A Method for Reasoning about Other Agents' Beliefs from Observations

Alexander Nittka[1]
Richard Booth[2]

[1] Institut für Informatik
Universität Leipzig
Johannisgasse 26
04103 Leipzig, Germany

[2] Faculty of Informatics
Mahasarakham University
Kantarawichai
Mahasarakham 44150, Thailand

nittka@informatik.uni-leipzig.de, richard.b@msu.ac.th

Abstract

This paper is concerned with the problem of how to make inferences about an agent's beliefs based on an observation of how that agent responded to a sequence of revision inputs over time. We collect and review some earlier results for the case where the observation is *complete* in the sense that *(i)* the logical content of all formulae appearing in the observation is known, and *(ii) all* revision inputs received by the agent during the observed period are recorded in the observation. Then we provide new results for the more general case where information in the observation might be distorted due to noise or some revision inputs are missing altogether. Our results are based on the assumption that the agent employs a specific, but plausible, belief revision framework when incorporating new information.

1 Introduction

1.1 Motivation

One of the overall goals of AI research is designing autonomous intelligent agents that are capable of acting successfully in dynamic environments. These environments may be artificial or even natural. In any case, it is very likely that they are "inhabited" by more than one agent. So, an agent will in general have to interact with (some of) the others. On the one hand, the agent—if it does not want to be purely reactive—needs a model of its environment in order to make informed choices of actions that change it in a way that brings the agent closer to achieving its goal. On the other, it

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7).* Texts in Logic and Games 3, Amsterdam University Press 2008, pp. 153–182.

also needs to model the other agents, making successful interaction more likely.

Much research has been done on formalising and reasoning about the effects of actions on an environment. Research on an agent's view of the world usually focuses on a first person perspective. How should the agent adapt its beliefs about the world in the light of new information? However, reasoning about *other agents'* beliefs or background knowledge is just as important. This work is intended to contribute to this latter question.

We will adopt a much narrower perspective than reasoning about other agents in their full complexity which includes goals, intentions, (higher order) beliefs, preferences, etc. and restrict our attention to their (propositional) beliefs about the world. We will also forget about the dynamic environment and assume a static world. That is, we will work in a very traditional belief revision setting. But rather than answering the question of how an agent *should* rationally change its beliefs in the light of new information, we address the question of what we can say about an agent we observe in a belief change process.

In [10], the authors use observable actions to draw conclusions about other agents' mental attitudes. But the beliefs of an agent manifest themselves not only in its actions. They may also be observed more directly, e.g., in communication. So indirectly we have access to parts of other agents' belief revision processes. Information they receive is their revision input, responses to that information are a partial description of their beliefs after the revision. From this information we may want to reason about the observed agent. Consider the following scenarios.

- We are directly communicating with another agent, i.e., we are the source of revision inputs for that agent. The feedback provided by the agent will not reflect its entire set of beliefs. To get a more complete picture we may want to infer what else was believed by the agent, what its *background knowledge* might be.

- We observe a dialogue between two or more agents. Beliefs one agent expresses are revision inputs for the others. Due to noise, private messages etc., we might not have access to the entire dialogue—possibly missing some inputs completely. So we have to deal with partial information about the revision inputs.[1] As we might have to deal with the observed agents later, forming a picture of them will be useful.

The information at our disposal for reasoning about another agent $\mathcal{A}$ will be of the following form. We are given a (possibly incomplete) sequence of

[1] This is of course possible in the first case, as well. The communication might take place in several sessions and we do not know which inputs the agent received in between.

(partially known) revision inputs that were received by $\mathcal{A}$. Further we are given information on what the agent believed and did not believe after having received each input. All this information constitutes an observation of the agent. First we will briefly recall results for the case where observations are complete with respect to the revision inputs received by $\mathcal{A}$. These are then used for dealing with the more general case.

The general approach to reasoning about an agent based on observations will be as follows. We assume $\mathcal{A}$ to employ a particular belief revision framework for incorporating revision inputs. We will then try to find a possible initial state of $\mathcal{A}$ that best explains the observation. By initial state we mean $\mathcal{A}$'s epistemic state at the time the observation started. As we do not know the true initial state, we will have to select a reasonable one. This state explains the observation if it yields the beliefs and non-beliefs recorded in the observation given the revision inputs received by the agent. The meaning of *best* in this context will be explained later. The initial state, which can be interpreted as $\mathcal{A}$'s background knowledge, will allow us to reason about beliefs not recorded in the observation.

Many approaches for reasoning about action, belief revision, etc. assume the initial belief state being given and deal with the case of progression through sequences of actions/revision inputs. They say little or nothing about the case where the initial state is not known. In particular with respect to the belief revision literature this work is intended to be a step towards filling this gap.

1.2 Simplifying assumptions

We make several simplifying assumptions which will naturally limit the applicability of the methods developed in this work but at the same time allow a focused analysis of the problem we approach.

As mentioned above, we assume a static world in the sense that the revision inputs and the information about the agent's beliefs refer to the same world. However, it is essential for our work that the revision inputs were received over time. One central point is to exploit having intermediate steps at our disposal. The observed agent itself may only be interested in the final picture of the world. We in contrast want to extract information about the agent from the process of its arriving there.

We restrict ourselves to propositional logic, and all components of an observation are already provided in propositional logic generated from a finite language. That is, we assume that revision inputs, beliefs and non-beliefs are (and are directly observed as) propositional formulae. Agents are assumed to be sincere, i.e., they are not deceptive about their beliefs, although the information may be partial. The observed agent will be referred to as $\mathcal{A}$. We will disregard concepts like (preferences for) sources, competence, context, etc. $\mathcal{A}$ will be assumed to employ a particular belief revision

framework which we describe in detail in Section 2. The only thing that happens during the time of observation is that $\mathcal{A}$ incorporates the revision inputs. In particular, it does not change its revision strategy or learns in any other way. In that sense, we consider the observations to be short term.

We do not investigate *strategies* for extracting as much information about $\mathcal{A}$ as possible. The observing agent simply uses the information provided to reason along the way, being passive in that sense. That is, our focus is not on the *elicitation* of information about other agents; the question of optimising the reasoning process by putting agents in a setting where observations yield the most precise results is another interesting topic which we do not pursue.

From the choice of revision framework it will become apparent that we equate recency with reliability of the information. We are well aware that this is highly debatable. We will briefly address this issue in the conclusion.

For real world applications many of these assumptions have to be dropped or weakened. Many of the issues we disregarded will have to be taken into account. But for the moment we try to keep the number of free variables low in order to give more precise formal results. We hope to convince the reader that even in this very restricted setting we will be able to draw interesting, non-trivial conclusions. Also, we will show that even if these assumptions are correct, there are very strict limitations to what we can *safely* conclude about $\mathcal{A}$.

1.3 Preliminaries

As stated above, the observed agent will be denoted by $\mathcal{A}$. L will be used to denote a propositional language constructed from a finite set of propositional variables $p, q, r, \ldots$, the connectives $\wedge, \vee, \neg, \rightarrow, \leftrightarrow$ and the symbols $\bot$ for some contradiction and $\top$ for some tautology. $\alpha, \beta, \delta, \theta, \lambda, \varphi, \phi, \psi$, and $\blacktriangle$ (often with subscript) will denote propositional formulae, i.e., particular elements of L. In Section 3, χ will be used as placeholder for an unknown formula. $\vdash$ is the classical entailment relation between a set of formulae and a formula, where we abbreviate $\{\alpha\} \vdash \beta$ by $\alpha \vdash \beta$ for singleton sets. $\mathsf{Cn}(S)$ denotes the set of all logical consequences of a set of formulae S, i.e., $\mathsf{Cn}(S) = \{\alpha \mid S \vdash \alpha\}$.

The revision operation $*$ introduced will be left associative and consequently $K * \varphi_1 * \varphi_2$ is intended to mean $(K * \varphi_1) * \varphi_2$. σ and ρ are used to denote sequences of formulae, () being the empty sequence. The function $\cdot$ denotes concatenation, so $\sigma \cdot \rho$ and $\sigma \cdot \alpha$ represents sequence concatenation and appending a formula to a sequence, respectively.

The structure of the paper will be as follows. Section 2 will introduce the assumed agent model as well as the formal definition of an observation. It further recalls the central results for the case where all revision inputs received by $\mathcal{A}$ during the time of observation are completely known, i.e., in particular the method for calculating the best explaining initial state and its

properties. The section thus summarises [5, 6, 7]. It extends these papers by also discussing the question of how safe conclusions we draw about $\mathcal{A}$ are. Section 3 uses these results to deal with the case where the observation is allowed to be more partial. In particular, some inputs may not have been recorded in the observation (see also [23]) and the logical content of parts of the observation may only be partially known. We show how this lack of information can be represented and dealt with. This paper is intended to give a broad overview over our proposed method for reasoning about an observed agent. Hence, we give only short proofs sketches. Full proofs are available in the first author's PhD thesis [24].

2 Belief Revision Framework, Observation and Explanation

2.1 The assumed belief revision framework

We already mentioned that we will assume the agent to employ a particular belief revision framework. The first thing we will do is describe it. As we consider observations of $\mathcal{A}$'s belief revision behaviour over time, it is obvious that such a framework needs to support iterated revision [12, 17, 21]. Further, an observation may imply that a revision input was in fact not accepted. For example it might be explicitly recorded that after being informed that Manchester is the home of the Beatles, the agent does not believe this statement. Consequently, the assumed revision framework should also account for non-prioritised revision [16, 19], i.e., revision where the input is not necessarily believed after revising.

We will assume $\mathcal{A}$ to employ a belief revision framework [3] that is conceptually similar to the approaches in [4, 9, 20, 25] but is able to handle non-prioritised revision as well. The agent's epistemic state $[\rho, \blacktriangle]$ is made up of two components: *(i)* a sequence ρ of formulae and *(ii)* a single formula $\blacktriangle$, all formulae being elements of L. $\blacktriangle$ stands for the agent's set of core beliefs—the beliefs of the agent it considers "untouchable". One main effect of the core belief is that revision inputs contradicting it will not be accepted into the belief set. ρ is a record of the agent's revision history. Revision by a formula is carried out by simply appending it to ρ. The agent's full set of beliefs $\mathsf{Bel}([\rho, \blacktriangle])$ in the state $[\rho, \blacktriangle]$ is then determined by a particular calculation on ρ and $\blacktriangle$ which uses the function f which maps a sequence σ of propositional formulae to a formula. This is done by starting off with the last element of σ and then going backwards through the sequence collecting those formulae that can be consistently added and forgetting about the remaining ones.

Definition 2.1.

$$f(\beta_k,\dots,\beta_1)=\begin{cases}\beta_1 & k=1\\ \beta_k\wedge f(\beta_{k-1},\dots,\beta_1) & k>1\,\&\,\beta_k\wedge f(\beta_{k-1},\dots,\beta_1)\nvdash\bot\\ f(\beta_{k-1},\dots,\beta_1) & \text{otherwise}\end{cases}$$

As hinted at above, iterated revision is handled quite naturally by the framework. All revision steps are simply recorded and the problem of what $\mathcal{A}$ is to believe after each revision step, in particular whether the input just received is accepted, i.e., is believed, is deferred to the calculation of the beliefs in an epistemic state. In order to calculate them the agent starts with its core belief $\blacktriangle$ and then goes backwards through ρ, adding a formula as an additional conjunct if the resulting formula is consistent. If it is not, then the formula is simply ignored and the next element of ρ is considered. The belief set of $\mathcal{A}$ then is the set of logical consequences of the formula thus constructed.

Definition 2.2. The revision operator $*$ is defined for any epistemic state $[\rho,\blacktriangle]$ and formula φ by setting $[\rho,\blacktriangle]*\varphi=[\rho\cdot\varphi,\blacktriangle]$. The belief set $\mathsf{Bel}([\rho,\blacktriangle])$ in any epistemic state $[\rho,\blacktriangle]$ is $\mathsf{Bel}([\rho,\blacktriangle])=\mathsf{Cn}(f(\rho\cdot\blacktriangle))$.

Note, that we do not prohibit the core belief $\blacktriangle$ to be inconsistent in which case $\mathcal{A}$'s belief set is inconsistent. This is the essential difference of to the linear base-revision operator in [22]. From the definition, it is easy to see that $\mathsf{Bel}([\rho,\blacktriangle])$ is inconsistent if and only if $\blacktriangle$ is inconsistent.

Example 2.3. Consider the epistemic state $[(),\neg p]$ of an agent. The beliefs of the agent in this state are $\mathsf{Cn}(f(\neg p))=\mathsf{Cn}(\neg p)$. If q is received as a new input, we get $[(),\neg p]*q=[(q),\neg p]$ as the new epistemic state. The corresponding beliefs are $\mathsf{Cn}(f(q,\neg p))=\mathsf{Cn}(q\wedge\neg p)$.

A further input $q\to p$ changes the epistemic state to $[(q,q\to p),\neg p]$. Note, that $f(q,q\to p,\neg p)=(q\to p)\wedge\neg p$ and q cannot be consistently added, so now the agent believes the logical consequences of $\neg q\wedge\neg p$.

The revision input p changes the epistemic state to $[(q,q\to p,p),\neg p]$ but the beliefs remain unchanged, as p contradicts the core belief.

Given the state $[\rho,\blacktriangle]$ of $\mathcal{A}$ and a sequence $(\varphi_1,\dots,\varphi_n)$ of revision inputs received in that state we can define the *belief trace* of the agent. This is a sequence of formulae characterising the beliefs of $\mathcal{A}$ after having received each of the inputs starting with the beliefs in $[\rho,\blacktriangle]$.

Definition 2.4. Given a sequence $(\varphi_1,\dots,\varphi_n)$ the *belief trace* $(\mathsf{Bel}_0^\rho,\mathsf{Bel}_1^\rho,\dots,\mathsf{Bel}_n^\rho)$ of an epistemic state $[\rho,\blacktriangle]$ is the sequence of formulae $\mathsf{Bel}_0^\rho=f(\rho\cdot\blacktriangle)$ and $\mathsf{Bel}_i^\rho=f(\rho\cdot(\varphi_1,\dots,\varphi_i,\blacktriangle))$, $1\le i\le n$.

The belief trace in the above example is $(\neg p,q\wedge\neg p,\neg q\wedge\neg p,\neg q\wedge\neg p)$.

2.2 Observations

After having formalised the assumptions about any observed agent, we now turn to the specific information we receive about a particular agent $\mathcal{A}$—some observation on its belief revision behaviour. An observation contains information about revision inputs $\mathcal{A}$ received, what it believed and did not believe upon receiving them.

Definition 2.5. An *observation* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ is a sequence of triples $(\varphi_i, \theta_i, D_i)$, where for all $1 \leq i \leq n$: φ_i, θ_i, and all $\delta \in D_i$ (D_i is finite) are elements of a finitely generated propositional language L.

The intuitive interpretation of an observation is as follows. After having received the revision inputs φ_1 up to φ_i starting in some initial epistemic state, $\mathcal{A}$ believed at least θ_i but did not believe any element of D_i. In this section, we assume that during the time of the observation $\mathcal{A}$ received exactly the revision inputs recorded in o, in particular we assume that no input was received between φ_i and φ_{i+1}, the observation being correct and complete in that sense. For the θ_i and D_i we assume the observation to be correct but possibly partial, i.e., the agent did indeed believe θ_i and did not believe any $\delta \in D_i$, but there may be formulae ψ for which nothing is known. In this case we have both $\theta_i \nvdash \psi$ and $\psi \nvdash \delta$ for any $\delta \in D_i$. Note that complete ignorance about what the agent believed after a certain revision step can be represented by $\theta_i = \top$ and complete ignorance about what was not believed by $D_i = \varnothing$.

The observation does not necessarily give away explicitly whether a revision input was actually accepted into $\mathcal{A}$'s belief set or not. If $\theta_i \vdash \varphi_i$ then the revision input φ_i must have been accepted and if $\theta_i \vdash \neg\varphi_i$ or $\varphi_i \vdash \delta$ for some $\delta \in D_i$ then it must have been rejected. But if none of these conditions hold, it is not obvious whether an input has been accepted or rejected. Often, none of these two cases can be excluded. One of the aims of our investigation is to draw more precise conclusions with respect to this question.

A given observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ covers only a certain length of time of the agent's revision history. When the observation started, $\mathcal{A}$ already was in some epistemic state $[\rho, \blacktriangle]$. We will give the formal conditions for an initial state to explain an observation o. The intuitive interpretation of o is formally captured by the system of relations in the second condition of the definition.

Definition 2.6. Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. Then $[\rho, \blacktriangle]$ *explains* o (or *is an explanation for* o) if and only if the following two conditions hold.

1. $\blacktriangle \nvdash \bot$
2. For all i such that $1 \leq i \leq n$:

$$\mathsf{Bel}([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \vdash \theta_i$$

and

$$\forall \delta \in D_i : \mathsf{Bel}([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \nvdash \delta$$

We say $\blacktriangle$ is an *o-acceptable core* iff $[\rho, \blacktriangle]$ explains o for *some* ρ.

For us, an explanation of a given observation o is an epistemic state that verifies the information in o and has a consistent core belief. It is (conceptually) easy to check whether an epistemic state $[\rho, \blacktriangle]$ is an explanation for o. It suffices to confirm that the conditions in Definition 2.6 are satisfied, i.e., that $\blacktriangle$ is consistent and that for all i we have $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \vdash \theta_i$ and $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \nvdash \delta$ for all $\delta \in D_i$. A state with an inconsistent core belief satisfies the second condition if and only if $D_i = \varnothing$ for all i, so there are observations that *could* be explained by such a state. However, we do not consider claiming the agent to be inconsistent worthy of being called an explanation.

Example 2.7. Let $o = \langle (p, q, \varnothing), (q, r, \varnothing) \rangle$ which states that $\mathcal{A}$ after receiving p believes q and after then receiving q believes r. It does not inform us about any non-beliefs of the agent.

$[\rho, \blacktriangle] = [(p \rightarrow q), r]$ explains o because $f(p \rightarrow q, p, r)$ entails q and $f(p \rightarrow q, p, q, r)$ entails r (both are equivalent to $p \wedge q \wedge r$). $[(p \rightarrow q), \top]$ does not explain o because $f(p \rightarrow q, p, q, \top) \equiv p \wedge q \nvdash r$. $[(), p \wedge q \wedge r]$, $[(p \rightarrow q \wedge r), \top]$, $[(\neg p, q, r), s]$, and $[(q \wedge r), \neg p]$ are some further possible explanations for o.

There is never a unique explanation for o, in fact there are infinitely many in case o can be explained. This is why our proposed method for reasoning about $\mathcal{A}$ is to choose one explanation $[\rho, \blacktriangle]$. Using $\blacktriangle$ and the belief trace we then draw our conclusions as follows. Revision inputs consistent with $\blacktriangle$ will be accepted by $\mathcal{A}$, those inconsistent with $\blacktriangle$ are rejected. $\mathcal{A}$'s beliefs after receiving the ith input are characterised by Bel_i^ρ. In Section 2.4 we will discuss the quality of these conclusions and present a method for improving them. But first we have to say how to actually choose one explanation.

2.3 The rational explanation

This section recalls the essential results from [5, 6, 7] for identifying and justifying the best of all possible explanations. A very important property

of the framework is that $\mathcal{A}$'s beliefs after *several* revision steps starting in an initial state can equivalently be expressed as the beliefs after a *single* revision on the same initial state. Intuitively, the agent merges its core belief and all revision inputs received using f into a single formula and then conditions its epistemic state using it.

Proposition 2.8. $\mathsf{Bel}([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = \mathsf{Bel}([\rho, \blacktriangle] * f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$.

Proof (Sketch). Note that by Definition 2.4 it suffices to show that

$$f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \equiv f(\rho \cdot (f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle)).$$

One property of f that follows from its recursive definition is $f(\sigma \cdot \sigma') \equiv f(\sigma \cdot f(\sigma'))$. If we can show that $f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \equiv f(f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle)$ we are done as then in both cases equivalent formulae have been collected before processing ρ. We can restrict our attention to consistent $\blacktriangle$, in which case $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ is consistent and entails $\blacktriangle$. Hence

$$f(f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle) = f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \wedge \blacktriangle \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle).$$

Q.E.D.

How does that help to reason about the observed agent $\mathcal{A}$? Recall that an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ expresses the information that "revision by φ_1 in the initial state leads to a *new* state (where θ_1 but no element of D_1 is believed) in which revision by φ_2 leads to..." That is, the observation contains bits of information concerning beliefs and non-beliefs in *different* (if related) epistemic states. This proposition now allows us to translate the observation into information about a single state—the initial epistemic state we are after. Note however, that $\blacktriangle$ needs to be known for applying the proposition as otherwise $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ cannot be calculated. So, given a core belief $\blacktriangle$, o yields that $\mathcal{A}$ would believe θ_i (and would not believe any $\delta \in D_i$) in case it revised its initial epistemic state by $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$. This is nothing but conditional beliefs held and not held by $\mathcal{A}$ in its initial state $[\rho, \blacktriangle]$. That is, o is a partial description of $\mathcal{A}$'s conditional beliefs in $[\rho, \blacktriangle]$. The proposition further entails that if we had a *full* description of its conditional beliefs we could calculate the beliefs after any sequence of revision inputs.

It turns out that the assumed belief revision framework allows us to apply existing work ([18] and in particular [8]) on completing partial information about conditional beliefs[2] and to construct a suitable ρ such that

[2] [8] presents a *rational closure* construction that takes into account both positive and negative information as is necessary in our case. It extends the case of positive-only information studied in [18]. These papers also inspired the name rational explanation.

$[\rho, \blacktriangle]$ is indeed an explanation for o in case $\blacktriangle$ is o-acceptable. $\rho_R(o, \blacktriangle)$ denotes the sequence thus constructed. The construction even reveals *if* a given core belief is o-acceptable.

We further showed that the set of o-acceptable cores is closed under disjunction. If $\blacktriangle_1$ and $\blacktriangle_2$ are o-acceptable, then so is $\blacktriangle_1 \vee \blacktriangle_2$.[3] This entails that—if o can be explained at all—there is a unique logically weakest o-acceptable core belief, which we denote by $\blacktriangle_\vee(o)$. Consequently $\blacktriangle \vdash \blacktriangle_\vee(o)$ for *any* o-acceptable $\blacktriangle$. The rationale behind choosing $\blacktriangle_\vee(o)$ for an explanation is that any input we predict to be rejected by $\mathcal{A}$ will indeed be rejected. Furthermore, it can be shown that adding beliefs or non-beliefs to o by strengthening some θ_i or enlarging some D_i as well as appending observations to the front or the back of o to get an observation o' cannot falsify this conclusion as $\blacktriangle_\vee(o') \vdash \blacktriangle_\vee(o)$. For any other core belief explaining o, a revision input predicted to be rejected by the agent might in fact be accepted. In this sense, we consider $\blacktriangle_\vee(o)$ to be optimal.

The choice of $\rho_R(o, \blacktriangle_\vee(o))$, which we call the *rational prefix*, as the sequence in the agent's initial epistemic state is justified by showing that it yields an optimal belief traces. Let $\rho = \rho_R(o, \blacktriangle_\vee(o))$ and σ be the sequence of any other explanation $[\sigma, \blacktriangle_\vee(o)]$ for o, $(\mathsf{Bel}_0^\rho, \mathsf{Bel}_1^\rho, \ldots, \mathsf{Bel}_n^\rho)$ and $(\mathsf{Bel}_0^\sigma, \mathsf{Bel}_1^\sigma, \ldots, \mathsf{Bel}_n^\sigma)$ be the corresponding belief traces. Then the following holds: If $\mathsf{Bel}_j^\rho \equiv \mathsf{Bel}_j^\sigma$ for all $j < i$ then $\mathsf{Bel}_i^\sigma \vdash \mathsf{Bel}_i^\rho$.[4] This tells us that the formulae we predict the agent to believe initially will indeed be believed (although some further formulae might be believed as well)—provided the agent's core belief really is $\blacktriangle_\vee(o)$. And if our predicted belief trace exactly captures the agent's beliefs up to the ith input then again all beliefs predicted after the next input will indeed be believed. The assumption that the two explanations use the same core belief causes this criterion, which we will refer to as the optimality criterion for the rational prefix, to be a rather weak one as we will see shortly.

In [7], we defined $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ to be the *rational explanation* of an observation o—if there is an explanation at all. That paper and [5] contain more results about the rational explanation but these are the most important ones which justify the claim that the rational explanation is the best explanation for a given observation o. An algorithm which calculates the rational explanation is given below and described in more detail in [6]. The problem with calculating $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ is that $\blacktriangle_\vee(o)$ has to be known, which it is not in the beginning. So the idea is to iteratively refine the core belief starting with the weakest possible of all $\top$.

[3] The proof is constructive and not deep but lengthy.

[4] This result, as almost all the others in this section, is proved in [5]. Note also that $\mathsf{Bel}_i^\rho \vdash \mathsf{Bel}_i^\sigma$ need not hold. Consider $[(\neg p), \top]$ and $[(p \wedge q, \neg p), \top]$. The belief traces when assuming a single input p are $(\neg p, p)$ and $(\neg p, p \wedge q)$. Although the beliefs are equivalent initially, they need not be after a revision step.

Algorithm 1: Calculation of the rational explanation.

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

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

Having calculated the rational explanation $[\rho, \blacktriangle]$ of an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$, we can make predictions concerning which inputs $\mathcal{A}$ accepts and rejects based on $\blacktriangle$ and our conclusions about its beliefs after having received each input are summarised by the corresponding belief trace $(\mathsf{Bel}^\rho_0, \mathsf{Bel}^\rho_1, \ldots, \mathsf{Bel}^\rho_n)$.

2.4 Safe conclusions and hypothetical reasoning

In the remainder of this section, we will illustrate some limitations of the rational explanation. In particular, we will show that predictions based on it will almost never be safe ones. However, this is inherent in the problem and not due to our solution to it.

As with many optimisation problems the quality of the solution and the conclusions we can draw from it depend heavily on the quality of the data and validity of the assumptions made. In our case, we clearly stated the assumptions made about the given observation as well as the agent's being ruled by the assumed framework. The optimality result for the best explaining core belief $\blacktriangle_\vee(o)$, i.e., that $\blacktriangle_\vee(o)$ is entailed by any o-acceptable core, depends on those. The optimality of the rational prefix $\rho_R(o, \blacktriangle)$ and therefore the conclusions about $\mathcal{A}'s$ further beliefs also depend on its actually employing the assumed core belief. That is, if we cannot be sure of the agent's actual core belief then most of what we can say about the agent's belief trace based on the rational explanation is merely justified guesses but not safe bets.

Example 2.9. *(i)* Let $o = \langle(\top, p, \varnothing), (\neg p, \top, \varnothing), (r \leftrightarrow \neg p, r \vee p, \varnothing)\rangle$. The rational explanation for o is $[(p), \top]$ and the corresponding belief trace is $(p, p, \neg p, r \wedge \neg p)$. That is, we conclude that $\mathcal{A}$ accepted the input $\neg p$ and believes $r \wedge \neg p$ after then receiving $r \leftrightarrow \neg p$.

Now assume the agent's real initial belief state was $[(), p]$—note that the core belief does not correspond to the one calculated by the rational explanation—and thus the belief trace in truth is $(p, p, p, \neg r \wedge p)$. That is, it did *not* accept the input $\neg p$ and believed $\neg r \wedge p$ after receiving $r \leftrightarrow \neg p$.

So except for the beliefs before the observation started and after receiving the tautology (where we are informed that the agent believes p and hence must have believed it initially) most of the conclusions about beliefs held by $\mathcal{A}$ we draw from the belief trace are wrong!

(ii) Let $o = \langle (p, p, \varnothing), (q, q, \varnothing), (r \leftrightarrow p, \top, \varnothing)\rangle$. The rational explanation for o is $[(), \top]$ and the belief trace implied by that explanation is $(\top, p, p \wedge q, p \wedge q \wedge r)$. Assuming that $[(), q \rightarrow \neg p]$ was $\mathcal{A}$'s true initial state, the belief trace in truth is $(q \rightarrow \neg p, p \wedge \neg q, q \wedge \neg p, q \wedge \neg p \wedge \neg r)$. Again, for large parts the conclusions we draw about the agent's beliefs based on the rational explanation are wrong. For example, we conclude that agent continues to believe p once it has been received. This is clearly not the case.

This strong dependence on the core belief can be easily explained. There are two main effects due to the core belief. First, it causes revision inputs to be rejected immediately. This is why the conclusions based on the rational explanation are off the mark in case *(i)* in the above example. Secondly, the core also accounts for interactions between revision inputs. An earlier input is eliminated from the belief set in the light of the core and some later inputs. This effect is illustrated in case *(ii)*. For one choice of the core belief, after having received the input φ_{i+j}, the agent may still believe the input φ_i received earlier, while for another core it may believe $\neg\varphi_i$.

Even if we got the core belief right and hence the agent really employs $\blacktriangle_\vee(o)$, conclusions based on the rational explanation of o should not be used without care. The optimality result for the rational prefix does not exclude mistakes. Correct conclusions about beliefs are guaranteed only up to the point in the belief trace where the beliefs we calculate and the agent's actual ones first fail to be equivalent. This can easily be the case already for the initial beliefs.

Consider $o = \langle (p, q, \varnothing), (r, \top, \varnothing)\rangle$ for which the rational explanation is $[(p \rightarrow q), \top]$, the corresponding belief trace being $(p \rightarrow q, p \wedge q, p \wedge q \wedge r)$. So we would conclude the agent to keep believing in q. If the agent's real initial epistemic state was $[(\neg q, \neg r \wedge q), \top]$ then the real belief trace would be $(\neg r \wedge q, p \wedge \neg r \wedge q, r \wedge \neg q)$. Although the correct core was calculated, we would still be wrong about q whose negation is in fact believed after having received the input r.

As stated above, using the rational explanation $[\rho, \blacktriangle_\vee(o)]$ we conclude that $\mathcal{A}$ believed $\mathsf{Bel}_i^\rho = f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_\vee(o)))$ after having received the first i revision inputs recorded in o. How safe is this conclusion? The above example showed that it is not very safe. So what can we do to further improve the results?

Here, we will consider only one very strong notion. We call the conclusion that $\mathcal{A}$ believes ψ after receiving the ith revision input recorded in o safe if and only if for *all* explanations for o we have $\mathsf{Bel}_i \vdash \psi$, where

Bel_i is the element of the belief trace corresponding to that input. In other words, every possible explanation predicts that belief (so in particular the one corresponding to the agent's real initial state). Analogously, we call the conclusion that the agent did *not* believe ψ at a certain point safe whenever *no* explanation predicts that formula to be believed. Note that a safe conclusion about an agent's belief does not mean that this belief is correct. The agent may have received and accepted unreliable information, but it means that given the observation, the agent must have held this belief.

We will now describe a way to calculate whether a conclusion of that form is safe, a method we call *hypothetical reasoning*. By this we mean modifying the given observation according to some conjecture and rerunning the rational explanation construction on the observation thus obtained. Note that any explanation for

$$o' = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i \wedge \psi, D_i), \ldots, (\varphi_n, \theta_n, D_n)\rangle \text{ or}$$
$$o' = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i, D_i \cup \{\psi\}), \ldots, (\varphi_n, \theta_n, D_n)\rangle$$

will also explain $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i, D_i), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. This follows directly from Definition 2.6. If $\theta_i \wedge \psi$ belongs to the beliefs after the ith revision step then so does θ_i and if none of the elements of $D_i \cup \{\psi\}$ is believed at that point, the same holds for any subset.

So in order to check whether the conclusion of $\mathcal{A}$ believing ψ after receiving the ith revision input is a safe one, we simply add ψ to D_i and test whether the observation thus obtained has an explanation.[5] If so, then the conclusion is not safe as there is an explanation where ψ is in fact not believed. However, if no such explanation exists then ψ must indeed be believed by $\mathcal{A}$. The non-belief of a formula ψ can be verified by replacing the corresponding θ_i by $\theta_i \wedge \psi$. If the observation thus obtained has an explanation then the agent may have believed ψ and consequently the conclusion is not safe.

With a small modification this method works also for hypothetical reasoning about the agent's initial beliefs, i.e., before receiving the first input. It does not work directly, as the observation does not contain an entry for the initial state. By appending $\langle(\varphi_0, \theta_0, D_0)\rangle = \langle(\top, \top, \varnothing)\rangle$ to the front of the observation o we create this entry. The point is that receiving a tautology as input leaves the beliefs unchanged. We can now add a formula ψ to D_0 or θ_0 as described above to verify conclusions about the initial state. Further, there is no restriction which formulae ψ can be used for hypothetical reasoning. It is even possible to add several ψ_j simultaneously to o to get a modified observation o'.

[5] The rational explanation algorithm always finds an explanation if there is one, and returns "no explanation" if there is none.

Hypothetical reasoning can also be used in order to improve the conclusions about $\mathcal{A}$'s core belief $\blacktriangle$. We already know that $\blacktriangle \vdash \blacktriangle_\vee(o)$, i.e., all inputs we predict to be rejected by $\mathcal{A}$ will indeed be rejected. This is because any o-acceptable core entails $\blacktriangle_\vee(o)$. But what about the other inputs, must they really have been accepted? Can we be sure that φ_i really was accepted if it is consistent with $\blacktriangle_\vee(o)$? Rejecting φ_i is equivalent to not believing the input after having received it. So, we simply add φ_i to D_i, i.e., replace $(\varphi_i, \theta_i, D_i)$ in o by $(\varphi_i, \theta_i, D_i \cup \{\varphi_i\})$ to get o'. If there is an o'-acceptable core, then $\mathcal{A}$ may in fact have rejected φ_i. However, if o' does not have an explanation then we know that $\mathcal{A}$ must have accepted that input.

It might be nice to be able to check whether conclusions about $\mathcal{A}$'s beliefs are safe, but can we ever be sure to have the correct core belief in order to apply the optimality results we gave for the rational prefix? The answer to this question is almost exclusively negative. Usually, there is more than one o-acceptable core. In a different context Sébastien Konieczny[6] suggested the additional assumption that the last belief θ_n recorded in the observation o is in fact *complete*. This assumption gives us an upper bound on the actual core belief $\blacktriangle$ as then $\theta_n \vdash \blacktriangle \vdash \blacktriangle_\vee(o)$ must hold and we can use the hypothetical reasoning methodology in order to get an improved core belief. As we know the exact belief θ_n at the end of the observation, we can iteratively add to D_n those formulae ψ which the rational explanation predicts to be believed but which are not entailed by θ_n. This method will yield an improved lower bound for the core belief of the agent, but it cannot guarantee uniqueness of the core.

Even if we assumed that *every* θ_i completely characterises the beliefs of the agent after receiving φ_i, we would not be guaranteed to get the real core belief. Consider $o = \langle (p, p, \varnothing), (q, p \wedge q, \varnothing), (r, p \wedge q \wedge r, \varnothing) \rangle$ to illustrate this. The rational explanation for o is $[(), \top]$. However, p is also an o-acceptable core, $[(), p]$ being one possible explanation. That is, the conclusion that an input $\neg p$ will be accepted by the agent is not safe. This illustrates that even using much more severe assumptions about a given observation, identifying the agent's real core belief is impossible.

3 Extension to Unknown Subformulae

Up to this point we considered observations that were complete with respect to the revision inputs received. We knew exactly which inputs were received during the time of observation. The scenarios in the introduction suggested that it is well possible that some of the inputs might have been missed. Further, the observer may not understand the complete *logical content* of all the revision inputs, $\mathcal{A}$'s beliefs and non-beliefs. Consider the following example

[6] Personal communication.

where the agent is observed to receive *exactly* two inputs p and q. After hearing p, the agent believed *something we cannot understand*, but after then hearing q, it did not believe that anymore. In the original framework this cannot be formalised as there is no means to represent the unknown belief. However, we should be able to conclude that $\mathcal{A}$ believed $\neg q$ after having received p. This is because the assumed belief revision framework satisfies (most of) the AGM postulates [1]. In particular, if the input is consistent with the current beliefs they have to survive the revision process (cf. the "Vacuity" postulate from AGM) which is clearly not the case in the example. The current section investigates how the previous results can still be used to reason about $\mathcal{A}$ if the observations are allowed to be less complete in this sense.

We want to emphasise that there is a big difference between knowing there was a revision input while being ignorant about its logical content and not even knowing whether there were one or more revision inputs. We will first deal with the former case which will provide results to deal with the latter one in Section 3.3.

3.1 Modelling unknown logical content

We will model partial information by allowing formulae appearing in the observation to contain unknown subformulae which are represented by n placeholders χ_j. $\lambda(\chi_1, \ldots, \chi_n)[(\chi_i/\phi_i)_i]$ denotes the result of replacing in λ every occurrence of χ_i by ϕ_i.

Definition 3.1. Let L be a propositional language and $\chi_1, \ldots, \chi_n$ be placeholders not belonging to L.

A "formula" $\lambda(\chi_1, \ldots, \chi_n)$ possibly containing $\chi_1, \ldots, \chi_n$ is called a *parametrised formula based on* L iff $\lambda(\chi_1, \ldots, \chi_n)[(\chi_i/\phi)_i] \in L$ whenever $\phi \in L$. $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_l, \theta_l, D_l)\rangle$ is a *parametrised observation based on* L iff all φ_i, θ_i, $\delta \in D_i$ are parametrised formulae based on L. We denote by $L(o)$ the smallest language L a parametrised observation o is based on.

To put it differently, a parametrised formula based on L is a formula from L in which some subformulae have been replaced by placeholders χ_i. This allows hiding parts of the logical content of a formula. So in order to model (even more) partial knowledge, we will consider parametrised observations. The example from the introductory paragraph can now be represented by $o = \langle(p, \chi, \varnothing), (q, \top, \{\chi\})\rangle$. We will often write λ rather than $\lambda(\chi_1, \ldots, \chi_n)$ to denote a parametrised formula in order to ease reading.

Unknown subformulae χ_i are allowed to appear in all components of an observation—revision inputs, beliefs and non-beliefs. The same χ_i can appear several times. In fact, this is when it contributes to the reasoning

process. It is not unreasonable to assume that this can happen. For example, the meaning of an utterance in a dialogue might not be understood as part of the language may not be known to the observing agent, but the utterance might be recognised when it appears again later. Analogous to a learner of a foreign language, the observer may be familiar with (parts of) the structure of the language while being ignorant about the meaning of certain "phrases". In case we are completely ignorant about the logical content the entire parametrised formula will simply be a placeholder.

Let o be a parametrised observation. $o[\chi_1/\phi_1, \ldots, \chi_n/\phi_n]$ and equivalently $o[(\chi_i/\phi_i)_i]$ denote the observation obtained by replacing in o every occurrence of the placeholder χ_i by a formula ϕ_i.

We still assume correctness of the information contained in the parametrised observation o, i.e., we assume the existence of instantiations ϕ_i of all unknown subformulae χ_i such that the observation $o[(\chi_i/\phi_i)_i]$ is a correct observation in the sense of Section 2—in particular, there must be an entry for every revision input received. The agent indeed received exactly the inputs recorded and beliefs and non-beliefs are correct if partial. Note that this implies that we are not yet able to deal with *missing* inputs. These will be considered in Section 3.3. One important technical restriction is that the instantiations of unknown subformulae χ_i must not contain unknown subformulae χ_j themselves, i.e., the instantiations must be elements of the underlying language—however, not necessarily elements of $L(o)$. That is, the true meaning of χ_i is not assumed to be expressible in the language of the known part of o. Abusing notation we will write that o has an explanation, meaning that there exist instantiations $\phi_1, \ldots, \phi_n$ for the unknown subformulae such that $o[(\chi_i/\phi_i)_i]$ has an explanation; similarly that $\blacktriangle$ is o-acceptable if $\blacktriangle$ is $o[(\chi_i/\phi_i)_i]$-acceptable.

3.2 Finding an acceptable core belief

In this section, we will present results on what can be said about $\mathcal{A}$'s core belief given a parametrised observation o. If an explanation exists at all, once more there will be a unique weakest o-acceptable core $\blacktriangle$. This may be surprising as there are many different possible instantiations for the unknown subformulae. But this will also allow us to choose them such that any o-acceptable core entails $\blacktriangle$. If we knew the instantiations of the unknown subformulae we could simply use the rational explanation algorithm, as in that case a parametrised observation could be transformed into a regular one. As we do not know them, we have to guess. The trick is to extend the language and treat every χ_i as a new propositional variable x_i.

Proposition 3.2. If $[\rho, \blacktriangle]$ explains $o[(\chi_i/\phi_i)_i]$ and $x_1, \ldots, x_n$ are propositional variables not appearing in o, $\blacktriangle$, ρ or any ϕ_i then $o[(\chi_i/x_i)_i]$ is explained by $[\rho, \blacktriangle \wedge \bigwedge_{1 \leq i \leq n}(x_i \leftrightarrow \phi_i)]$.

Proof (Sketch). $\lambda[(\chi_i/\phi_i)_i] \vdash \bot$ iff $\bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i) \wedge \lambda[(\chi_i/x_i)_i] \vdash \bot$ for any parametrised formula λ not containing x_i is the key to this result. As the x_i are not contained in $[\rho, \blacktriangle]$ or $o[(\chi_i/\phi_i)_i]$, requiring $\bigwedge(x_i \leftrightarrow \phi_i)$ ensures that the different instantiations have the same logical consequences—modulo entailment of irrelevant formulae containing the x_i. The (relevant) beliefs are the same for both explanations. Q.E.D.

The proposition formalises that given there is *some* instantiation for the unknown subformulae in o such that the resulting observation has an explanation, we can also replace them by new variables and still know that there is an explanation. However, this tells us that we can apply the rational explanation algorithm to $o[(\chi_i/x_i)_i]$ and be guaranteed to be returned an explanation if there is one. If this fails, i.e., we are returned an inconsistent core belief, then no explanation can exist using any instantiation of the unknown subformulae in o. The core belief $\blacktriangle \wedge \bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i)$ ensures that the new variables x_i behave exactly as the "correct" instantiations ϕ_i for the unknown subformulae in o.

In general, $\blacktriangle_\vee(o[(\chi_i/x_i)_i])$—the core belief returned by the rational explanation algorithm—will not be that particular formula. Note that this would be impossible as there can be several suitable instantiations such that $o[(\chi_i/\phi_i)_i]$ has an explanation. The core belief calculated will in general be weaker but may still contain (some of) the additional variables x_i. We will now go on to show that it is possible to eliminate these variables from the core belief by choosing different instantiations for the unknown subformulae.

The idea is to split the core $\blacktriangle$ calculated by the rational explanation construction into two parts, one $\blacktriangle'$ that talks only about $L(o)$ and not at all about the additional variables and one ψ part that talks also about those. Formally, we choose $\blacktriangle'$ and ψ such that $\blacktriangle \equiv \blacktriangle' \wedge \psi$ and $\mathsf{Cn}(\blacktriangle') = \mathsf{Cn}(\blacktriangle) \cap L(o)$, which is possible as we are in a finite setting.[7] Instead of x_i we then use $x_i \wedge \psi$ to instantiate the placeholders. This shifting of parts of the core to the new variables is possible because the part of the core belief that talks about the x_i becomes relevant in the calculation of the beliefs of an agent only when those variables themselves appear.

Proposition 3.3. If $[\rho, \blacktriangle]$ explains $o[(\chi_i/x_i)_i]$ then there exist $\blacktriangle'$ and ψ such that $\blacktriangle'$ contains no x_i and $[\rho \cdot \psi, \blacktriangle']$ explains $o[(\chi_i/x_i \wedge \psi)_i]$.

Proof (Sketch). Let $\psi = \blacktriangle$ and $\blacktriangle'$ such that $\mathsf{Cn}(\blacktriangle') = \mathsf{Cn}(\blacktriangle) \cap L(o)$. It can now be shown that $f(\varphi_1[(\chi_i/x_i)_i], \ldots, \varphi_j[(\chi_i/x_i)_i], \blacktriangle)$ is equivalent to $f(\blacktriangle, \varphi_1[(\chi_i/x_i \wedge \blacktriangle)_i], \ldots, \varphi_j[(\chi_i/x_i \wedge \blacktriangle)_i], \blacktriangle')$. Again, the proof of that is

[7] We can trivially choose $\psi = \blacktriangle$ and $\blacktriangle'$ to represent all logical consequences of $\blacktriangle$ in $L(o)$.

not deep but lengthy. The intuition is that $\blacktriangle'$ makes sure that with respect to $L(o)$ all formulae are treated correctly and using $x_i \wedge \blacktriangle$ rather than just x_i, the effect of $\blacktriangle$ with respect to new variables is maintained. Consequently, before processing ρ when calculating the beliefs, in both cases equivalent formulae have been constructed and the beliefs will thus be equivalent. Q.E.D.

To summarise what we know so far. Given a parametrised observation o has an explanation, we can construct one for $o[(\chi_i/x_i)_i]$. However, the corresponding core belief $\blacktriangle_\vee(o[(\chi_i/x_i)_i])$ may still contain variables that are not contained in $L(o)$ and thus we cannot claim that the agent had contact with them. The last proposition now showed that we can construct instantiations for the unknown subformulae such that the explaining core belief $\blacktriangle'$ is in $L(o)$. We can even go one step further and show that *any* $o[(\chi_i/\phi_i)_i]$-acceptable core $\blacktriangle''$ must entail the core $\blacktriangle'$ constructed as described above.

Proposition 3.4. Let $[\rho'', \blacktriangle'']$ be an explanation for $o[(\chi_i/\phi_i)_i]$ and $[\rho, \blacktriangle]$ be the rational explanation for $o[(\chi_i/x_i)_i]$, where x_i are additional propositional variables not appearing in any ϕ_i, $\blacktriangle''$ or the language $L = L(o)$. Further let $\blacktriangle'$ such that $\mathsf{Cn}(\blacktriangle') = \mathsf{Cn}(\blacktriangle) \cap L$. Then $\blacktriangle'' \vdash \blacktriangle'$.

Proof. By Proposition 3.2 $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i)$ is $o[(\chi_i/x_i)_i]$-acceptable and hence entails $\blacktriangle$ (any o-acceptable core entails $\blacktriangle_\vee(o)$). Obviously $\blacktriangle \vdash \blacktriangle'$, so $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i) \vdash \blacktriangle'$. Now assume $\blacktriangle''$ does not entail $\blacktriangle'$ which implies there is a model for $\blacktriangle'' \wedge \neg\blacktriangle'$. Neither $\blacktriangle''$ nor $\blacktriangle'$ contain any x_i so we can extend that model to one for $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i) \wedge \neg\blacktriangle'$ by evaluating x_i just as ϕ_i—contradicting $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i) \vdash \blacktriangle'$. Q.E.D.

There is an important consequence of that result. As in the original case there is a unique weakest o-acceptable core for a parametrised observation o. This follows directly from the last two propositions. $\blacktriangle'$, being constructed as described above, is o-acceptable and is entailed by any o-acceptable core, so in particular by the agent's real core belief. Hence, all formulae inconsistent with $\blacktriangle'$ will be rejected by $\mathcal{A}$. That is, $\blacktriangle'$ yields a safe conclusion with respect to which formulae must be rejected by $\mathcal{A}$—no matter what the instantiations of the unknown subformulae really are.

Example 3.5. Consider $o = \langle(\chi, \chi, \varnothing), (p, q \wedge \neg\chi, \varnothing)\rangle$. This parametrised observation expresses that the observed agent accepted an input whose meaning is unknown to us. After then receiving p, it believed q and the negation of the unknown input. The observation constructed according to Proposition 3.2, where χ is replaced by a new variable x, is $o[\chi/x] = \langle(x, x, \varnothing), (p, q \wedge \neg x, \varnothing)\rangle$. The rational explanation for $o[\chi/x]$ is

$[(p \wedge \neg x \rightarrow q), p \rightarrow \neg x]$ and $(p \rightarrow (\neg x \wedge q), x \wedge \neg p, p \wedge q \wedge \neg x)$ is the corresponding belief trace.

This indicates that after receiving the unknown input the agent believes $\neg p$. In order to test whether this is necessarily the case, we investigate the parametrised observation $o' = \langle(\chi, \chi, \{\neg p\}), (p, q \wedge \neg\chi, \varnothing)\rangle$. According to the hypothetical reasoning methodology, $\neg p$ was added to the non-beliefs. Applying the rational explanation algorithm yields that $o'[\chi/x]$ has no explanation. Proposition 3.2 now tells us that there cannot be an explanation for o'—no matter how χ is instantiated. That is, if the parametrised observation correctly captures the information about the agent, it must believe $\neg p$ after receiving the first input.

o is based on the language L constructed from the variables p and q and $\mathsf{Cn}(p \rightarrow \neg x) \cap L = \mathsf{Cn}(\top)$. To illustrate Propositions 3.3 and 3.4 note that $o[\chi/x \wedge (p \rightarrow \neg x)] = \langle(x \wedge \neg p, x \wedge \neg p, \varnothing), (p, q \wedge (\neg x \vee p), \varnothing)\rangle$ is explained by $[(p \wedge \neg x \rightarrow q, p \rightarrow \neg x), \top]$, the corresponding belief trace being $(p \rightarrow (\neg x \wedge q), x \wedge \neg p, p \wedge q \wedge \neg x)$. $\top$ is trivially entailed by any o-acceptable core.

In order to find an acceptable core for a parametrised observation o, we extended the language $L(o)$ with new variables. In [23], we gave an example—which we will not repeat here—illustrating that there are parametrised observations that have an explanation when language extension is allowed but which cannot be explained restricting the language to $L(o)$. In other words, the proposed algorithm of replacing each χ_i by a new variable x_i, running the rational explanation construction and then eliminating the x_i from the core belief (the result being $\blacktriangle'$) may yield an explanation, although none exists when restricting the instantiations of the χ_i to $L(o)$. Although we know that each acceptable core will entail $\blacktriangle'$, we cannot generally say that restricting the instantiations to $L(o)$ will allow the same core, a strictly stronger one or none at all to explain o.

Note that Proposition 3.2 makes no assumption about the language of the instantiations ϕ_i of the unknown subformulae. They may or may not belong to $L(o)$. They may contain arbitrarily (but finitely) many propositional variables not belonging to $L(o)$. However, that proposition has an interesting implication. It says if $o[(\chi_i/\phi_i)_i]$ has an explanation then so does $o[(\chi_i/x_i)_i]$, but $o[(\chi_i/x_i)_i]$ contains only variables from $L(o)$ and n additional variables x_i, one for each placeholder. As that observation has an explanation, the rational explanation construction will return one. However, that construction uses only formulae present in the observation. Consequently, it does not invent new variables. So, no matter how many variables not appearing in $L(o)$ were contained in the ϕ_i, n additional variables suffice for finding an explanation for the parametrised observation o. This yields an upper bound on additional variables needed.

In Section 2.4 we showed that assuming the wrong core belief greatly affects the quality of the conclusions about $\mathcal{A}$'s other beliefs. And even if the core is correct, the belief trace implied by the rational explanation does not necessarily yield only safe conclusions with respect to the beliefs of the agent during the observation.

These problems are obviously inherited by the current extension to partial information about the logical content of the formulae in an observation. They cannot be expected to become less when not even knowing what inputs the agent really received or when information about the beliefs and non-beliefs becomes even more vague. Much depends not only on the core belief but also on the instantiation of the unknown subformulae. So rather than just having to calculate a best initial epistemic state, we now would also have to find an optimal instantiation of the unknown subformulae. However, the limitations illustrated in Section 2.4 prevent us from even attempting to look for them. Instead, we propose to investigate the belief trace implied by the rational explanation of an observation $o[(\chi_i/x_i)_i]$ and reason hypothetically about beliefs and non-beliefs from $L(o)$ in that belief trace.

3.3 Intermediate inputs

Up to now, we assumed the (parametrised) observation o to contain an entry (φ, θ, D) for every revision input received by $\mathcal{A}$, even if some of the formulae are only partially known. This corresponds to the assumption of having an eye on the agent at all times during the observation. In this section, we want to drop this assumption. That is, we will allow for intermediate inputs between those recorded in o. In real applications this will be the norm rather than an exceptional case. $\mathcal{A}$ or the observing agent may leave the scene for a time, and if the observing agent is the source of information then o might have been gathered over several sessions between which $\mathcal{A}$ may have received further input.

Using our notation for observations, an intermediate input is one we have no information about, i.e., we do not know what the revision input is or what is believed or not believed after receiving it. Hence, we can represent it by $\langle(\chi, \top, \varnothing)\rangle$; χ again represents an unknown formula. Note that this is different from $\langle(\chi, \chi, \varnothing)\rangle$ as here the input would be required to be accepted by $\mathcal{A}$. In other words, the agent's core belief would have to be consistent with the instantiation of χ.

Example 3.6. Consider the following observation without intermediate inputs: $o = \langle(p, q, \varnothing), (p, \neg q, \varnothing)\rangle$. Assume $\blacktriangle$ was o-acceptable and thus consistent. Then either it is consistent or inconsistent with p. In both cases, the belief set does not change upon receiving the second input p. Either the first p was accepted and hence already believed or p was rejected (both

times) in which case the belief set never changes. So $\neg q$ must have been believed already after the first p was received. But it is not possible to believe $q \wedge \neg q$ consistently (the belief set is inconsistent if and only if the core belief is inconsistent). Consequently, there is no o-acceptable core.

Assuming a single intermediate input $\langle(\chi, \top, \varnothing)\rangle$, there is only one reasonable position yielding $o' = \langle(p, q, \varnothing), (\chi, \top, \varnothing), (p, \neg q, \varnothing)\rangle$. Instantiating the unknown formula χ with $p \rightarrow \neg q$, $[(p \rightarrow q), \top]$ is an explanation. Before receiving the first input the agent believes $p \rightarrow q$, after receiving the first p it believes $p \wedge q$ and after receiving the (assumed) intermediate input as well as after receiving the last input p it believes $p \wedge \neg q$. Hence $\top$ is o'-acceptable. That is, while o does not have an explanation, assuming an intermediate input allows the observation to be explained.

In the general case we do not know how many intermediate inputs were received at which points in a given (parametrised) observation o. In [23] we showed that number and positions of the intermediate inputs have an impact on the possible explanations of o. If number and positions are fixed then we deal with a parametrised observation (containing an entry for every revision input received) and can hence use the results of Section 3.2 in order to calculate the weakest acceptable core belief. To represent the intermediate inputs we simply have to introduce *further* unknown subformulae not contained in o. Assume we have the partial observation $o = \langle(p, q \wedge \chi_1, \varnothing), (r, \neg q, \varnothing), (p, q, \{\chi_1\})\rangle$ and the information that exactly two intermediate inputs have been received immediately after r. In order to reason about $\mathcal{A}$, we consider the partial observation $o' = \langle(p, q \wedge \chi_1, \varnothing), (r, \neg q, \varnothing), (\chi_2, \top, \varnothing), (\chi_3, \top, \varnothing), (p, q, \{\chi_1\})\rangle$ which now contains an entry for every input received. At this point we want to emphasise once more that intermediate inputs and partial information about inputs are related but distinct cases.

In the following we want to indicate what can be said about the agent's core belief depending on how much information we have concerning possible intermediate inputs. Naturally, the more specific our knowledge concerning number and positions, the more informative the conclusions can be. We will start with the case where we have no information at all, which means that any number of intermediate inputs may have been received any time. Then we will turn to the cases where the positions or the number are restricted.

Any number of intermediate inputs at any time. Consider an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. Assume $[\rho, \blacktriangle]$ explains the observation o' which is obtained from o by putting some arbitrary number of intermediate inputs at any position in o. It can be proved that then there are a sequence σ and $n-1$ intermediate inputs such that $[\sigma, \blacktriangle]$ explains o'' obtained from o by putting exactly one intermediate input between any two inputs φ_i and φ_{i+1} in o. Note that both explanations use the same core

belief. Intuitively, the intermediate input from o'' before the input φ_{i+1} is the conjunction of *all relevant* intermediate inputs from o' before that input.

Proposition 3.2 tells us that we can also use new variables x_i instead of those intermediate inputs[8] and the observation o''' thus obtained is guaranteed to have an explanation. However, o''' does not contain any unknown subformulae, so we can apply the rational explanation construction which will return some epistemic state with core belief $\blacktriangle'$. We can now construct the weakest possible core belief by taking $\blacktriangle''$ such that $\mathsf{Cn}(\blacktriangle'') = \mathsf{Cn}(\blacktriangle') \cap L(o)$. Any o'-acceptable core belief—o' being constructed as described above—will entail $\blacktriangle''$. That is from $\blacktriangle''$ we can safely conclude which formulae are rejected by $\mathcal{A}$, no matter how many intermediate inputs it received at any point during the observation.

What happens if we have further information about the positions or the number of intermediate inputs? The following proposition implies that we should always assume the maximal number of intermediate inputs. It says that an additional intermediate input, which we instantiate with a new variable for calculating the weakest possible core belief, can only make the core logically weaker. Conversely, not assuming the maximal number of intermediate inputs may lead to the conclusion that $\mathcal{A}$ rejects a formula which it actually does not reject simply because an additional intermediate input allows $\mathcal{A}$'s core belief to be logically weaker.

Proposition 3.7. If $\mathsf{Cn}(\blacktriangle) = \mathsf{Cn}(\blacktriangle_\vee(o_1 \cdot \langle(x, \top, \varnothing)\rangle \cdot o_2)) \cap L(o_1 \cdot o_2)$ and $x \notin L(o_1 \cdot o_2)$ then $\blacktriangle_\vee(o_1 \cdot o_2) \vdash \blacktriangle$.

Proof (Sketch). By showing $\blacktriangle_\vee(o_1 \cdot o_2) \equiv \blacktriangle_\vee(o_1 \cdot \langle(\top, \top, \varnothing)\rangle \cdot o_2)$, which holds because a tautologous input has no impact, we introduce the extra input which allows us to compare the cores. By Proposition 3.2, $\blacktriangle_\vee(o_1 \cdot \langle(\top, \top, \varnothing)\rangle \cdot o_2) \wedge x$ is $o_1 \cdot \langle(x, \top, \varnothing)\rangle \cdot o_2$-acceptable and hence entails $\blacktriangle_\vee(o_1 \cdot \langle(x, \top, \varnothing)\rangle \cdot o_2)$. We can now show that any formula from $L(o_1 \cdot o_2)$ entailed by that core as already entailed by $\blacktriangle_\vee(o_1 \cdot o_2)$. Q.E.D.

Fixed positions of intermediate inputs. Now assume we know the positions where intermediate inputs may have occurred. This is imaginable, for example, in scenarios where the observing agent gathers o in several sessions, but does not know if $\mathcal{A}$ receives further inputs between those sessions. How many intermediate inputs should be assumed at each of those points? We cannot allow an arbitrary number as this is computationally infeasible, so it would be helpful to have an upper bound which we could then use. We claim that it suffices to assume j intermediate inputs at a particular position in o, where j is the number of revision inputs recorded in o following

[8] That is, we put an entry $\langle(x_i, \top, \varnothing)\rangle$ with a new variable x_i between any two entries $\langle(\varphi_i, \theta_i, D_i)\rangle$ and $\langle(\varphi_{i+1}, \theta_{i+1}, D_{i+1})\rangle$ in o.

that position, i.e., ignoring possible intermediate inputs appearing later.[9] The intuition is as above. For every recorded revision input, we assume one intermediate input which collects all the relevant intermediate inputs that have really occurred.

If this claim is correct, we can introduce into o one entry $(\chi_i, \top, \varnothing)$ for every intermediate input. Thus we get a parametrised observation containing an entry for every revision input received. We can then construct a weakest acceptable core belief by instantiating each χ_i by x_i, calculating the rational explanation of the observation thus obtained and then eliminating the additional variables from the core belief. For example, given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_5, \theta_5, D_5)\rangle$ and the information that intermediate inputs have been received only after φ_2 and φ_4, we can calculate the weakest possible core starting with $o' = \langle(\varphi_1, \theta_1, D_1), (\varphi_2, \theta_2, D_2), (x_1, \top, \varnothing), (x_2, \top, \varnothing), (x_3, \top, \varnothing), (\varphi_3, \theta_3, D_3), (\varphi_4, \theta_4, D_4), (x_4, \top, \varnothing), (\varphi_5, \theta_5, D_5)\rangle$ and eliminating the x_i from $\blacktriangle_\vee(o')$. Again, all x_i are propositional variables not contained in $L(o)$.

The above claim for limiting the number of assumed intermediate inputs follows almost immediately from the following proposition.

Proposition 3.8. Let $\rho = (\varphi_1, \ldots, \varphi_n)$ and $\sigma = (\psi_1, \ldots, \psi_m)$. Then there exists a $\sigma' = (\psi'_1, \ldots, \psi'_n)$ such that for all $1 \leq i \leq n$

$$f(\sigma \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \equiv f(\sigma' \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) .$$

Proof (Sketch). The proof of this result uses the fact that for every sequence σ there is a logical chain σ' (a sequence of formulae where each formula is entailed by its successor) that behaves exactly like σ. That is $f(\sigma \cdot \rho') \equiv f(\sigma' \cdot \rho')$ for all sequences ρ'. However, for this result it suffices that σ and σ' behave equivalently for all prefixes of ρ. We then show that a suitable σ' exists, in fact using the rational explanation algorithm and hypothetical reasoning. Q.E.D.

Note that this result is not trivial, as m can be (much) greater than n and in this case we have to find a shorter sequence yielding equivalent formulae for all $1 \leq i \leq n$. This proposition tells us that we can replace one block of intermediate inputs σ by one of the proposed length and be guaranteed an equivalent formula being constructed in the calculation for each recorded revision input φ_i coming later in the observation.

We want to remark that some care has to be taken when considering the general case, where several blocks of intermediate inputs exist. Then ρ in the proposition may contain more elements than just the recorded revision

[9] The above result—that one intermediate input between any two recorded ones is enough—is not applicable here. Intermediate inputs may not be allowed at every position.

inputs; it also contains intermediate ones. And thus we have to find a sequence σ' not of length n but $j \leq n$ where j is the number of recorded inputs. We are currently investigating whether the number of intermediate inputs that have to be assumed can be reduced further without effect on the core belief calculated.

Fixed number of intermediate inputs. If we are given a maximal (or exact) number n of intermediate inputs that may have occurred we can draw conclusions about the core belief of the agent using the following method. Due to Proposition 3.7 we should indeed assume the maximal number of intermediate inputs—n. So let o be the observation containing only recorded inputs. If o has less than $n+2$ recorded inputs and there are no restrictions as to the positions of the intermediate inputs, we can use the result that one intermediate input between any two recorded ones suffices to explain o; otherwise, there are not enough intermediate inputs for this result to be applicable. In this case, we create the set of all possible observations o' where n intermediate inputs have been inserted in o:

$$O' = \{o_1 \cdot \langle (x_1, \top, \varnothing)\rangle \cdot o_2 \cdot \ldots \cdot \langle (x_n, \top, \varnothing)\rangle \cdot o_{n+1} \mid o = o_1 \cdot \ldots \cdot o_{n+1}\}.$$

Here we have already replaced the unknown formulae by new variables. If we have information about the positions of the intermediate inputs we can also take this into account when constructing O'. The observation o_j may be empty, so consecutive intermediate inputs are explicitly allowed. Now any possible core belief will entail $\bigvee\{\blacktriangle \mid \mathsf{Cn}(\blacktriangle) = \mathsf{Cn}(\blacktriangle_\vee(o')) \cap L(o), o' \in O'\}$. Note that this formula itself need not be an o'-acceptable core, i.e., it may not really explain the observation using n intermediate inputs. Conclusions about beliefs and non-beliefs can only be safe if they are safe for every observation in O'.

3.4 Summary

In this section, we showed what can still be said about $\mathcal{A}$ if some of the completeness assumptions about the observation o are weakened. We started by allowing unknown subformulae χ_i to appear in o. This can happen as the logical content of the revision inputs or the beliefs need not be completely known. In case the observation still contains a record for every input received, the calculation of an optimal core belief is still possible. The proposed method for dealing with such parametrised observations was to instantiate the unknown subformulae χ_i with new variables and apply the rational explanation construction to the observation thus obtained. From this explanation we can safely conclude which beliefs must belong to the agents core belief no matter what the real instantiation of the χ_i was.

We showed in [23] that although we can construct a core belief from $L(o)$ this does not guarantee that o can be explained without extending the

language. The unknown subformulae may still have to contain variables not belonging to $L(o)$. We claim that it is not useful to look for an optimal instantiation of the unknown subformulae. Weakest core belief and belief trace heavily depend on the choice of the instantiation of the χ_i and even if we had the correct ones, Section 2.4 showed that the conclusions drawn from the belief trace implied by our explanation are of limited use. Instead we argue that the χ_i should be instantiated with x_i and reasoning be done based on the rational explanation. This allows us to draw correct conclusions about the actual core belief of the agent, which must entail the one calculated that way. Further, we can use hypothetical reasoning to verify other beliefs and non-beliefs (restricted to $L(o)$) implied by the explanation thus obtained.

The additional assumption that the belief corresponding to the last revision input in the (parametrised) observation completely characterises $\mathcal{A}$'s beliefs at that point once more need not help. It might not even convey additional information about the language of the agent's epistemic state or of the unknown subformulae. Consider the parametrised observation $\langle(p \wedge \chi, \top, \varnothing), (\neg p, \neg p, \varnothing)\rangle$. It might not be very interesting but it illustrates the point. As $\neg p$ is inconsistent with the first input, χ could be instantiated with any formula and still $\neg p$ would completely characterise the agent's final beliefs.

We then further allowed intermediate inputs, i.e., the original observation does not contain a record for every input received. Some observations can be explained only when assuming that intermediate inputs have occurred. When fixing their number and positions, the problem is reduced to partially known inputs. If the observing agent does not have this information, we sketched procedures for drawing conclusions about what $\mathcal{A}$'s core belief must entail.

4 Conclusion, Future and Related Work

In this paper, we departed from the traditional belief revision setting of investigating what an agent should believe after receiving (a sequence of pieces of) new information in a given initial state. Instead, we place ourselves in the position of an observer trying to reason about another agent in the process of revising its beliefs. Coming up with models of other agents is useful in many application areas as informed decisions may improve the personal or group outcome of interactions.

The basic and admittedly oversimplified setting we consider is that we are given an observation containing propositional information about the revision inputs received by an agent $\mathcal{A}$ and about its beliefs and non-beliefs following each input. We investigated several degrees of incompleteness of the information provided. From such an observation we try to get a clearer

picture of $\mathcal{A}$. Assuming $\mathcal{A}$ to employ a particular belief revision framework, the general approach for reasoning about the agent is to "regress" the information contained in the observation to arrive at a possible initial state of the agent. This state completely determines the revision behaviour and therefore allows to draw conclusions about $\mathcal{A}$'s beliefs at each point in time during the observation as well as future beliefs. Even under the very strict assumptions we impose, hardly any safe conclusions can be drawn. Intuitively, this is because coming up with $\mathcal{A}$'s true initial state is virtually impossible. The observing agent can only try to extend and refine the observation and reason hypothetically in the sense of testing conjectures about $\mathcal{A}$ in order to improve the model.

It should be clear that the general question does not require the use of the belief revision framework we assumed. For future work, it might be interesting to see if similar results can be obtained when assuming $\mathcal{A}$ to employ a different framework. It would be interesting to see how different revision frameworks compare with respect to their power to explain an observation and whether there is a significant difference in the quality of the conclusions that can be drawn. Another important question is whether there is a way to actually find out which revision framework an observed agent employs or whether other assumptions can be verified. We claimed that it is not reasonable to look for the optimal instantiation of the unknown subformulae but rather do hypothetical reasoning restricted to $L(o)$. However, in some applications it might be interesting to know what the actual revision input was that triggered a certain reaction in the agent. So comparing potential instantiations (possibly from a fixed set of potential formulae) could be a topic for future research.

We want to remark that the methodology illustrated in this paper can also be applied in slightly modified settings. It is possible to construct an initial state that explains several observations in the sense that different revision sequences start in the same state. This is reasonable, e.g., when thinking about an expert reasoning about different cases (the initial state representing the expert's background knowledge) or identical copies of software agents being exposed to different situations. Our work is focused on reasoning using observations of other agents, but observing oneself can be useful as well. By keeping an observation of itself an agent may reason about what other agents can conclude about it, which is important when trying to keep certain information secret. The results can also be applied for slight variations of the assumed belief revision framework. For example, it is possible to allow the core belief to be revised or to relax the restriction that new inputs are always appended to the end of ρ in an epistemic state $[\rho, \blacktriangle]$. The interested reader is referred to [24].

Our work has contact points to many other fields in AI research. Most obvious is its relation to belief revision. The intuitive interpretation we

used for the assumed revision framework is incorporation of evidence [13]. However, the representation of the epistemic state as a sequence of formulae does not distinguish between background knowledge and evidence. When applying the results, a more detailed analysis of the intended meaning of the concepts involved and a corresponding interpretation of the results would be needed. Reasoning about other agents is central for many areas, e.g., multi-agent systems, user modelling, goal and plan recognition, etc. Here we investigated one specific aspect. In reasoning about action and change, the question is often to find an action sequence that would cause a particular evolution of the world—either to achieve some goal (planning), or to find out what happened (abduction). Often, the initial state and the effects of an action are specified. In our setting, the effect of a revision input is not quite clear. It might be accepted by the agent or not and beliefs triggered by the input heavily depend on the initial state. Trying to come up with hypotheses about the inner mechanisms of an observed system, which could be interpreted as its initial state that determines its future behaviour, is a topic treated also in induction.

We are not aware of work that investigates reasoning about the evolution of an observed agent's beliefs matching our setting. So we want to conclude by mentioning some papers that investigate similar questions. [11] considers a much richer belief revision framework in a dialogue context. However, the focus is on progressing beliefs through a sequence of speech acts starting in a given initial state of the agents. This and many other publications utilise modal logics for representing agents' beliefs, [14] being another example also handling the *dynamics* of these beliefs. Often there are proof systems or model checkers for the logics presented, but model generation, which is what we are doing in this paper, generally seems to be a problem. This means that if the initial state is not given, hypotheses can only be tested (via proofs) but not systematically generated. However, this is what calculating a potential initial state and the corresponding belief trace is.

The papers [2, 26], which are dealing with update rather than belief revision, start from a sequence of partial descriptions of an evolving world and try to identify preferred trajectories explaining this sequence. [2] intends to sharpen the information about the last state of the world and concentrates on a particular preference relation, giving a representation result. [26] compares different possible preference relations among trajectories, positioning the approach with respect to revision and update. However, both allow for arbitrary changes at any point in time, i.e., they do not allow to integrate information about which actions were performed nor reason about possible outcomes of an action. Recall that although our observation contains the revision input received, this does not mean that it is actually accepted.

Acknowledgments

We thank the editors and anonymous reviewers for very constructive comments that helped to improve an earlier version of this paper. We also want to thank (in lexicographic order) Gerhard Brewka, James P. Delgrande, Didier Dubois, Andreas Herzig, Gabriele Kern-Isberner, Sébastien Konieczny, Jérôme Lang, Wiebe van der Hoek, and Hans van Ditmarsch for helpful discussions on the topic.

References

[1] C. Alchourrón, P. Gärdenfors & D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] S. Berger, D. Lehmann & K. Schlechta. Preferred history semantics for iterated updates. *Journal of Logic and Computation*, 9(6):817–833, 1999.

[3] R. Booth. On the logic of iterated non-prioritised revision. In G. Kern-Isberner, W. Rödder & F. Kulmann, eds., *Conditionals, Information, and Inference. International Workshop, WCII 2002, Hagen, Germany, May 13–15, 2002, Revised Selected Papers*, vol. 3301 of *Lecture Notes in Artificial Intillegence*, pp. 86–107. Springer, 2005.

[4] R. Booth, T. Meyer & K.-S. Wong. A bad day surfing is better than a good day working: How to revise a total preorder. In Doherty et al. [15], pp. 230–238.

[5] R. Booth & A. Nittka. Reconstructing an agent's epistemic state from observations about its beliefs and non-beliefs. *Journal of Logic and Computation*. Forthcoming.

[6] R. Booth & A. Nittka. Beyond the rational explanation. In J. Delgrande, J. Lang, H. Rott & J.-M. Tallon, eds., *Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics*, no. 05321 in Dagstuhl Seminar Proceedings. 2005.

[7] R. Booth & A. Nittka. Reconstructing an agent's epistemic state from observations. In L.P. Kaelbling & A. Saffiotti, eds., *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30–August 5, 2005*, pp. 394–399. Professional Book Center, 2005.

[8] R. Booth & J.B. Paris. A note on the rational closure of knowledge bases with both positive and negative knowledge. *Journal of Logic, Language and Information*, 7(2):165–190, 1998.

[9] C. Boutilier. Revision sequences and nested conditionals. In R. Bajcsy, ed., *Proceedings of the 13th International Joint Conference on Artificial Intelligence. (IJCAI-93) Chambéry, France, August 28–September 3, 1993.*, pp. 519–525. Morgan Kaufmann, 1993.

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

[11] L.F. del Cerro, A. Herzig, D. Longin & O. Rifi. Belief reconstruction in cooperative dialogues. In F. Giunchiglia, ed., *Artificial Intelligence: Methodology, Systems, and Applications, 8th International Conference, AIMSA '98, Sozopol, Bulgaria, September 21–13, 1998, Proceedings*, vol. 1480 of *Lecture Notes in Computer Science*, pp. 254–266. Springer, 1998.

[12] A. Darwiche & J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89(1–2):1–29, 1997.

[13] J.P. Delgrande, D. Dubois & J. Lang. Iterated revision as prioritized merging. In Doherty et al. [15], pp. 210–220.

[14] H. van Ditmarsch, W. van der Hoek & B.P. Kooi. *Dynamic Epistemic Logic.*, vol. 337 of *Synthese Library*. Springer-Verlag, 2007.

[15] P. Doherty, J. Mylopoulos & C.A. Welty, eds. *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning, Lake District of the United Kingdom, June 2–5, 2006*. AAAI Press, 2006.

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

[17] D. Lehmann. Belief revision, revised. In C.S. Mellish, ed., *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal, Québec, Canada, August 20–25 1995*, pp. 1534–1540. Morgan Kaufmann, 1995.

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

[19] D. Makinson. Screened revision. *Theoria*, 63(1–2):14–23, 1997.

[20] A. Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41(3):353–390, 1994.

[21] A. Nayak, M. Pagnucco & P. Peppas. Dynamic belief revision operators. *Artificial Intelligence*, 146(2):193–228, 2003.

[22] B. Nebel. Base revision operations and schemes: Semantics, representation and complexity. In A.G. Cohn, ed., *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI-94), Amsterdam, The Netherlands, August 8–12, 1994*, pp. 342–345. John Wiley and Sons, 1994.

[23] A. Nittka. Reasoning about an agent based on its revision history with missing inputs. In M. Fisher, W. van der Hoek, B. Konev & A. Lisitsa, eds., *Logics in Artificial Intelligence, 10th European Conference, JELIA 2006, Liverpool, UK, September 13–15, 2006, Proceedings*, vol. 4160 of *Lecture Notes in Computer Science*, pp. 373–385. Springer, 2006.

[24] A. Nittka. *A Method for Reasoning About Other Agents' Beliefs from Observations*. Ph.D. thesis, Leipzig University, 2008.

[25] O. Papini. Iterated revision operations stemming from the history of an agent's observations. In M.-A. Williams & H. Rott, eds., *Frontiers of Belief Revision*, pp. 279–301. Kluwer Academic Press, 2001.

[26] F.D. de Saint-Cyr & J. Lang. Belief extrapolation (or how to reason about observations and unpredicted change). In D. Fensel, F. Giunchiglia, D.L. McGuinness & M.-A. Williams, eds., *Proceedings of the Eights International Conference on Principles and Knowledge Representation and Reasoning (KR-02), Toulouse, France, April 22–25, 2002*, pp. 497–508. Morgan Kaufmann, 2002.