Reconstructing an Agent's Epistemic State from Observations

Richard Booth
Macquarie University
Dept. of Computing
Sydney NSW 2109 Australia
rbooth@ics.mq.edu.au

Alexander Nittka
University of Leipzig
Dept. of Computer Science
Leipzig 04109 Germany
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.

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 [Gärdenfors, 1988]. 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 this paper we change viewpoint and instead cast ourselves in the role of an *observer* of the agent. Imagine the following scenario. Suppose 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 suppose 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. Throughout the paper, 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. A couple of questions now suggest themselves:

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

One area in which such questions might arise is in human-machine dialogues [del Cerro *et al.*, 1998], where the ϕ_i correspond to inputs given to $\mathcal{A}$ – the user – by a machine, and the θ_i are the user's responses. Here, it can be useful for the machine to keep a model of the evolution of a user's beliefs during a dialogue. Another setting where reconstruction of a prior belief set is needed is in law cases, where guilt and innocence often depend on who knew what, when. Inquiry is a possible method of approaching this type of problem. The above mentioned sequences can be seen as the information the inquiry yields and which is now used for drawing conclusions. Our aim in this paper is to answer these questions.

Our strategy for dealing with these questions 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., [Brafman and Tennenholtz, 1997; Lang, 2004].) 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 [Darwiche and Pearl, 1997; Lehmann, 1995b; Nayak *et al.*, 2003], and preferably also *non-prioritised* revision [Hansson *et al.*, 2001; Makinson, 1997], i.e., revision in which new inputs are allowed to be *rejected*. In this paper, we restrict the investigation to one such framework that has been studied in [Booth, 2005]. 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. 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 $o = \langle(\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n)\rangle$ we have made of $\mathcal{A}$.

The plan of the paper is as follows. In Sect. 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 Sect. 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 Sect. 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 Sect. 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 Sect. 6. In Sect. 7 we briefly mention a related piece of work by Dupin de Saint-Cyr and Lang, before concluding and giving some pointers for future research.

2 Modelling the Agent

We assume sentences $\phi_i, \theta_i, \blacktriangle$, etc. are elements of some finitely-generated propositional language L. In our examples, p, q, r 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 set of all possible observations $o = \langle \iota, \tau \rangle$ which can be made of $\mathcal{A}$, where $\iota = (\phi_1, \ldots, \phi_n)$ and $\tau = (\theta_1, \ldots, \theta_n)$ are two finite sequences of sentences of the same length, is denoted by O. The operation $\cdot$ on sequences denotes sequence concatenation.

As indicated in the introduction, we follow [Booth, 2005] by assuming that, at any given moment in time, an agent's epistemic state is represented by a pair $[\rho, \blacktriangle]$. ([Konieczny and Pérez, 2000; Lehmann, 1995b] 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. f is defined by induction on the length m of σ: if $m = 1$ then $f(\sigma) = \alpha_1$. If $m > 1$ then

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

In other words $f(\sigma)$ 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 [Nebel, 1994] and the "basic memory operator" of [Konieczny and Pérez, 2000].) The belief set associated to the state $[\rho, \blacktriangle]$ is then given by $Bel([\rho, \blacktriangle]) = f(\rho \cdot \blacktriangle)$. Hence when calculating its 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.

Example 2.1. 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 $\neg p$, $f(q \rightarrow p, \neg p) = (q \rightarrow p) \wedge \neg p \equiv \neg q \wedge \neg p$. So q is inconsistent with $f(q \rightarrow p, \neg p)$. Consequently we get $f(q, q \rightarrow p, \neg p) = f(q \rightarrow p, \neg p)$ 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 setting, 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 $\blacktriangle$ then it will not be accepted, and in fact in this case the agent's belief set will remain unchanged (c.f. *screened* revision [Makinson, 1997]). Note also that $\blacktriangle$ remains unaffected by a revision input, i.e., $*$ is a *core-invariant* revision operator [Booth, 2005].[1] 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 [Booth, 2005], the above revision method satisfies several natural properties. In particular, it stays largely faithful to the AGM postulates [Gärdenfors, 1988] (leaving aside the "success" postulate, which forces all new inputs to be accepted), and satisfies slight,"non-prioritised" variants of several postulates for iterated revision which have been proposed, including those of [Darwiche and Pearl, 1997]. One characteristic property of this method is the following variant of the rule "Recalcitrance" from [Nayak *et al.*, 2003]:

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

This entails if the agent *accepts* 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 $\blacktriangle$.

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

$$f(\rho \cdot (\phi_1, \ldots, \phi_i) \cdot \blacktriangle) \vdash \theta_i \qquad i = 1, \ldots, n \qquad (1)$$

We make the following definitions:

Definition 2.2. *Let $o = \langle (\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n) \rangle \in O$. 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.*

[1] In fact the model of [Booth, 2005] allows the core itself to be revisable. We do not explore this possibility here.

Example 2.3. *(i)* $[\rho, \blacktriangle] = [(p \rightarrow q), r]$ explains $\langle (p,q),(q,r)\rangle$ because $f(p \rightarrow q, p, r) \equiv p \wedge q \wedge r \vdash q$ and $f(p \rightarrow q, p, q, r) \equiv p \wedge q \wedge r \vdash r$.
(ii) $[(p \rightarrow q), \top]$ does not explain $\langle (p,q),(q,r)\rangle$ because $f(p \rightarrow q, p, q, \top) \equiv p \wedge q \nvdash r$.

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 $f(\rho \cdot (\phi_1, \ldots, \phi_n, \phi_{n+1}) \cdot \blacktriangle)$, before receiving the first input $\mathcal{A}$ believes $f(\rho \cdot \blacktriangle)$, and the beliefs after the i^{th} input are $f(\rho \cdot (\phi_1, \ldots, \phi_i) \cdot \blacktriangle)$.

Note for any $o \in O$ there always exists *some* explanation $[\rho, \blacktriangle]$ for 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.

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 $o = \langle (\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n)\rangle$, 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 $[\rho, \blacktriangle]$ 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 $\lambda \Rightarrow_{[\rho,\blacktriangle]} \chi$.[2] 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

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

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 may be translated into a partial description of the set of conditional beliefs that $\mathcal{A}$ has in its initial epistemic state:

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

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 beliefs has been well-studied and several solutions have been proposed, e.g., [Geffner and Pearl, 1992; Lehmann, 1995a]. One particularly elegant and well-motivated solution is to take the *rational closure* of $\mathcal{C}$ [Lehmann and Magidor, 1992]. Furthermore, as is shown in, e.g., [Freund, 2004], 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 $\mathcal{C}_{\blacktriangle}(o)$. First let us describe the general construction.

3.1 The rational closure of a set of conditionals

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., $\tilde{\mathcal{C}} = \{\lambda_i \rightarrow \chi_i \mid i = 1, \ldots, l\}$. Then a sentence ν is *exceptional* for $\mathcal{C}$ iff $\tilde{\mathcal{C}} \vdash \neg\nu$, and a conditional $\nu \Rightarrow \mu$ is exceptional for $\mathcal{C}$ iff its antecedent ν is. To find the (sequence corresponding to the) rational closure $\rho_R(\mathcal{C})$ of $\mathcal{C}$, we first define a decreasing sequence of sets of conditionals $\mathcal{C}_0 \supseteq \mathcal{C}_1 \supseteq \cdots \supseteq \mathcal{C}_m$ by setting *(i)* $\mathcal{C}_0 = \mathcal{C}$, *(ii)* $\mathcal{C}_{i+1}$ equals the set of conditionals in $\mathcal{C}_i$ which are exceptional for $\mathcal{C}_i$, and *(iii)* m is minimal such that $\mathcal{C}_m = \mathcal{C}_{m+1}$. Then we set

$$\rho_R(\mathcal{C}) = (\bigwedge \tilde{\mathcal{C}}_m, \bigwedge \tilde{\mathcal{C}}_{m-1}, \ldots, \bigwedge \tilde{\mathcal{C}}_0).$$

Writing α_i for $\bigwedge \tilde{\mathcal{C}}_i$, the rational closure of $\mathcal{C}$ is then 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 3.1. *Let $o \in O$ and $\blacktriangle \in L$. We call $\rho_R(\mathcal{C}_{\blacktriangle}(o))$ the* rational prefix of *o with respect to $\blacktriangle$, and will denote it by $\rho_R(o, \blacktriangle)$.*

Example 3.2. Let $o = \langle (p,q),(r,\neg p)\rangle$ and $\blacktriangle = \neg p$. Then

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

Since neither of the individual conditionals are exceptional for $\mathcal{C}_{\blacktriangle}(o)$ we get $\mathcal{C}_0 = \mathcal{C}_{\blacktriangle}(o)$ and $\mathcal{C}_1 = \emptyset$. Clearly then also $\mathcal{C}_2 = \emptyset = \mathcal{C}_1$ so we obtain $\rho_R(o, \blacktriangle) = (\bigwedge \emptyset, \bigwedge \tilde{\mathcal{C}}_{\blacktriangle}(o))$. 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. 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 3.3. *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 w.r.t. $\blacktriangle$. 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 this proposition gives us a necessary and sufficient condition for $\blacktriangle$ to be an o-acceptable core. This will be used in the algorithm of Sect. 5.

In Example 3.2 $\rho_R(o, \blacktriangle) = (\top, \neg p \rightarrow r)$ was calculated. The above proposition implies $[(\top, \neg p \rightarrow r), \neg p]$

[2] The relation $\Rightarrow_{[\rho,\blacktriangle]}$ *almost* satisfies all the rules of a rational inference relation [Lehmann and Magidor, 1992]. More precisely the modified version does, viz., $\lambda \Rightarrow'_{[\rho,\blacktriangle]} \chi$ iff $[\blacktriangle \vdash \neg\lambda$ or $\lambda \Rightarrow_{[\rho,\blacktriangle]} \chi]$.

is an explanation for $o = \langle (p, q), (r, \neg p) \rangle$. This is verified by $f(\top, \neg p \rightarrow r, p, \neg p) \equiv \neg p \wedge r \vdash r$ and $f(\top, \neg p \rightarrow r, p, q, \neg p) \equiv \neg p \wedge q \wedge r \vdash \neg p$.

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, $[\rho_R(o, ▲), ▲]$ 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 $\Sigma = \{\sigma \mid [\sigma, ▲] \text{ explains } o\}$.

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 $[\sigma, ▲]$). 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 3.4. Let o, ▲ and $\rho_R(o, ▲)$ be as in Example 3.2. Then the belief trace is $(\neg p \wedge r, \neg p \wedge r, \neg p \wedge q \wedge r)$.

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 $(\beta_0, \ldots, \beta_n) \leq_{\text{lex}} (\gamma_0, \ldots, \gamma_n)$ iff, for all $i = 0, \ldots, n$, $\big[\beta_j \equiv \gamma_j \text{ for all } j < i \text{ implies } \gamma_i \vdash \beta_i\big]$. 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., reflexive and transitive) 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 $\rho_R(o, ▲)$ is a best element in Σ under this ordering.

Proposition 3.5. $\rho_R(o, ▲) \preceq_1 \sigma$ *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.

$$\rho \preceq_2 \sigma \quad \text{iff} \quad Bel([\sigma, ▲] * \phi_1 * \cdots * \phi_n * \lambda) \vdash Bel([\rho, ▲] * \phi_1 * \cdots * \phi_n * \lambda) \text{ for all } \lambda$$

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. It turns out $\rho_R(o, ▲)$ is a most preferred element under $\preceq_2$ *amongst all minimal elements under* $\preceq_1$.

Proposition 3.6. *For all* $\sigma \in \Sigma$, *if* $\sigma \preceq_1 \rho_R(o, ▲)$ *then* $\rho_R(o, ▲) \preceq_2 \sigma$.

Thus if we take a lexicographic combination of $\preceq_1$ and $\preceq_2$ (with $\preceq_1$ being considered as more important), $\rho_R(o, ▲)$ emerges overall as a best, most preferred, member of Σ. Having provided a method for finding the best explanation $[\rho, ▲]$ *given* ▲, we now turn our attention to finding the best ▲ itself.

4 Minimising ▲

As argued earlier, core beliefs are needed, 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

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

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.

Proposition 4.1. *If* $▲_1$ *and* $▲_2$ *are o-acceptable then so is* $▲_1 \vee ▲_2$.

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

(Acceptability) $▲(o)$ is an o-acceptable core

What other properties does $▲_\vee(o)$ satisfy? Clearly, $▲_\vee(o)$ will always be consistent provided at least one consistent o-acceptable core exists:

(Consistency) If $▲(o) \equiv \bot$ then $▲' \equiv \bot$ for every o-acceptable core $▲'$.

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 $▲_\vee$ we can say more. Given two observations $o = \langle \iota, \tau \rangle$ and $o' = \langle \iota', \tau' \rangle$, let us denote by $o \cdot o'$ the concatenation of o and o', i.e., $o \cdot o' = \langle \iota \cdot \iota', \tau \cdot \tau' \rangle$. 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) $o'' \in O$, 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 $▲_\vee$ satisfies the following 2 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 $▲(o') \vdash ▲(o)$

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

Right- and Left Monotony provide ways of expressing that $▲(o)$ 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, $▲_\vee$ will always lead to a logically stronger core. Consider $o_1 = \langle (p, q), (p, \neg p) \rangle$ and $o_2 = \langle (p, \neg p, q), (p, \neg p, \neg p) \rangle$, i.e., $\langle (\neg p), (\neg p) \rangle$ was inserted in the middle of o_1. $▲_\vee(o_1) \equiv q \rightarrow \neg p$ whereas $▲_\vee(o_2) \equiv \top$. So although o_2 extends o_1 in a sense, the corresponding $▲_\vee$ 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 Sect. 2). However, looking at o_2, the information that $\mathcal{A}$ received (and accepted) 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 $o = \langle (\phi_1, \ldots, \phi_n), (\theta_1, \ldots, \theta_n) \rangle$ 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. This will be a topic for future work.

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 $\blacktriangle : O \to L$ 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. Then we call $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ the* rational explanation *for o.*

In Sect. 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 w.r.t. $\blacktriangle_0$. We then check whether α_m is a tautology. If it is then we know by Prop. 3.3 that $[\rho_0, \blacktriangle_0]$ is an explanation for o and so we stop and return this as output. If it isn't then Prop. 3.3 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 w.r.t. 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
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 $[\rho, \blacktriangle]$

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 $\blacktriangle \equiv \bot$. However in this case it can be shown the rational prefix of o w.r.t. $\bot$ is just $(\top)$, and so the termination condition will be satisfied at the very next step.

Now, to show the output matches the rational explanation, 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. Since obviously ρ_k is the rational prefix of o w.r.t. $\blacktriangle_k$ by construction, we have:

Proposition 5.3. *Given input observation o, the algorithm outputs the rational explanation for o.*

Example 5.4. Let $o = \langle (p, q), (r, \neg p) \rangle$. Starting with $\blacktriangle = \top$, in the first run $\tilde{\mathcal{C}}_0 = \{f(p, \top) \to r, f(p, q, \top) \to \neg p\} = \{p \to r, p \wedge q \to \neg p\}$. Only the second conditional is exceptional, so $\tilde{\mathcal{C}}_1 = \{p \wedge q \to \neg p\}$. Now the remaining conditional is exceptional for itself, so $\tilde{\mathcal{C}}_2 = \tilde{\mathcal{C}}_1$. $\blacktriangle$ is updated to $\blacktriangle \equiv p \to \neg q$ because $\rho = (p \to \neg q, p \to (r \wedge \neg q))$.

The next calculation yields $\tilde{\mathcal{C}}_0 = \{f(p, p \to \neg q) \to r, f(p, q, p \to \neg q) \to \neg p\} = \{p \wedge \neg q \to r, q \wedge \neg p \to \neg p\}$. This time none of the conditionals are exceptional, so $\tilde{\mathcal{C}}_1 = \emptyset$. As this means $\alpha_1 = \top$, no further run is necessary and the result is $\rho = (\top, (p \wedge \neg q) \to r)$, $\blacktriangle = p \to \neg q$. $[(\top, (p \wedge \neg q) \to r), p \to \neg q]$ is the rational explanation for o.

6 Some Examples

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

For $o = \langle (p), (q) \rangle$, the rational explanation is $[(\top, p \to q), \top]$. So we infer $\mathcal{A}$'s initial belief set is $p \to q$. Indeed to explain $\mathcal{A}$'s belief in q following receipt of p it is clear $\mathcal{A}$ *must* initially believe *at least* $p \to q$ 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 $\mathcal{A}$ *accepts* 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 accept 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 $o = \langle (p), (\neg p) \rangle$, 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 $\neg p$.

In some cases the rational explanation gives only the trivial explanation, i.e., $\blacktriangle_{\vee}(o) \equiv \bot$. One of the simplest examples extends the prior one: $o = \langle (p, \neg p), (\neg p, p) \rangle$. The first part of the observation tells us that $\neg p$ must be a core belief, but when confirmed of that belief $\mathcal{A}$ changes its opinion. This behaviour of always believing the opposite of what one is told can be called rational only in very specific circumstances that are not in the scope of this investigation. Hence, failing to provide a satisfactory explanation for this example is not to be seen as a failure of the method.

7 Conclusion

Before concluding, one paper which deserves special mention as having similarities with the present one is [de Saint-Cyr and Lang, 2002] on *belief extrapolation* (itself an instance of the general framework of [Friedman and Halpern, 1999]). A belief extrapolation operator takes as input a sequence of sentences representing a sequence of partial observations of a possibly changing world, and outputs another sequence which "completes" the input. These operators proceed by trying to determine some history of the world which "best fits" (according to various criteria) the observations. A fundamental difference between that work and ours is that belief extrapolation is, like traditional operators of revision and update, an "agent's perspective" operator – it is concerned with how an agent should form a picture of how the external world is evolving, whereas we are interested in forming a picture of how an *observed agent's beliefs* are evolving. Nevertheless the precise connections between these two works seems worthy of further study.

To conclude, in this paper we made an attempt at reconstructing an agent's initial epistemic state in order to explain a given observation of the agent and make predictions about its future beliefs. We did so by assuming a simple yet powerful model for epistemic states allowing for iterated non-prioritised revision. The algorithm we provided constructs a best explanation based on the rational closure of conditional beliefs. This answer should be applicable to problems requiring the modelling of agents' beliefs, for example in the area of user modelling.

Generalisations of this approach which are object of future work include *(i)* relaxing the assumption that we are given an unbroken sequence of revision inputs, i.e., allowing also for intermediate inputs as an explanation for what the core accounts for now, *(ii)* allowing our observations to incorporate information about what the agent did *not* believe after a given revision step, and *(iii)* allowing the core beliefs to be revised. Further, it is of interest to compare our results with what other models of epistemic states would yield as explanation.

Acknowledgements

Thanks are due to the reviewers for helpful comments. A.N. acknowledges support by the EC (IST-2001-37004, WASP).

References

[Booth, 2005] 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*, pages 86–107. Springer's LNAI 3301, 2005.

[Brafman and Tennenholtz, 1997] R. I. Brafman and M. Tennenholtz. Modeling agents as qualitative decision makers. *Artificial Intelligence*, 94(1-2):217–268, 1997.

[Darwiche and Pearl, 1997] A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89:1–29, 1997.

[de Saint-Cyr and Lang, 2002] F. Dupin de Saint-Cyr and J. Lang. Belief extrapolation (or how to reason about observations and unpredicted change). In *Proceedings of KR'02*, pages 497–508, 2002.

[del Cerro *et al.*, 1998] L. Fariñas del Cerro, A. Herzig, D. Longin, and O. Rifi. Belief reconstruction in cooperative dialogues. In *Proceedings of AIMSA'98*, pages 254–266. Springer's LNCS 1480, 1998.

[Freund, 2004] M. Freund. On the revision of preferences and rational inference processes. *Artificial Intelligence*, 152(1):105–137, 2004.

[Friedman and Halpern, 1999] N. Friedman and J. Halpern. Modeling belief in dynamic systems, part II: Revision and update. *Journal of Artificial Intelligence Research*, 10:117–167, 1999.

[Gärdenfors, 1988] P. Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.

[Geffner and Pearl, 1992] H. Geffner and J. Pearl. Conditional entailment: Bridging two approaches to default entailment. *Artificial Intelligence*, 53:209–244, 1992.

[Hansson *et al.*, 2001] S. O. Hansson, E. Fermé, J. Cantwell, and M. Falappa. Credibility-limited revision. *Journal of Symbolic Logic*, 66(4):1581–1596, 2001.

[Konieczny and Pérez, 2000] S. Konieczny and R. Pino Pérez. A framework for iterated revision. *Journal of Applied Non-Classical Logics*, 10(3-4):339–367, 2000.

[Lang, 2004] J. Lang. A preference-based interpretation of other agents' actions. In *Proceedings of KR'04*, pages 644–653, 2004.

[Lehmann and Magidor, 1992] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.

[Lehmann, 1995a] D. Lehmann. Another perspective on default reasoning. *Annals of Mathematics and Artificial Intelligence*, 15(1):61–82, 1995.

[Lehmann, 1995b] D. Lehmann. Belief revision, revised. In *Proceedings of IJCAI'95*, pages 1534–1540, 1995.

[Makinson, 1997] D. Makinson. Screened revision. *Theoria*, 63:14–23, 1997.

[Nayak *et al.*, 2003] A. Nayak, M. Pagnucco, and P. Peppas. Dynamic belief revision operators. *Artificial Intelligence*, 146:193–228, 2003.

[Nebel, 1994] B. Nebel. Base revision operations and schemes: Semantics, representation and complexity. In *Proceedings of ECAI'94*, pages 342–345, 1994.