates matters slightly. These constructions cannot guarantee that none of the negative conditionals in $\mathcal{N}_{\blacktriangle}(o)$ follow from $\mathcal{C}_{\blacktriangle}(o)$, i.e. they do not incorporate negative information. This generalisation was introduced in [3] and will be illustrated in the next section.

The rational closure for both positive and negative conditionals

Given a set of conditionals $\mathcal{C} = \{\lambda_i \Rightarrow \chi_i \mid i = 1, \dots, l\}$ we denote by $\tilde{\mathcal{C}} = \{\lambda_i \rightarrow \chi_i \mid i = 1, \dots, l\}$ the set of *material counterparts* of all the conditionals in $\mathcal{C}$.

We define two types of exceptionality of a conditional $\lambda \Rightarrow \chi$ with respect to a set of sentences U, one for positive and one for negative conditionals.

Definition 4. *A conditional $\lambda \Rightarrow \chi$ is*

- *p-exceptional for a set of sentences U iff $U \vdash \neg\lambda$.*
- *n-exceptional for U iff $U \cup \{\lambda\} \vdash \chi$.*

The intuition is as follows, a conditional is p-exceptional for U if it is not possible to consistently add its antecedent to the set of sentences. It is n-exceptional if adding the antecedent to U will make the consequent inferable – for a negative conditional this is exactly what is not wanted.

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 fix-point (lfp) construction. Those sets have to satisfy the following conditions:

1. $\mathcal{C}_0 = \mathcal{C}$ and $\mathcal{N}_0 = \mathcal{N}$
2. $U_i = \tilde{\mathcal{C}_i} \cup lfp\,(\{\neg\lambda \mid \lambda \Rightarrow \chi \in \mathcal{N}_i \wedge \lambda \Rightarrow \chi \text{ is n-exeptional for } U_i\})$
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 $\mathcal{N}_m = \mathcal{N}_{m+1}$

Then we set[3]

$$\rho_R(\mathcal{C},\mathcal{N}) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0).$$

The calculation of U_i starts off with $\tilde{\mathcal{C}_i}$ and then adds negated antecedents of conditionals in $\mathcal{N}_i$ that are n-exceptional for the growing U_i.

We remark that for $\mathcal{N} = \emptyset$ the entire construction reduces to the original rational closure construction of [6,12], utilised in [2]. This is because $U_i = \tilde{\mathcal{C}_i}$ for all i, as there are no conditionals that could be n-exceptional.

To get an intuition of what the U_i mean, let us take a look at how $\rho_R(\mathcal{C},\mathcal{N})$ will be used and what happens during the calculation. First note that $\rho_R(\mathcal{C},\mathcal{N})$ is a logical chain with the logically strongest sentence at the right-hand end. It will be used as the sequence part ρ in the epistemic state $[\rho, \blacktriangle]$ of an agent. In the calculation of the belief set using $f(\cdot)$ after the i^{th} revision step before processing $\rho_R(\mathcal{C},\mathcal{N})$, $f(\phi_1,\ldots,\phi_i,\blacktriangle)$ will have been collected. Note that this corresponds to the antecedent of at least one conditional. The processing of $\rho_R(\mathcal{C},\mathcal{N})$ then reduces to the choice of exactly one U_j – the first one to be consistent with $f(\phi_1,\ldots,\phi_i,\blacktriangle)$. This is because the ones with greater index are inconsistent with $f(\phi_1,\ldots,\phi_i,\blacktriangle)$ while the ones with smaller index are all entailed by U_j.

Any U_i basically contains all the positive conditionals (or rather the corresponding implications) that still need to be satisfied (so U_0 contains all of them).[4]

We now have to find out which positive conditionals are satisfied at the current stage and which have to be dealt with later on. It does not suffice to check that the antecedent λ of a positive conditional $(\lambda \Rightarrow \chi)$ can be consistently added (which then gives us the desired consequent). It is possible that adding that λ also makes the consequent of a *negative* conditional inferable, which is not wanted. So we have to make sure that in this case λ cannot be consistently added. We do so by adding $\neg\lambda$ to U_i. This addition, however, can have an effect

[3] We define $\bigwedge \emptyset$ to be the tautology.

[4] Informally, a positive conditional is satisfied by U_i if its antecedent can be consistently added to U_i yielding the desired consequent – which usually is ensured as the material counterpart of the conditional belongs to U_i – without contradicting the negative information in the observation. This means a conditional not satisfied by U_i still needs to be dealt with by a later U_j otherwise the epistemic state cannot be an explanation.

on other negative conditionals, so we check all of them again until U_i does not change anymore. With U_i now settled, we can determine which conditionals still need to be dealt with, i.e. which were still not yet satisfied by U_i and have to be satisfied by U_{i+1} or even later.

Writing α_i for $\bigwedge U_i$, this construction defines the relation $\Rightarrow_R$ given by $\lambda \Rightarrow_R \chi$ iff either $\alpha_m \vdash \neg\lambda$ or $\big[\alpha_j \wedge \lambda \vdash \chi$ where j is minimal such that $\alpha_j \nvdash \neg\lambda\big]$. Since $\alpha_m \dashv \cdots \dashv \alpha_0$ it easy to check that in fact this second disjunct is equivalent to $f(\alpha_m, \ldots, \alpha_0, \lambda) \vdash \chi$.

We now make the following definition:

Definition 5. *Let $o \in O^{\pm}$ 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 4. Let $o = \langle (p,q), (r,\neg p), (\{p\}, \emptyset)\rangle$ and $\blacktriangle = \neg p$. Then

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

$U_0 = \tilde{\mathcal{C}}_0 = \{\neg p \rightarrow r, \neg p \wedge q \rightarrow \neg p\}$. This is because the only negative conditional is not n-exceptional with respect to this set. Since neither of the individual conditionals in $\mathcal{C}_0$ is p-exceptional for U_0 we get that $\mathcal{C}_1 = \emptyset$. Also $\mathcal{N}_1 = \emptyset$ as the negative conditional is not n-exceptional for U_0. Clearly, then also $\mathcal{C}_2 = \emptyset = \mathcal{C}_1$ and $\mathcal{N}_2 = \emptyset = \mathcal{N}_1$, so we obtain $\rho_R(o, \blacktriangle) = (\bigwedge \emptyset, \bigwedge U_0)$. Rewriting the sequence using logically equivalent sentences we get $\rho_R(o, \blacktriangle) = (\top, \neg p \rightarrow r)$.

Now, an interesting thing to note about the rational prefix construction is that it actually goes through *independently* of whether $\blacktriangle$ is o-acceptable. In fact a useful side-effect of the construction is that it actually *reveals* whether $\blacktriangle$ is o-acceptable. For $\blacktriangle \equiv \bot$ Proposition 1 tells us if $\blacktriangle$ is o-acceptable – if $D_i = \emptyset$ for all i, then $\blacktriangle$ is o-acceptable, otherwise it is not. For $\blacktriangle \not\equiv \bot$, 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 2. *Let $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$ be the rational prefix for $o \in O^{\pm}$ and $\blacktriangle \in L$ such that $\blacktriangle \not\equiv \bot$. Then*
(i) *if $\alpha_m \equiv \top$ then $[\rho_R(o, \blacktriangle), \blacktriangle]$ is an explanation for o.*
(ii) *if $\alpha_m \not\equiv \top$ then $\blacktriangle$ is not an o-acceptable core.*

Thus these propositions give us a necessary and sufficient condition for $\blacktriangle$ to be an o-acceptable core. This will be used in the algorithm of Section 3.3.

In Example 4 $\rho_R(o, \blacktriangle) = (\top, \neg p \rightarrow r)$ was calculated. The above proposition implies that $[(\top, \neg p \rightarrow r), \neg p]$ explains $o = \langle (p,q), (r,\neg p), (\{p\}, \emptyset)\rangle$. This is verified by $f(\top, \neg p \rightarrow r, p, \neg p) \equiv \neg p \wedge r \vdash r$, $f(\top, \neg p \rightarrow r, p, q, \neg p) \equiv \neg p \wedge q \wedge r \vdash \neg p$, and $f(\top, \neg p \rightarrow r, p, \neg p) \equiv \neg p \wedge r \nvdash p$.

Justification for using the rational prefix

Given some o-acceptable core belief ▲ there will be several sequences σ, such that $[\sigma, ▲]$ explains o. So, the question is why we should choose $\rho_R(o, ▲)$ rather than any other of those sequences. In [2] we gave an answer to that question for the case that $o \in O$, i.e. there was no information about what was not believed after a revision step.

We did so by comparing the belief traces of the possible solutions. The belief trace of a state $[\sigma, ▲]$ is the sequence of sentences $(Bel_0^\sigma, Bel_1^\sigma, \ldots, Bel_n^\sigma)$, where Bel_i^σ is defined to be the beliefs after the i^{th} input in o. In other words $Bel_i^\sigma = f(\sigma \cdot (\phi_1, \ldots, \phi_i) \cdot ▲)$. So Bel_0^σ gives the initial belief set.

Example 5. Let o, ▲ and $\rho_R(o, ▲)$ be as in Example 4. Then the belief trace is $(\neg p \wedge r, \neg p \wedge r, \neg p \wedge q \wedge r)$.

Given any two possible belief traces $(\beta_0, \ldots, \beta_n)$ and $(\gamma_0, \ldots, \gamma_n)$, let us write $(\beta_0, \ldots, \beta_n) \leq_{\text{lex}} (\gamma_0, \ldots, \gamma_n)$ iff, for all $i = 0, \ldots, n$, $\big[\beta_j \equiv \gamma_j$ for all $j < i$ implies $\gamma_i \vdash \beta_i\big]$. Elements lower down in the ordering are considered better. So, we look at the initial beliefs first. If one trace has a logically weaker initial belief than another trace, the former is preferred. If the initial beliefs are equivalent, we go on to the beliefs after the first revision step. Again, the trace with the weaker belief there is preferred, and so on. This preference relation between traces naturally defines a preference relation between the sequences σ that explain an observation o given a fixed o-acceptable core.

It turns out that the rational prefix gives rise to a belief trace that is at least as preferred (using this preference relation) as the belief trace of any other solution. Furthermore, among the most preferred solutions the rational prefix will predict the logically weakest belief after a further revision no matter what the input is. Whether the same or a similar result holds for the more general case considered in this paper, is still under investigation.

3.2 Minimising ▲

In this section, we will restrict our attention to those observations $o \in O^{\pm}$ that have an o-acceptable core and we will denote this set of observations by $O^{\pm}\!\downarrow$. [5] Core beliefs are needed to allow non-prioritised revision, but at the same time we try to minimise the assumptions about the agent's beliefs. This includes minimising ▲. The first idea would be to simply take the disjunction of all possible o-acceptable cores, i.e., to take $▲_\vee(o)$, defined by

Definition 6. $▲_\vee(o) \equiv \bigvee\{▲ \mid ▲$ *is an* o*-acceptable core*$\}$.

But is $▲_\vee(o)$ itself o-acceptable? 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.

[5] Note, that it makes no sense to try and identify a best solution if there is none. Further, recall that every observation without negative information has a solution, i.e. $O \subset O^{\pm}\!\downarrow$.

Proposition 3. *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) $\blacktriangle(o)$ is an o-acceptable core

What other properties does $\blacktriangle_\vee(o)$ satisfy? Clearly, $\blacktriangle_\vee(o)$ will always be consistent provided at least one consistent o-acceptable core exists:
(Consistency) If $\blacktriangle(o) \equiv \bot$ then $\blacktriangle' \equiv \bot$ for every o-acceptable core $\blacktriangle'$.

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. However for $\blacktriangle_\vee$ we can say more. Given two observations $o = \langle \iota, \tau, D \rangle$ and $o' = \langle \iota', \tau', D' \rangle$, let us denote by $o \cdot o'$ the concatenation of o and o', i.e., $o \cdot o' = \langle \iota \cdot \iota', \tau \cdot \tau', D \cdot D' \rangle$. 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) $o'' \in O^\pm$, and $o \sqsubseteq_{\text{left}} o'$ to denote o' *left extends* o, i.e., $o' = o'' \cdot o$ for some (possibly empty) $o'' \in O^\pm$.

Proposition 4. *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. As pointed out in [2], it is *not* the case that by inserting any observation *anywhere* in o, $\blacktriangle_\vee$ will always lead to a logically stronger core. Intuitively, an intermediate input can possibly explain a change in the belief set that would have to be attributed to the core belief, otherwise.

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 5. *Let $\blacktriangle : O^\pm\!\downarrow \to L$ be any function which returns a sentence given any $o \in O^\pm\!\downarrow$. 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 $o \in O^\pm\!\downarrow$.*

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 7. *Let $o \in O^{\pm}\downarrow$ be an observation. Then we call $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ the* rational explanation *for o.*

But how might we find it in practice? The next section gives an algorithm for just that.

3.3 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 corresponding 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 2 that $[\rho_0, \blacktriangle_0]$ is an explanation for o and so we stop and return this as output. If it isn't then Proposition 2 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 $o = \langle(\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n), (D_1, \ldots, D_n)\rangle$
Output: the rational explanation for o

```
▲ ⇐ ⊤
repeat
    ρ ⇐ ρ_R(o, ▲)          {ρ = (α_m, ..., α_0)}
    ▲ ⇐ ▲ ∧ α_m
until α_m ≡ ⊤
```

Return If $\blacktriangle \equiv \bot$ and $\exists i : D_i \neq \emptyset$ then "no explanation", $[\rho, \blacktriangle]$ otherwise

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

Lemma 1. *Let $\blacktriangle$ and α_m be as after the calculation of $\rho_R(o, \blacktriangle)$ in the repeat loop of the algorithm. 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 $\blacktriangle \equiv \bot$. However in this case it can be shown the rational prefix of o with respect to $\bot$ is just $(\top)$, and so the termination condition will be satisfied at the very next step.

Now we turn to the correctness of the output of the algorithm. Assume there is no explanation for o. Then in this case the algorithm will terminate returning an inconsistent core $\blacktriangle$. This is because if the returned $\blacktriangle$ was consistent we would have calculated an explanation – note that $\alpha_m \equiv \top$, as otherwise the

algorithm would not have terminated, hence by Proposition 2 $[\rho_R(o, \blacktriangle), \blacktriangle]$ is an explanation for o, contradicting the assumption there is none. Also, Proposition 1 tells us there must be a non-empty D_i. Hence if no explanation exists the algorithm correctly tells us so.

Now assume there is an explanation for o. To show the output matches the rational explanation, consider the sequence $[\rho_0, \blacktriangle_0], \dots, [\rho_k, \blacktriangle_k]$ 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 $i = 0, \dots, k$: the case $i = 0$ clearly holds since $\blacktriangle_0 \equiv \top$. The inductive step uses the following property:

Lemma 2. *Let $0 < i \leq k$ and suppose $\rho_{i-1} = (\alpha_m, \dots, \alpha_0)$. 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. Since obviously ρ_k is the rational prefix of o with respect to $\blacktriangle_k$ by construction, we have:

Proposition 6. *Given input observation o, the algorithm outputs the rational explanation for o, if an explanation for o exists. If no explanation exists it outputs "no explanation".*

3.4 An Example

Let $o = \langle (p, q, r), (s, \top, \neg q), (\emptyset, \{s\}, \emptyset) \rangle$. So o is saying that after $\mathcal{A}$ receives p then it believes s. Then, receiving q leads $\mathcal{A}$ to drop this belief in s. Finally, after receiving r, $\mathcal{A}$ believes $\neg q$.

The calculation of the rational explanation starts off with the core $\blacktriangle = \top$, so our conditionals have the antecedents $f(p \cdot \top) \equiv p$, $f(p \cdot q \cdot \top) \equiv p \wedge q$, and $f(p \cdot q \cdot r \cdot \top) \equiv p \wedge q \wedge r$. We get

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

By construction $\tilde{\mathcal{C}_0} \subseteq U_0$ and as $p \wedge q \Rightarrow s$ is n-exceptional for U_0, $\neg(p \wedge q)$ has to be added as well. 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, $p \wedge q \Rightarrow s$ is still n-exceptional for U_0, so

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

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. Note that adding $p \wedge q$ to U_1 does not make s inferable. In fact $p \wedge q \wedge r \Rightarrow \neg q$ is exceptional for *itself* because $p \wedge q \wedge r$ is inconsistent with $p \wedge q \wedge r \rightarrow \neg q$. So we have

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

As $\alpha_m = \alpha_2 \not\equiv \top$, we know, that $\blacktriangle = \top$ is not o-acceptable and has to be adapted. The new core is the old one conjoined with α_2, so the new core is $\neg p \vee \neg q \vee \neg r$. This means, that we get the following sets of conditionals.

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

As in the first run the negative conditional is n-exceptional, so

$U_0 = \{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r \rightarrow \top, \neg p \wedge q \wedge r \rightarrow \neg q, \neg(p \wedge q \wedge \neg r)\}$

$\mathcal{C}_1 = \{p \wedge q \wedge \neg r \Rightarrow \top, \neg p \wedge q \wedge r \Rightarrow \neg q\}$
$\mathcal{N}_1 = \{p \wedge q \wedge \neg r \Rightarrow s\}$
$U_1 = \{p \wedge q \wedge \neg r \rightarrow \top, \neg p \wedge q \wedge r \rightarrow \neg q\}$

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

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

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

The n-exceptionality of $p \wedge q \wedge \neg r \Rightarrow s$ is easily verified, so that

$U_0 = \{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r \rightarrow \top, p \wedge \neg q \wedge r \rightarrow \neg q, \neg(p \wedge q \wedge \neg r)\}$

None of the positive conditionals is p-exceptional for U_0, but $p \wedge q \wedge \neg r \Rightarrow s$ is n-exceptional. So we get

$\mathcal{C}_1 = \emptyset$
$\mathcal{N}_1 = \{p \wedge q \wedge \neg r \Rightarrow s\}$
$U_1 = \emptyset$

$\mathcal{C}_2 = \mathcal{C}_3 = \emptyset$
$\mathcal{N}_2 = \mathcal{N}_3 = \emptyset$
$U_2 = U_3 = \emptyset$

Now, we are done as $\alpha_m = \top$ and as $\blacktriangle \not\equiv \bot$, we indeed have an explanation:

$$[\sigma, \blacktriangle] = [(\top, (p \wedge (\neg q \vee \neg r) \rightarrow s) \wedge (\neg p \vee \neg q \vee r)), \neg q \vee \neg r]$$

Let us verify, that $o = \langle (p, q, r), (s, \top, \neg q), (\emptyset, \{s\}, \emptyset) \rangle$ is indeed explained by $[\sigma, \blacktriangle]$.[6] We do so by calculating the belief trace, omitting Bel_0^σ.

$$(Bel_1^\sigma, Bel_2^\sigma, Bel_3^\sigma) = (p \wedge \neg q \wedge s, p \wedge q \wedge \neg r, p \wedge \neg q \wedge r \wedge s)$$

[6] As we know the core $\blacktriangle$ to hold at each step, we can simplify the belief state to $[(\top, (p \rightarrow s) \wedge (\neg p \vee \neg q)), \neg q \vee \neg r]$ and get $Bel_0^\sigma = (\neg p \vee s) \wedge (\neg p \vee \neg q) \wedge (\neg q \vee \neg r)$ as the beliefs the agent held before receiving any input.

We see that, indeed, s is believed after receiving p, and not believed anymore after then receiving q. After receiving the last input r, $\neg q$ is believed.

In fact, if we look at the observation it is clear that $\blacktriangle$ has to entail at least $\neg q \vee \neg r$. If it did not, it would not be possible to keep q out of the belief set when constructing it for the case after incorporating r into the epistemic state. However, q has to be kept out, as $\mathcal{A}$ believes $\neg q$ after receiving r. The revision operator assumed in this paper satisfies the AGM-postulates (given an input does not contradict the core belief). Hence, it is reasonable that after the first revision step $\neg q$ is believed. If q was consistent with the belief set at that stage, the revision would simply have to add q. But then s would still have to be believed after revising with q, which it is not. In other words, in order for s to be dropped from the belief set, q has to contradict it. A core belief is not needed in this case (in contrast to the one involving q and r), because nothing in the observation prevents us from assuming that p is still believed after receiving q. The same argument supports that $\neg r$ must be believed after receiving q, because the belief set does not simply expand after receiving r in the next step.

So far, we have justified the core belief, all of Bel_1^σ (input p believed, $\neg q$ as argued above, s forced by the observation), all of Bel_2^σ (no reason why p would have to be dropped, input q accepted, $\neg r$ as argued above, s not believed anymore as forced by the observation), and most of Bel_3^σ (no reason why p would have to be dropped, $\neg q$ as forced by the observation, input r believed). It remains to explain, why s reappears in the belief set.

The observation indicates that p is a reason to believe s and it implies that q is a (stronger) reason not to believe s. After receiving r the reason not to believe s is gone, while the reason for believing s remains. So it seems plausible to infer that s is indeed believed. Note, that if p is *not* a reason to believe s, it would have to be believed even before p arrived. In this case it would still make sense to assume s to be believed in the end, because for all we know q is the only reason for s not to be believed.

4 Discussion

In this section, we briefly want to discuss some limitations of our approach and comment on a few issues that came up during the Dagstuhl seminar.

4.1 Core constraints

We motivated the use of core beliefs by the need for non-prioritised revision. However, core beliefs do more than block certain inputs from being believed – they cause the negations of these inputs to be believed. This is much stronger than just not believing an input. As a consequence, the agent cannot remain agnostic about a sentence. Consider the following observation which basically says that the agent received p followed by $\neg p$ as inputs, but did not believe either: $o = \langle (p, \neg p), (\top, \top), (\{p\}, \{\neg p\}) \rangle$.

The belief revision framework assumed in this paper forces us to assign an inconsistent ▲ to the agent, which of course does not explain the observation. There is a rather natural generalisation of the framework that prevents this problem. Rather than assuming a core belief – a sentence the agent will always commit to, we allow the agent to have core constraints – sentences it wants to remain consistent with, i.e. $\mathcal{A}$ will not believe their negations.

The reason to allow several sentences and not just one is the same as for generalising the observations allowing sets of sentences not to be believed. A single sentence would not allow agnosticism.

Let $\sigma = (\alpha_m, \ldots, \alpha_1)$ be a sequence of sentences and CC the set of core constraints. These two components now make up the epistemic state of an agent. $f(\cdot, \cdot)$ is the function that calculates the sentence corresponding to the beliefs of an agent and is defined inductively on the length of σ as follows:

$$f((), CC) = \top$$

$$f(\sigma, CC) = \begin{cases} \varphi = \alpha_m \wedge f(\alpha_{m-1}, \ldots, \alpha_1, CC) & \text{if } \forall \psi \in CC : \varphi \wedge \psi \nvdash \bot \\ f(\alpha_{m-1}, \ldots, \alpha_1, CC) & \text{otherwise} \end{cases}$$

We remark that this definition still allows the belief set of an agent to be inconsistent. This is possible if $CC = \emptyset$. This fact shows that core constraints are indeed a generalisation of core beliefs. For every epistemic state $[\sigma, ▲]$ using core beliefs there is one using core constraints that behaves equivalently, i.e. the two yield the same belief sets, even after an arbitrary sequence of revisions.[7] Abusing notation, $[\sigma, ▲]$ can be represented by $[\sigma \cdot ▲, \{▲\}]$ if ▲ is consistent, and by $[(▲), \emptyset]$ if ▲ is inconsistent. The singleton core constraint makes sure that after all revision steps the agent remains consistent with what earlier was the core belief, so that the corresponding sentence can be added when it comes to its initial sequence of beliefs.

This shows that the new framework can explain all observations the original belief revision framework could, however it can also explain the above observation $o = \langle (p, \neg p), (\top, \top), (\{p\}, \{\neg p\}) \rangle$. $[(), \{p, \neg p\}]$, causes both inputs to be rejected as incorporating them would cause an inconsistency with individual core constraints.[8] Hence, this approach would be strictly more general. That is the good news. The bad news is that, so far, we were not able to transfer any of our results to this framework. Although we have an idea of how the weakening (analogous to the construction in Section 3.2) of two sets of core constraints might be, it is far from obvious how it can be proved that the weakened constraints provide a solution, as well. The problem is that the proofs would require extensive reasoning with and inferring from what is not believed by the agent, i.e. drawing conclusions from non-entailment of a set of sentences. One possibility to attack this problem might be to incorporate the notion of not believing into the object language, e.g. by using a logic for disbeliefs ([4,16]).

[7] Of course, we assume here that revision is again defined by $[\rho, CC] * \lambda = [\rho \cdot \lambda, CC]$.

[8] Similarly, the observation $o = \langle (p), (\top), (\{p, \neg p\}) \rangle$ from Example 3 can now be explained by $[(), \{\neg p\}]$.

4.2 In defense of the belief revision framework assumed and temporal aspects

During the Dagstuhl seminar on belief change in rational agents some points regarding belief revision in general and the assumptions in our work in particular were brought up. Some of them deserve discussion here.

Often it is not clear (or not clearly specified) what a new piece of information actually represents. If it is a piece of new or modified background knowledge of the world, it should be allowed to modify the agent's internal structure. If it is merely an observation about the current state of the world that impact might not be wanted. It would suffice to use it to enrich conclusions that the agent can draw from its background knowledge. If the information is of the second type then iterated revision seems not to be necessary. Further, it does not make sense that the incoming information can be contradictory.

In fact, core beliefs can be interpreted as knowledge of the world the agent possesses. The original framework in [1] distinguishes between two revision functions, one for core beliefs and one for regular ones. In this respect, it allows the input to be of both types. So the implicit assumption in our studies is, that the inputs received by the agent are mere observations about the world and our job is to identify which world knowledge ($\blacktriangle$) the agent possesses and which prior observations (ρ) it has made.

It might be interesting to investigate the question of not only reconstructing the agent's initial epistemic state but also make a justified guess about the type of revision an input has caused. By default it could be of the second type but in some cases assuming that the input changed the core beliefs might allow a more satisfactory explanation.

The claim that iterated revision is not necessary for regular inputs is based on the following argument. Rather than incorporating each piece of evidence at a time, it is more rational to collect all evidence, select the relevant and reliable pieces and then incorporate them at once. This is what is ideally done in legal cases. If we assume all the sources to be faithful to the real state of the world, all pieces of evidence had to be consistent. But this assumption is far from realistic. Even in the case of only one source, sensors might be noisy and give different information at different times (although the world did not change).

Our framework indeed carries out a selection of inputs at each point of time. We agree that a preference based solely on recency of information is not the most realistic one. However, the raw data we have does not provide any other information on which inputs should be preferred. Further, recency of information can be argued to be an indicator of reliability. More time and effort has been spent into investigating the status of a sentence.

In general the following can be said. The temporal aspect might not have the primary impact on the overall reliability of a piece of information, but it still is highly relevant to what was believed at which point of time. This is because an agent can only select from information that has been received up to that point of time.

4.3 Conclusion

In this paper, we generalised previous work on reconstructing an agent's epistemic state from observations by allowing further information to be incorporated. In addition to information about what the agent believed after revising by certain inputs, the present approach allows processing information about what the agent did not believe, as well. The model of the agent remained the same, but the calculation of the initial belief state was generalised. Most of the results carry over the setting presented here. Whether the minimality property we gave as justification for the rational prefix in the special case (cf. Section 3.1) holds as well is subject to further investigation.

References

1. Booth, R.: On the logic of iterated non-prioritised revision. In: Conditionals, Information and Inference – Selected papers from the Workshop on Conditionals, Information and Inference, 2002, Springer's LNAI 3301 (2005) 86–107
2. Booth, R., Nittka, A.: Reconstructing an agent's epistemic state from observations. In: Proceedings of IJCAI '05. (2005) 394–399
3. Booth, R., Paris, J.B.: A note on the rational closure of knowledge bases with both positive and negative knowledge. Journal of Logic, Language and Information **7** (1998) 165–190
4. Chopra, S., Heidema, J., Meyer, T.A.: Some logics of belief and disbelief. In Gedeon, T.D., Fung, L.C.C., eds.: Australian Conference on Artificial Intelligence. Volume 2903 of Lecture Notes in Computer Science., Springer (2003) 364–376
5. Darwiche, A., Pearl, J.: On the logic of iterated belief revision. Artificial Intelligence **89** (1997) 1–29
6. Freund, M.: On the revision of preferences and rational inference processes. Artificial Intelligence **152** (2004) 105–137
7. Gärdenfors, P.: Knowledge in Flux. MIT Press (1988)
8. Geffner, H., Pearl, J.: Conditional entailment: Bridging two approaches to default entailment. Artificial Intelligence **53** (1992) 209–244
9. Konieczny, S., Pérez, R.P.: A framework for iterated revision. Journal of Applied Non-Classical Logics **10(3-4)** (2000) 339–367
10. Lehmann, D.: Another perspective on default reasoning. Annals of Mathematics and Artificial Intelligence **15(1)** (1995) 61–82
11. Lehmann, D.: Belief revision, revised. In: Proceedings of IJCAI'95. (1995) 1534–1540
12. Lehmann, D., Magidor, M.: What does a conditional knowledge base entail? Artificial Intelligence **55** (1992) 1–60
13. Makinson, D.: Screened revision. Theoria **63** (1997) 14–23
14. Nayak, A., Pagnucco, M., Peppas, P.: Dynamic belief revision operators. Artificial Intelligence **146** (2003) 193–228
15. Nebel, B.: Base revision operations and schemes: Semantics, representation and complexity. In: Proceedings of ECAI'94. (1994) 342–345
16. Nittka, A.: A 3-valued approach to disbelief. Diplomarbeit, Leipzig University (2003)