A Method for Reasoning about other Agents' Beliefs from Observations

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

INFORMATIK

vorgelegt

von **Dipl.-Inf. Alexander Nittka**

geboren am 13.09.1977 in Rathenow, Deutschland

Die Annahme der Dissertation haben empfohlen:

1. Prof. Dr. Gerhard Brewka (Universität Leipzig)
2. Prof. Dr. Gabriele Kern-Isberner (Universität Dortmund)
3. Dr. habil. Jérôme Lang (IRIT, Frankreich)

Die Verleihung des akademischen Grades erfolgt auf Beschluss des Rates der Faukltät für Mathematik und Informatik vom 18.02.2008 mit dem Gesamtprädikat magna cum laude.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 26.03.08

CV and Relevant Publications

Curriculum Vitae

seit 06/03	wissenschaftlicher Mitarbeiter am Institut für Informatik der Universität Leipzig
09/02–12/02	3-monatiges Praktikum an der Universität von New South Wales in Sydney (Knowledge Systems Group, Artificial Intelligence Laboratory), betreut durch Professor Norman Foo
10/98–04/03	Informatik-Studium an der Universität Leipzig, Diplomarbeit: A 3-Valued Approach to Disbelief
1998	Abitur am Friedrich-Ludwig-Jahn-Gymnasium Rathenow

Publications

- R. Booth and A. Nittka. Reconstructing an agent's epistemic state from observations about its beliefs and non-beliefs. *Journal of Logic and Computation*, 2008. ([13])
- A. Nittka and R. Booth. A method for reasoning about other agents' beliefs from observations. *Texts in Logic and Games*, 2008. ([75])
- A. Nittka and R. Booth. A method for reasoning about other agents' beliefs from observations. In *Formal Models of Belief Change in Rational Agents*, number 07351 in Dagstuhl Seminar Proceedings, 2007. ([74])
- A. Nittka. Reasoning about an agent based on its revision history with missing inputs. In *Proceedings of JELIA'06*, pages 373–385, 2006. ([73])
- R. Booth and A. Nittka. Beyond the rational explanation. In *Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics*, number 05321 in Dagstuhl Seminar Proceedings, 2005. ([11])
- R. Booth and A. Nittka. Reconstructing an agent's epistemic state from observations. In *Proceedings of IJCAI'05*, pages 394–399, 2005. ([12])

Abstract

Belief revision traditionally deals — from a first person perspective — with the question of what an agent *should* believe given an initial state and a revision input. This question is approached in two main ways: (i) formulating general properties a belief revision operator should satisfy and (ii) constructing specific revision operators. Reasoning about what another agent does *in fact* believe during a sequence of revisions is equally important as agents are not alone in the world and have to interact successfully with each other. This third person perspective, which we look at in this thesis, has received much less attention so far.

In order to allow for a focused investigation, we assume the observed agent to employ a particular framework for iterated non-prioritised revision, i.e., a framework that allows for dealing with sequences of revision inputs that are not necessarily accepted by the agent. One important component of the agent's epistemic state is its core belief — a formula determining which revision inputs are accepted and which are not, a belief the agent commits to at all times.

The task is to draw conclusions about the agent based on an observation. This observation contains information about which revision inputs the agent received and what it believed and did not believe upon receiving them. We are particularly interested in conclusions concerning whether inputs are accepted or rejected, i.e., conclusions about the agent's core belief, and its unrecorded beliefs. The general method will be to construct a potential initial epistemic state of the agent and progress the inputs recorded in the observation starting in that state in order to generate hypotheses about the beliefs. We call a state an explanation if it verifies the information contained in the observation. There are generally many possible explanations. In order to select one explanation, we will present and justify a set of preference criteria. It turns out that there is a unique logically weakest core belief that can be used for explaining an observation. A second criterion expresses the preference of explanations that minimise the beliefs attributed to the agent.

We introduce the assumed belief revision framework and show that any epistemic state defines a rational consequence relation — a relation with particular closure properties. In the current setting it can be interpreted as containing information of the form "If the agent were

to receive and accept the revision input φ it would believe θ." It turns out that an observation can be translated into a partial description of such a relation and that we can make use of existing work on completing partial information about a rational consequence relation in order to construct an explanation. This is done by an iterative refinement of the assumed core belief. The explanation thus obtained, which we call the *rational explanation*, is optimal with respect to the preference criteria mentioned before. However, still not all conclusions we draw based on the rational explanation of an observation need to be correct and we present a way for distinguishing them. The idea is to modify the original observation such that the new one contradicts a conclusion. An explanation for the modified observation would consequently be a counterexample to the conclusion. We can use the rational explanation construction to find such an explanation or prove that no such explanation exists.

In the first part of the work, we assume that the observation is complete in the sense that all revision inputs received during the time of observation are indeed recorded and their logical content is completely known. These assumptions are essential to the optimality of the rational explanation. So far, this prevents us from dealing with cases where revision inputs are missed by the observer or where they are not all completely understood, which will be the case in most realistic settings. The second part of the thesis investigates what can be said in such cases. We model unknown logical content by allowing the formulae recorded in the observation to contain unknown subformulae which are instantiated by new propositional formulae. We can then use (variants of) the original rational explanation algorithm to reason about the agent. Again, we are able to prove the existence of a unique weakest core belief yielding safe conclusions as to which revision inputs must be rejected by the agent. Missing revision inputs can be dealt with by assuming the observation to contain additional entries where the revision inputs are formulae whose logical content is completely unknown. As we may not be informed about the positions or number of the additional inputs, further care needs to be taken when reasoning about the agent. We sketch algorithms for a number of cases differing in the detail of information available to the observer.

In the third part of the thesis, we look at the application of the methods in slightly different settings. Reasoning about different observations starting in the same state, which has applications in accessing expert knowledge or reasoning about software agents, is particularly interesting. We also consider variants of the belief revision framework that do not prioritise new over older revision inputs.

Computational complexity does not play a major role in this thesis but we are able to give a number of complexity results for the main decision problems we are interested in.

Acknowledgements

I would like to thank everybody who has helped me in starting, carrying out, and finalising this project. First and foremost, I want to express my deepest gratitude to Richard Booth who has accompanied me since the very beginning of my research career. I owe the topic of this thesis to him. He has been a constant source of inspiration, encouragement, constructive criticism, and detailed comments at every stage of the work. It was a great pleasure to co-author papers with him. I am grateful for his exceptional talent for pinpointing and improving buggy formulations. His efforts and support have been invaluable.

The role of my supervisor Gerhard Brewka has been no less important. Although the core of his research interests is different, he was generous with the time spent on discussions, reading of and commenting on various versions of this thesis and other papers. He has given me the freedom to pursue my interests and has backed my work in many ways. His support and guidance is gratefully acknowledged.

There are a number of people whose support was less permanent but no less valuable. I want to thank Jérôme Lang for numerous discussions, for the invitation to Toulouse and for his hospitality during that stay. I also want to thank Gabriele Kern-Isberner, Sébastien Konieczny, Didier Dubois, Wiebe van der Hoek, Hans van Ditmarsch, Andreas Herzig, James Delgrande, Samir Chopra, Thomas Meyer, Norman Foo and Maurice Pagnucco for insightful discussions. Thanks to all I fail to mention here.

This thesis has been proofread and commented on at different stages by a number of people. I deeply thank Richard Booth, Margit Brause, Gerhard Brewka as well as Gabriele Kern-Isberner and Jérôme Lang for doing this tedious work. Being the one who put them in, I am responsible for all errors that remain or have been added while making last-minute changes.

I want to thank all my family for their support. Thanks also to our neighbours Katrin and Frank Loebe for sharing their time and home appliances. I am grateful to my daughter Elke for providing a deadline for submitting this thesis.

Contents

Chapter 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 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 [18], the authors use observable actions to draw conclusions about other agents' mental attitudes. We argue that the beliefs of an agent manifest themselves not only in its actions.

They can also be observed more directly, e.g. in communication.[1] 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 three 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. This scenario is related to user modelling.
- 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 may have to cope with partial information about the revision inputs.[2] As we might have to deal with the observed agents later, forming a picture of them will be useful.
- By observing an expert reasoning in one or more cases we try to get access to the background knowledge. Assuming the expert to receive a sequence of facts concerning a particular case, beliefs resulting from the revision process might reveal knowledge the expert cannot state explicitly. If we manage to deal with such a case we also contribute to the field of knowledge acquisition.

The common element in all those scenarios is that given information about the revision process of another agent we[3] are interested in a more complete picture of the observed agent. Was a particular input definitely accepted or rejected by the agent? What did it believe at a certain point during the observation? What is its background knowledge, i.e., which general rules does the agent believe to hold? What will be the impact of a further input? Answers to these and similar questions might improve future interaction or enhance our own knowledge of the world.

The information at our disposal for reasoning about another agent will be of a particular form. We will not investigate how observations made in the real world can be transformed into that form although this is clearly an important problem. We are given a (possibly

[1] There may be further ways for beliefs to be observed, but this question is not the central one in our work.

[2] 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.

[3] In this context "we" is always equivalent to "the observing agent", as this work is intended to investigate what it can conclude about the other agent.

incomplete) sequence of (partially known) revision inputs that were received by the agent. Further we are given information on what the agent believed and did not believe after having received each input. All this (propositional) information constitutes an *observation*[4] of the agent. First, we will investigate observations which are complete with respect to the revision inputs received by the agent. That is, we assume to know exactly which inputs were received during the time of observation. The results obtained for this case will then be used for dealing with the more general one.

Our general approach to reasoning about an agent based on observations will be as follows. We assume the agent to employ a particular belief revision framework for incorporating revision inputs. It is clear that the question of drawing conclusions about another agent based on an observation is interesting independently of which revision framework that agent really uses. However, assuming a particular one allows for a more focused investigation. We will then try to find an initial state of the agent that best explains the observation. The notion of *best* in this context will be defined later. By initial state we mean the agent's epistemic state at the time the observation started. As we do not know the true initial state we will 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. Usually, there will be many explaining states. This initial state represents the background knowledge of the agent and will allow us to reason about beliefs of the agent not recorded in the observation.[5]

Many approaches for reasoning about action, belief revision, etc. assume the initial belief state to be given and deal with the case of progression through sequences of actions/revision inputs. They say little or nothing about the case where it 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 Context

Our work has contact points with a large number of research areas and of course it draws on a number of existing results. In this section, we briefly want to relate our work to these. We will start by giving an overview of the work in belief revision that is relevant to our approach. Then we will point out relevant work on conditional beliefs. Finally, we will

[4]This term as we use it will be made precise in Section 2.3.

[5]So our approach has similarities to regressing a known sequence of actions starting from some partially known states to arrive at a possible initial state and then progressing them starting in that state in order to get more information about the partially known states.

mention a number of research areas that deal with related questions. For now, we assume familiarity with notation from set theory and propositional logic, which we will nevertheless recall in Section 1.3.1.

1.2.1 belief revision

Our work is tightly connected to belief revision as the investigations started out as an attempt to take a third person perspective and to reason about an agent that revises its beliefs. Note that we are not interested in characterising or constructing yet another belief revision operator. Given a set of propositional beliefs K which is closed under logical consequence Cn and a new piece of information φ which is supposed to be incorporated into the beliefs, the question is what should happen in case K is inconsistent with φ. Which beliefs should be given up, which should persist? In particular, [1, 31] set off investigations for describing principled ways for answering those questions. As one part of what is now known as the AGM-framework for belief revision, the authors propose a set of postulates that should be satisfied by any revision operation. These postulates basically only relate the beliefs before and after the revision. We give the six basic and two supplementary postulates they suggest for a belief revision operator $\dot{+}$. $+$ denotes the expansion operator which simply adds the formula and closes the resulting set under logical consequence ($K + \varphi = Cn(K \cup \{\varphi\})$).

closure	$K \dot{+} \varphi$ is a belief set
success	$\varphi \in K \dot{+} \varphi$
inclusion	$K \dot{+} \varphi \subseteq Cn(K \cup \{\varphi\})$
preservation	If $\neg\varphi \notin K$ then $Cn(K \cup \{\varphi\}) \subseteq K \dot{+} \varphi$
consistency	$K \dot{+} \varphi$ is inconsistent if and only if φ is inconsistent
equivalence	If $\varphi \equiv \psi$ then $K \dot{+} \varphi = K \dot{+} \psi$
superexpansion	$K \dot{+} (\varphi \wedge \psi) \subseteq (K \dot{+} \varphi) + \psi$
subexpansion	If $\neg\psi \notin K \dot{+} \varphi$ then $(K \dot{+} \varphi) + \psi \subseteq K \dot{+} (\varphi \wedge \psi)$

Several formalisms have been proposed for characterising the set of all revision operations satisfying the postulates and constructing actual belief revision operators. Essential for our work is Grove's notion of a "System of Spheres" [36], which basically is a total preorder on — or ranking of — possible worlds (truth assignments). The preorder states for every pair of worlds whether they are considered to be equally plausible or which of the two is considered more plausible than the other. The preorder can also be interpreted as assigning a rank to each world, the ranks being linearly ordered. Worlds having the same rank are considered equally plausible, worlds with lower rank are considered more likely than worlds with higher ranks. The state of an agent in this setting is not given by its set of beliefs but by this ordering. The belief set K corresponds to whatever is true in all worlds with minimal

rank. The belief set resulting from the revision by φ is now determined by the ranking. It corresponds to what is true in all minimal worlds satisfying φ.[6] In case K is consistent with φ there are φ-worlds in the lowest rank. If K is inconsistent with φ, then the most plausible φ-worlds have a higher rank. The richer structure allows the result of revision to be different although the belief set is the same. In fact, it has been shown in [36] that any AGM revision function can be captured by a system of spheres, which is employed as described above, and vice versa.

Example 1.1. *Consider the propositional language generated from the variables p, q and r. We will use both $\neg p$ and $\overline{p}$ to denote the negation of p. There are eight possible worlds: $p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}, \overline{p}\,q\,r, \overline{p}\,q\,\overline{r}, \overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$. Assume $p\,q\,r$, $p\,q\,\overline{r}$, $p\,\overline{q}\,r$ and $p\,\overline{q}\,\overline{r}$ are considered equally plausible and more plausible than $\overline{p}\,q\,r$ and $\overline{p}\,q\,\overline{r}$ which in turn are more plausible than $\overline{p}\,\overline{q}\,r$ and $\overline{p}\,\overline{q}\,\overline{r}$. This is illustrated in the leftmost table below. The logical consequences of p are precisely the formulae true in all minimal worlds, so these make up the belief set. When revising by $p \leftrightarrow q$, we look for the minimal worlds consistent with this formula[7] and there are two such worlds in the lowest rank. That is, in this case the beliefs would be $Cn(\{p \wedge q\})$ after revision. When revising by $\neg p \wedge r$, there are no worlds satisfying that formula in the lowest rank. There is only one minimal world satisfying $\neg p \wedge r$: $\overline{p}\,q\,r$. So the beliefs after revision by $\neg p \wedge r$ would be the consequences of $\neg p \wedge q \wedge r$.*

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

$K = Cn(\{p\})$

$\boldsymbol{\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}}$
$\overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$\boldsymbol{p\,q\,r, p\,q\,\overline{r}}$, $p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

$K \dot{+} (p \leftrightarrow q)$

$\boldsymbol{\overline{p}\,\overline{q}\,r}$, $\overline{p}\,\overline{q}\,\overline{r}$
$\boldsymbol{\overline{p}\,q\,r}$, $\overline{p}\,q\,\overline{r}$
$p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

$K \dot{+} (\neg p \wedge r)$

If the initial ranking looked slightly different, then revision by $\neg p \wedge r$ would have a different outcome. In this case q would not be believed.

$\overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,\overline{q}\,r, \overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

$K = Cn(\{p\})$

$\overline{p}\,\overline{q}\,\overline{r}$
$\boldsymbol{\overline{p}\,\overline{q}\,r, \overline{p}\,q\,r}$, $\overline{p}\,q\,\overline{r}$
$p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

$K \dot{+} (\neg p \wedge r)$

As the lowest rank determines the beliefs, the AGM postulates tell us what the lowest rank should look like after the revision. They leave open what happens to the remaining ranks. However, this is an essential point if the revision process is to be *iterated*. Following the principle of minimal change, [15] suggests to leave the ranking completely unchanged, except for moving the minimal φ-worlds into a new lowest rank when revising by φ. [21] shows

[6] Considering a finitely generated language, as we do, they are guaranteed to exist.

[7] In the tables we have emphasised all worlds consistent with the revision input.

that this leads to counterintuitive results and proposes a set of further postulates describing the desired effect of a revision step on the entire ranking. A refinement ([9, 45]) of these postulates claims that a revision operator should modify the ranking as follows. When revising by φ, the relative order among φ-worlds and among $\neg\varphi$-worlds should persist but φ-worlds should be preferred to $\neg\varphi$-worlds that had the same rank before the revision.

In [67], Nayak proposes to split the entire ranking into one for the φ-worlds and one for the $\neg\varphi$-worlds and move the φ-ranking below the $\neg\varphi$-one. As a result every φ-world is now preferred over any $\neg\varphi$-world. [77] suggests to move the φ-worlds below $\neg\varphi$-worlds in each rank separately. [10] presents a framework subsuming these two as special cases.

Example 1.2. *The following tables are to illustrate the different approaches. We use the ordering used in Example 1.1 and revise it by $q \wedge r$. Note that the lowest rank after revision is the same in all approaches. But further revision steps may lead to different beliefs.*

"natural revision" [15]: select all minimal $q \wedge r$-worlds and place them into a new lowest rank.

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$\boldsymbol{p\,q\,r}$, $p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$
$\boldsymbol{p\,q\,r}$

"lexicographic revision" [67]: split the entire ranking into one for the $q \wedge r$-worlds and one for the $\neg(q \wedge r)$-worlds, move the former ranking below the latter one.

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

	$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r$	$\overline{p}\,q\,\overline{r}$
$p\,q\,r$	$p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

	$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
	$\overline{p}\,q\,\overline{r}$
	$p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r$	
$p\,q\,r$	

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,\overline{r}$
$p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r$
$p\,q\,r$

[77]: split each rank into $q \wedge r$-worlds and $\neg(q \wedge r)$-worlds, then move the former ones below the latter ones in each rank separately.

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r, \overline{p}\,q\,\overline{r}$
$p\,q\,r, p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

	$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,r$	$\overline{p}\,q\,\overline{r}$
$p\,q\,r$	$p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$

	$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
	$\overline{p}\,q\,\overline{r}$
$\overline{p}\,q\,r$	
	$p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$
$p\,q\,r$	

$\overline{p}\,\overline{q}\,r, \overline{p}\,\overline{q}\,\overline{r}$
$\overline{p}\,q\,\overline{r}$
$\overline{p}\,q\,r$
$p\,q\,\overline{r}, p\,\overline{q}\,r, p\,\overline{q}\,\overline{r}$
$p\,q\,r$

Obviously, when observing an agent over time, we will need to use a framework supporting iterated revision. The AGM postulates have not only been criticised for being restricted to one revision step. For example, the success postulate, which states that the revision input must be among the resulting beliefs, has found opposition as it is not clear why the new information should always be accepted. Indeed, there are many cases where an input should be rejected. This has led to investigations concerning so-called *non-prioritised* revision, i.e., revision in which the input does not automatically have *priority* over the old beliefs (see, e.g., [39, 40, 60, 64]). In defence of the postulate, it can be said that it should hold once the decision has been made that the input is to be incorporated into the beliefs. However, we cannot expect to have this information for the observed agent. That is, we will need a revision framework that supports not only iterated but also non-prioritised revision. We will describe the framework we assume the observed agent to employ in detail in Section 2.2. It will turn out to be a non-prioritised version of Nayak's lexicographic revision.

1.2.2 conditionals

Interrelations between conditional beliefs, belief revision, default reasoning and other forms of non-monotonic inference have been studied extensively (see, e.g., [7, 47, 53, 55, 61]). By $\varphi \Rightarrow \theta$ we denote a conditional assertion. Its meaning is that if φ holds then normally θ holds as well, and a set of such assertions can be interpreted as an inference (or consequence) relation. The question is which properties such a relation should satisfy in order to be reasonable. As one of several proposals, the authors of [54] give a list of properties to describe a class of inference relations they call *rational*.[8]

reflexivity	$\theta \Rightarrow \theta$
left logical equivalence	$\theta \Rightarrow \phi$ and $\theta \equiv \psi$ implies $\psi \Rightarrow \phi$
right weakening	$\theta \Rightarrow \phi$ and $\phi \models \psi$ implies $\theta \Rightarrow \psi$
and	$\theta \Rightarrow \phi$ and $\theta \Rightarrow \psi$ implies $\theta \Rightarrow \phi \wedge \psi$
or	$\theta \Rightarrow \phi$ and $\psi \Rightarrow \phi$ implies $\theta \vee \psi \Rightarrow \phi$
cautious monotonicity	$\theta \Rightarrow \phi$ and $\theta \Rightarrow \psi$ implies $\theta \wedge \phi \Rightarrow \psi$
rational monotonicity	$\theta \Rightarrow \phi$ and $\theta \not\Rightarrow \neg\psi$ implies $\theta \wedge \psi \Rightarrow \phi$

The first result that is particularly important for our work is that total preorders on worlds exactly characterise rational consequence relations [54].[9] Now, an important question is

[8] Our notation for conditionals neglects the distinction between syntax and semantic. Each line is to be read as a universally quantified closure property.

[9] $\theta \Rightarrow \phi$ holds with respect to a given total preorder on worlds if and only if all minimal θ-worlds are also ϕ-worlds. Every total preorder gives rise to a rational consequence relation and for every rational consequence relation there is a corresponding total preorder on worlds.

which assertions should follow from a set of assertions; and the authors of [54] answer this question by constructing a preferred rational consequence relation which extends that set. In other words, given a partial description of the total preorder on worlds, their rational closure construction completes that information.[10] We will not describe the construction here but give a detailed account of its extension to negative conditionals [14] in Section 2.5. This extension also allows for incorporating information of the form $\varphi \not\Rightarrow \theta$ (it is not the case that if φ then normally θ).

We make use of the correspondence of belief revision and conditionals via total preorders on worlds as follows. We will show that the belief revision framework which we assume the observed agent to employ allows to translate the observation about the agent's revision history into a partial description of a rational consequence relation. This relation will then be completed using the results from [14] and reinterpreted as the agent's initial state. Given this state we can then draw further conclusions about the agent.

1.2.3 further areas

As mentioned in the motivation, one of the main issues of AI research is reasoning about dynamic domains. This involves the question of how to represent the state of the world and the effects of actions on that world and how to reason about them. For doing so, a number of formalisms have been proposed, e.g. situation calculus [63, 85], fluent calculus [90, 91], event calculus [52], and the action language $\mathcal{A}$ [33]. However, it is obvious that it is also necessary to reason about the agents involved. The actions taken by an agent will generally depend on its goals and its beliefs about the world. So all the formalisms we mentioned were extended to allow for expressing beliefs and belief change of agents (see, e.g., [41, 43, 44, 56, 62, 88]), but specific methods for reasoning about the revision history of an observed agent are not discussed. In [86], Sandewall also identifies *chronicle completion*, i.e., completion of partial information about an evolving world, as one of the important reasoning tasks for dynamical systems. This is what we attempt in this thesis (interpreting the beliefs of an agent as the dynamic system).

Reasoning about other agents is not new. For example, much work has been done on interpreting observed actions of agents (as opposed to their beliefs) in order to identify the plans they follow. Possible applications of plan ascription [2, 51] and plan recognition [48, 49, 87] include inferring the goals of an agent or predicting its next action.

[10] The underlying idea is as follows. The total preorder on worlds represents their respective plausibility and the rational closure construction attempts to make every world as plausible as the given set of conditional assertions allows.

We stated before that our general method will be to identify an initial state of the agent that best explains the observation. One area whose main aim is to find explanations is abduction [80]. Generally, there is a background theory and an observation that cannot be accounted for by the theory alone. So the task is to identify possible explanations (see, e.g., [78] for an overview of abductive methods). Normally, a set of abducibles to select from is given. Clearly, the mere fact of being an explanation is not enough — the interest is in best or at least good explanations (see [81] for an account on properties of preference relations). There are also approaches to abduction that employ belief revision [16, 95]. Causal reasoning goes a step further. Here, the aim is to identify the *actual* causes for an observation [37, 38]. Our work differs from these approaches for identifying explanations. The observations we start off with are different, a background theory is not given — only general properties satisfied by it — and the explanation we come up with is not a fact (literal) but a more complex structure.

1.3 Preliminaries

1.3.1 notation

The observed agent this thesis is focused on will be denoted by $\mathcal{A}$. We will restrict our attention completely to propositional logic. So, L will usually 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. $\alpha, \beta, \theta, \lambda, \mu, \varphi, \phi, \psi$ (often with subscript) will denote propositional formulae, i.e., particular elements of L. Later, χ will be used as *placeholder* for an unknown formula.

σ and ρ are used to denote sequences of formulae, () being the empty sequence. The function $\cdot$ in $\sigma \cdot \rho$ and $\sigma \cdot \varphi$ denotes sequence concatenation and appending a formula to a sequence, respectively. In fact, $\cdot$ will be used to denote concatenation of any type of sequences and corresponding elements. $\vdash$ 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. $Cn(S)$ denotes the set of all logical consequences of a set of formulae S, i.e., $Cn(S) = \{\beta \mid S \vdash \beta\}$. Again $Cn(\alpha)$ abbreviates $Cn(\{\alpha\})$ for singleton sets. $\equiv$ is the relation of logical equivalence between formulae.

In some of the proofs we will use the relation $\models$ between a truth assignment m and a formula φ evaluated to true by m. Given a subset P of the propositional variables in our language, $m \sim_P m'$ denotes that two truth assignments m and m' agree on all variables except those contained in P, which may be evaluated differently by m and m'. If P is a singleton set we

will write $\sim_p$ instead of $\sim_{\{p\}}$. All revision operations $*$ introduced will be left associative and consequently $K * \varphi_1 * \varphi_2$ is intended to mean $(K * \varphi_1) * \varphi_2$. As is common, we will sometimes abbreviate "if and only if" by "iff".

Computational complexity will only play a marginal role in this thesis. Here we will only name the most important concepts we use and refer the reader to [32, 76, 94] for more detail. P is the class of decision problems that can be decided by a deterministic Turing machine (TM) in polynomial time (in the size of the input) and NP the class of problems that can be decided by a non-deterministic TM in polynomial time. coX is the class of problems whose complements are in X. $EXPTIME$ contains those problems that can be decided in exponential time by a deterministic TM. The polynomial hierarchy allows for a more fine-grained distinction between (a number of) problems in this class. It uses the idea of oracles. Intuitively, the decision procedure can solve problems of a given class instantaneously. A problem is in P^{NP} if it can be decided in polynomial time querying an NP-oracle, e.g., by a polynomial number of satisfiability tests. The polynomial hierarchy is defined as follows. $\Delta_0^P = P$, $\Sigma_0^P = P$ and $\Pi_0^P = P$. $\Delta_{n+1}^P = P^{\Sigma_n^P}$, $\Sigma_{n+1}^P = NP^{\Sigma_n^P}$ and $\Pi_{n+1}^P = coNP^{\Sigma_n^P}$. A problem M is X-hard for a complexity class X if every problem M' in X can be reduced to M in polynomial time ($M' \leq_p M$). $M' \leq_p M$ iff there is a total function g — computable in polynomial time — such that $x \in M' \leftrightarrow g(x) \in M$. That is, via g a decision procedure for M can be applied for deciding M'. M is X-complete if it is in X and X-hard.

1.3.2 structure of this thesis

In the remainder of this chapter, we will present the simplifying assumptions we make for our investigation. Chapter 2 will introduce the assumed agent model as well as the formal definition of an observation. It further contains 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. We indicate what we can conclude about the observed agent using this explanation and propose a method for verifying those conclusions. These results were in large parts obtained in joint work with Richard Booth and published in [11, 12, 13]. Chapter 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 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. Parts of these results were published in [73, 75]. Chapter 4 illustrates the applicability of the results for slight modifications of the setting in which the reasoning is done, as well as for variants of $\mathcal{A}$'s assumed belief revision framework. Chapter 5 briefly discusses some pieces of related work before we conclude and indicate possibilities

of future continuations of this work. An appendix contains all proofs, a section on aspects of computational complexity and a compilation of the algorithms (in pseudo-code), which are only verbally described in the main text.

1.3.3 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 for a focused analysis of the problem we approach.

We restrict ourselves to propositional logic, and all components of an observation o, which is the starting point of our investigation, are already provided in a propositional logic generated from a finite set of variables. That is, we assume that revision inputs, beliefs and non-beliefs are directly observed as propositional formulae. We disregard the question of how sensory data representing the actual physical input received by the observed agent is transformed into the needed format. Agents are assumed to be sincere, i.e., they are not deceptive about their beliefs, but the information may be partial. In other words, we assume that the function mapping the actual physical observation, recorded communication, etc. into an observation o ensures that these requirements are met.

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. [30] argues that it suffices that the description of the world is static, i.e., that the evaluation of the propositions describing the world does not change. This allows for handling dynamic worlds in a belief revision setting using, e.g., time-stamped variables. As the focus of our work is a different one, we will not distinguish between these notions.

It is essential for our investigation that the revision inputs were received over time. [24] argues that in case of a static world the order of the revision inputs should not matter (in contrast to what is implied by the success postulate), but rather all input should be merged. This merging can be done according to different priority criteria, and the recency of an input is just one possible one. A central point of our work is to exploit having intermediate steps at our disposal — we explicitly need and use the information about when an input was received. The observed agent $\mathcal{A}$ 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. $\mathcal{A}$ will be assumed to employ a particular belief revision framework which will be described in detail in Section 2.2. It equates recency with reliability of information — concepts like (preferences for) sources, competence, context, etc. are completely disregarded. This is to simplify the investigation, not because we think it to be the only possible and

correct revision framework. In Chapter 4, we will show that some of the results for this very restricted framework can be used in more flexible ones.

The formal results will be given for one observed agent. We will briefly sketch the multi-agent case in Section 4.1. Basically, it can be handled by constructing one observation for each agent. These observations may or may not have been constructed independently of one another. One example of them being related is dialogue settings — beliefs expressed by one agent are the revision inputs for the others.

We consider the observations to be short term in the sense that learning, modification of the background knowledge, change of revision strategy etc. do not take place. The only thing that happens during the time of observation is that the agent incorporates the revision inputs. With respect to the third scenario (reasoning about an expert's background knowledge), if we observe the expert reasoning about several cases, the expert starts in the same initial state in each case. That is, the different cases are treated independently and we assume that the expert does not learn from case to case.

We do not investigate *strategies* for extracting as much information 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 providing the expert with interesting cases or putting agents in a setting where observations yield the most precise results is another interesting topic which we do not pursue.

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. Further, even in this very restricted setting we will able to draw interesting conclusions. Also, we will show (in particular in Section 2.8) that even if these assumptions are correct, there are very strict limitations to what we can *safely* conclude about the observed agent.

Chapter 2

The Rational Explanation

2.1 The function f

In this section, we will introduce the function f. It will be used for the definition of the belief revision framework we assume the observed agent to employ. The properties of f play a central role for the results presented in this thesis. They will be used in many of the proofs which is why we collect them in a separate section.

The argument of f is a non-empty sequence of formulae and it returns a single formula. For readability we omit the outer parentheses of the sequence and write $f(\beta_k, \ldots, \beta_1)$ instead of $f((\beta_k, \ldots, \beta_1))$. The reversed order of subscripts in the definition just indicates that formulae later in the sequence take priority over those before. The order of formulae in a sequence σ is not changed when passed to $f(\sigma)$ as an argument.

Definition 2.1.

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

First note that $f(\sigma)$ returns the conjunction of a subset of the formulae contained in $\sigma = (\beta_k, \ldots, \beta_1)$. A first important property is that the resulting formula is inconsistent exactly in those cases where the last element β_1 of its argument sequence was inconsistent. In all other cases the result is consistent — even if $\{\beta_1, \ldots, \beta_k\}$ is an inconsistent set of formulae. It is easy to see that $Cn(f(\sigma)) = Cn(\{\beta_1, \ldots, \beta_k\})$ in case the latter set is consistent. f can thus be seen as a particular form of reasoning from an inconsistent knowledge base (see, e.g., [4, 5, 27, 84]).

Proposition 2.2. *$f(\beta_k, \ldots, \beta_1) \equiv \bot$ if and only if $\beta_1 \equiv \bot$.*

The next proposition illustrates how $f(\sigma)$ operates on the sequence σ. It takes the last element (even if it is inconsistent) of the sequence and then goes backwards through σ, adding a formula as a new conjunct if this can be done consistently. If a formula is inconsistent with what has been collected so far then it is simply left out.

Proposition 2.3. *For all formulae β, β_i and all sequences σ and σ':*
(i) $f(\beta \cdot \sigma) = f(\beta, f(\sigma))$, implying
(ii) $f(\beta_k, \ldots, \beta_1) = f(\beta_k, f(\beta_{k-1}, f(\ldots, f(\beta_1)\ldots)))$ and
(iii) $f(\sigma \cdot \sigma') = f(\sigma \cdot f(\sigma'))$.

This procedure basically corresponds to linear base revision [70], which we denote here by $*_L$.[1] $(\alpha_1, \ldots, \alpha_m) *_L \alpha$ is equivalent to $f(\alpha_1, \ldots, \alpha_m, \alpha)$. However, the mode of the calculation has a subtle difference. Whereas $f(\alpha_1, \ldots, \alpha_m, \alpha)$ starts off with α and adds from $(\alpha_1, \ldots, \alpha_m)$ whatever can be added consistently, $*_L$ starts with α_m proceeding as follows. A formula is added in case its addition does not make $\neg\alpha$ inferable. Finally, α is added to the set thus constructed. But this means that the formula revised by is always believed, which corresponds to prioritised revision. We will later use f in a way that allows us to model non-prioritised revision — simply by not placing α in the last position of the sequence.

Proposition 2.4 expresses that f is syntax-independent. We can substitute any formula in a sequence by a logically equivalent one and f will return an equivalent formula for that sequence. This property will often be used implicitly without reference to the proposition.

Proposition 2.4. *If $\beta \equiv \beta'$ then $f(\alpha, \beta) \equiv f(\alpha, \beta')$ and $f(\beta, \alpha) \equiv f(\beta', \alpha)$.*

Example 2.5. *We want to calculate $f(r, p, p \rightarrow q, \neg q)$. By Proposition 2.3 this is equivalent to $f(r, f(p, f(p \rightarrow q, f(\neg q))))$. $f(\neg q) = \neg q$ corresponding to the first case in definition 2.1. $f(p \rightarrow q, f(\neg q)) = (p \rightarrow q) \wedge \neg q \equiv \neg p \wedge \neg q$ as this is consistent, which requires the second case of that definition to be applied. So, up to now $f(r, f(p, f(p \rightarrow q, f(\neg q))))$ reduces to $f(r, f(p, \neg p \wedge \neg q))$. However, p is inconsistent with $\neg p \wedge \neg q$, so it is left out (case 3) and we finally arrive at $f(r, \neg p \wedge \neg q) = r \wedge \neg p \wedge \neg q$.*

Proposition 2.6. *$f(\beta_k, \ldots, \beta_1) \vdash \beta_i$ or $f(\beta_k, \ldots, \beta_1) \vdash \neg\beta_i$ for all $1 \leq i \leq k$.*

Proposition 2.6 tells us that for any formula β_i appearing in the sequence, β_i itself or its negation $\neg\beta_i$ is entailed by the resulting formula. So $f(\sigma)$ cannot be *agnostic* about any element of σ. Proposition 2.7 expresses that extending a sequence by a formula β that is already entailed has no immediate impact.

[1] It is thus the special case of preferred subtheories [19] where all ranked sets of formulae are singletons.

Proposition 2.7. *For consistent α: If $\alpha \vdash \beta$ then $f(\alpha, \beta) \equiv f(\alpha)$.*

For tautologies and contradictions there is a much more general result. They can be inserted (or deleted) *anywhere* in the proper prefix of an argument sequence without any impact. Even if the argument sequence is extended at the end in an arbitrary way, the formula constructed will be equivalent. The restriction that the tautology generally may not be inserted or deleted at the very end is due to the treatment of contradictions in the last position. Appending a tautology will turn the resulting formula consistent, appending a contradiction will turn it inconsistent.

Proposition 2.8. *For all sequences of formulae σ, ρ_1 and ρ_2 (ρ_2 being non-empty) we have $f(\rho_1 \cdot \top \cdot \rho_2 \cdot \sigma) \equiv f(\rho_1 \cdot \rho_2 \cdot \sigma)$ and $f(\rho_1 \cdot \bot \cdot \rho_2 \cdot \sigma) \equiv f(\rho_1 \cdot \rho_2 \cdot \sigma)$.*

A similar result exists also for the case that some formula β is inconsistent with the last element of the sequence. The presence or absence of β in a proper prefix of that sequence has no impact on the formula constructed.

Proposition 2.9. *If $\alpha \vdash \neg\beta$ then $f(\rho \cdot \beta \cdot \sigma \cdot \alpha) \equiv f(\rho \cdot \sigma \cdot \alpha)$ for all sequences ρ, σ.*

Proposition 2.10 tells us that if the formulae collected from a sequence do not contradict a formula α then α can be inserted anywhere in the sequence. The results are all logically equivalent. It is not possible to conclude that in these cases the added formula has no impact whatsoever — when considering a *further* formula added to the sequence, presence and position of α do indeed matter!

Proposition 2.10. *If $f(\sigma_1 \cdot \sigma_2) \nvdash \neg\alpha$ then $f(\sigma_1 \cdot \alpha \cdot \sigma_2) \equiv f(\sigma_1 \cdot \sigma_2) \wedge \alpha$.*

Example 2.11. *As $f(p \wedge q) \vdash q$, $f(p \wedge q, q) \equiv f(p \wedge q) = p \wedge q$, but $f(p \wedge q, q, \neg p) = q \wedge \neg p$ and $f(p \wedge q, \neg p) = \neg p$ are not equivalent.*

$f(p, q) \nvdash \neg r$, so $f(r, p, q) \equiv f(p, r, q) \equiv f(p, q, r) \equiv p \wedge q \wedge r$, but $f(p, q, \neg r, r)$, $f(p, q, r, \neg r)$ and $f(p, q, \neg r) \wedge r$ are all logically different.

The first part of this example shows that appending a formula already entailed is not completely without effect as might be carelessly concluded from Proposition 2.7. The second part illustrates that similar care is necessary when applying Proposition 2.10.

The next proposition applies Proposition 2.10, showing that the formula not contradicting $f(\sigma)$ cannot only be inserted into the sequence but also be conjoined with its last element. The limitation illustrated in the above example still applies, of course. For further modifications of the sequence, the equivalence need not carry over.

Proposition 2.12. *If* $f(\sigma \cdot \alpha) \nvdash \neg\beta$ *then* $f(\sigma \cdot \alpha \wedge \beta) \equiv f(\sigma \cdot \alpha) \wedge \beta$

Proposition 2.13 shows the impact of transforming the first element of a sequence into an implication for the case that the antecedent or its negation is entailed by what is collected from the remainder of the sequence. In the first case the transformation has no impact, i.e., an equivalent formula is returned. In the second, the modified formula could simply have been omitted.

Proposition 2.13. *(i) If* $f(\sigma) \vdash \alpha$ *then* $f(\alpha \rightarrow \beta \cdot \sigma) \equiv f(\beta \cdot \sigma)$
(ii) If $f(\sigma) \vdash \neg\alpha$ *then* $f(\alpha \rightarrow \beta \cdot \sigma) \equiv f(\sigma)$

Example 2.14. *(i)* $f(r, p) = r \wedge p \equiv (p \rightarrow r) \wedge p = f(p \rightarrow r, p)$.

(ii) $f(\neg p \rightarrow r, p) = (\neg p \rightarrow r) \wedge p \equiv p = f(p)$

The last property of f which we want to present relates a sequence of formulae to any of its prefixes, given that a common formula is appended to both of them.

Proposition 2.15. *Either* $f(\sigma \cdot \rho \cdot \alpha) \vdash \neg f(\sigma \cdot \alpha)$ *or* $f(\sigma \cdot \rho \cdot \alpha) \vdash f(\sigma \cdot \alpha)$

2.2 $\mathcal{A}$'s assumed Belief Revision Framework

We already mentioned that we will assume the observed agent $\mathcal{A}$ to employ a particular belief revision framework. As we deal with sequences of revisions it will obviously need to allow for iterated revision. Further, we should not assume that all inputs received will be believed by the agent. Some inputs may contradict strong beliefs or knowledge of the agent and will hence be rejected. We cannot require an agent to accept a revision input like "Manchester is the home of the Beatles". Also, we cannot expect to be provided with the information whether a particular input is accepted or rejected by $\mathcal{A}$. Consequently, the framework also needs to allow for non-prioritised revision, i.e., revision not satisfying the success postulate.

In Section 1.2, we recalled total preorders on worlds as a possible representation of an agent's epistemic state and some corresponding revision operators. In this work we will assume $\mathcal{A}$ to employ a revision framework that is a special instance of a family of frameworks — whose underlying structure is also total preorders on worlds — that was studied in [8]. However, we will not reason about the preorders directly but work with a syntactic representation suggested there. We will establish the link between the two at the end of this section.

An agent's epistemic state $[\rho, \blacktriangle]$ is made up of two components: (i) a sequence ρ of formulae and (ii) a single formula $\blacktriangle$. $\blacktriangle$ stands for the agent's set of core beliefs — the beliefs of the agent it considers "untouchable". We will see that one main effect of the core belief is

that revision inputs contradicting it will not be accepted into the belief set of the agent and hence that non-prioritised revision is possible. ρ represents the sequence of revision inputs the agent has received thus far, so it can be interpreted as the record of the agent's revision history.[2] Revision by a formula is carried out by simply appending it to ρ. The agent's full set of beliefs $Bel([\rho, \blacktriangle])$ in the state $[\rho, \blacktriangle]$ is then determined by a particular calculation on ρ and $\blacktriangle$ using the function f which we introduced in the previous section.

Definition 2.16. *Given an epistemic state* $[\rho, \blacktriangle]$ *and a formula* φ*, the revision operator* $*$ *is defined by*

$$[\rho, \blacktriangle] * \varphi = [\rho \cdot \varphi, \blacktriangle]$$

This way, iterated revision is handled quite naturally by the framework. All revision steps are simply recorded and the problem of what the agent 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.

Definition 2.17. *The* set of beliefs $Bel([\rho, \blacktriangle])$ *in the epistemic state* $[\rho, \blacktriangle]$ *is* $Bel([\rho, \blacktriangle]) = Cn(f(\rho \cdot \blacktriangle))$.

So in order to calculate its set of beliefs, 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. The belief set of the agent then is the set of logical consequences of the formula thus constructed. Note that we do not prohibit the core belief $\blacktriangle$ to be inconsistent in which case the belief set of the agent is inconsistent. From the definition and Proposition 2.2 it follows immediately that $Bel([\rho, \blacktriangle])$ is inconsistent if and only if $\blacktriangle$ is inconsistent.

While f basically corresponds to linear base revision $*_L$ [70], $*$ is slightly different. Our belief revision operation places the input φ in the last but one and not in the last position. Thus, $[\rho, \blacktriangle]*\varphi$ corresponds to $\rho \cdot \varphi *_L \blacktriangle$ or equivalently to $\rho *_L f(\varphi, \blacktriangle)$. The core belief is always more important than the latest revision input φ. Consequently, in terms of linear base revision, the core belief can be interpreted as an input received after every regular revision input. Note also that the latest revision input, provided it is consistent with the core belief, is always considered to be more important than any revision input received before. This means that the belief revision framework we assume $\mathcal{A}$ to employ equates recency with importance. The following example illustrates the revision process and the calculation of the agent's beliefs.

[2] We believe the interpretation of ρ as revision history to be not as important. Essential for our work is the fact that the epistemic state is another way of looking at a total preorder on worlds which in turn is one representation for conditionals. In the first part of this thesis, we will propose a way to construct a state $\mathcal{A}$ *may* have had immediately before the observation started. This state will obviously not contain the agent's true revision history, but one whose future behaviour is as close to the true one as possible.

Example 2.18. *Consider the epistemic state* $[(), \neg p]$ *of an agent. Its beliefs in this state are* $Cn(f(\neg p)) = Cn(\neg p)$*. If* q *is received as a new input then the resulting epistemic state is* $[(), \neg p] * q = [(q), \neg p]$*. The corresponding beliefs are* $Cn(f(q, \neg p)) = Cn(q \wedge \neg p)$*.*

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

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

Upon receiving a revision input φ, it is believed by the agent if and only if it is consistent with the core belief $\blacktriangle$ (provided $\blacktriangle$ is consistent). This is the direct effect of the core belief and the reason why the assumed belief framework is one for non-prioritised revision. But the example also illustrates an indirect effect of $\blacktriangle$. After receiving q, the agent believed q, but after then receiving $q \rightarrow p$ it believed $\neg q$ — in the light of the core belief $\neg p$, the belief in q cannot be maintained. So, although revision inputs may not contradict each other directly, they may in presence of the core belief. In other words, the core belief not only blocks some inputs from being introduced into the belief set, but it also accounts for interaction between different inputs.

Definition 2.19. *Given a sequence of revision inputs* $(\varphi_1, \ldots, \varphi_n)$ *the* belief trace $(Bel_0^{[\sigma,\blacktriangle]}, Bel_1^{[\sigma,\blacktriangle]}, \ldots, Bel_n^{[\sigma,\blacktriangle]})$ *of an epistemic state* $[\sigma, \blacktriangle]$ *is the sequence of formulae* $Bel_0^{[\sigma,\blacktriangle]} = f(\sigma \cdot \blacktriangle)$ *and* $Bel_i^{[\sigma,\blacktriangle]} = f(\sigma \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))$*,* $1 \leq i \leq n$*.*

This definition will allow us to talk about the evolution of the agent's beliefs. The belief trace characterises the beliefs of an agent at every point of the revision process starting in its initial state $[\sigma, \blacktriangle]$. We will abbreviate $Bel_i^{[\sigma,\blacktriangle]}$ by Bel_i^{σ}, provided $\blacktriangle$ is fixed. Also, we will refer to the belief trace of the "agent" rather than the "epistemic state". Note that Bel_i^{σ} is not the *set of beliefs* of the agent after the i^{th} revision step but a *formula* uniquely determining those beliefs, i.e., $Bel([\sigma, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = Cn(Bel_i^{\sigma})$. The agent's belief trace in example 2.18 is $(\neg p, q \wedge p, \neg q \wedge \neg p, \neg q \wedge \neg p)$.

An interesting property — and possibly one of the drawbacks — of the assumed framework is the following one. It expresses that the agent will always be clear about the status of a revision input it once received. The input itself or its negation will be believed at any point in the future. $\mathcal{A}$ will never forget about an input. It may, however, forget about some logical consequences of inputs.

Proposition 2.20. $Bel([(\varphi_1, \ldots, \varphi_n), \blacktriangle]) \vdash \varphi_i$ *or* $Bel([(\varphi_1, \ldots, \varphi_n), \blacktriangle]) \vdash \neg\varphi_i$ *for all* $1 \leq i \leq n$*.*

Example 2.21. *Consider the epistemic state* $[(p \wedge q), \top]$*. Clearly,* $\mathcal{A}$*'s beliefs are* $Cn(p \wedge q)$ *in this state. After revising by* $\neg p$*, the state is* $[(p \wedge q, \neg p), \top]$ *and the agent's beliefs are* $Cn(\neg p)$*, so it ceased to believe* q*. However, the negation* $\neg p \vee \neg q$ *of the first input* $p \wedge q$ *is in the belief set.*

Proposition 2.22. *If* $Bel([\rho, \blacktriangle]) \vdash \varphi$ *then* $Bel([\rho, \blacktriangle] * \varphi) = Bel([\rho, \blacktriangle])$*.*

This proposition tells us that revising by a formula the agent already believes has no immediate impact on the *belief set*. This does not mean that the input has *no* effect. As it is included into the agent's epistemic state, it may make implicit beliefs explicit and thereby strengthen them against elimination from the belief set.

Example 2.23. *Given the epistemic state* $[(p \wedge q), \top]$ *of an agent, its belief set is* $Cn(p \wedge q)$*. Proposition 2.22 tells us that after revising by* p *the belief set remains unchanged. The belief state after a further revision step with input* $\neg q$ *is* $[p \wedge q, p, \neg q, \top]$ *and the corresponding belief set is* $Cn(p \wedge \neg q)$*. However, if the agent had not received* p*, the final state would be* $[p \wedge q, \neg q, \top]$*, the corresponding belief set being* $Cn(\neg q)$*, so in this case* p *is not believed.*

Proposition 2.10 which is used in the proof allows us to go even a step further. With respect to the agent's beliefs immediately after revising by a formula φ, it plays no role at all where φ is inserted in the sequence of its received formulae — provided that φ is consistent with its current beliefs. However, for further revision steps it plays an essential role which priority (which is reflected by the position within the sequence) each formula has.

In case the revision input is a tautology, Proposition 2.8 tells us that not only the beliefs after receiving it are the same but also that the beliefs after any sequence of further inputs will be the same. That is, a tautology has no impact. Applicability of that proposition is guaranteed, as the last element of the sequence passed to f for calculating the belief set is the core belief, so the revision input $\top$ is placed in a proper prefix of that sequence.

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. Note that this can be interpreted — in the spirit of [24] — as merging the revision inputs received over time into a single one and then revising by the resulting formula. In our framework the merging and revising is done using the same method. Note that this property talks only about the beliefs. The resulting epistemic states will generally be different.

Proposition 2.24. $Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = Bel([\rho, \blacktriangle] * f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$

Example 2.25. $Bel([(), \neg p] * q * q \rightarrow p * p * q \wedge r) = Bel([(q, q \rightarrow p, p, q \wedge r), \neg p])$ *which is equivalent to* $Cn(\neg p \wedge q \wedge r)$*.* $Bel([(), \neg p] * f(q, q \rightarrow p, p, q \wedge r, \neg p)) = Bel([(\neg p \wedge q \wedge r), \neg p])$

yielding the same set of beliefs. However, if we consider a further input $\neg r$ *w.r.t. the resulting epistemic states, the belief sets differ.* $Bel([(q, q \rightarrow p, p, q \wedge r), \neg p] * \neg r) = Cn(\neg p \wedge \neg r \wedge \neg q)$ *while* $Bel([(\neg p \wedge q \wedge r), \neg p] * \neg r) = Cn(\neg p \wedge \neg r)$*. In connection with Proposition 2.41, we will give a hint as to why the equivalence does not carry over when considering a further revision step. The epistemic states after a sequence of revision steps and the corresponding single revision are generally different.*

One characteristic property of the agent model is the following variant of the rule "Recalcitrance" from [68] which is closely related to the property given in Proposition 2.20. This variant is suggested in [8] and states that if the agent believes an input φ_1, then it does so *wholeheartedly*, in that the only way it can be dislodged from the belief set by a succeeding input φ_2 is if that input contradicts it given the core beliefs $\blacktriangle$. We have already seen this happening in Example 2.18. Proposition 2.24 entails that φ_2 might also be the result of a series of revision inputs, i.e., φ_1 is believed until several succeeding inputs contradict it in the light of $\blacktriangle$.

Proposition 2.26. *If* $\blacktriangle \nvdash (\varphi_2 \rightarrow \neg\varphi_1)$ *then* $Bel([\rho, \blacktriangle] * \varphi_1 * \varphi_2) \vdash \varphi_1$

If we are not given the epistemic state of the agent, we cannot say anything about the revision history. We can prove no general property of the interaction of inputs received. This is because any input may have been received at any point in time. The following proposition tells us that no matter what the revision history ρ of the agent has been up to now, there is another revision history σ that will yield exactly the same revision behaviour *from now on*. But σ has a particular characteristic — up to now the agent has received inputs becoming logically increasingly stronger. In other words, each input logically entails its predecessor, i.e., the revision history σ is a logical chain. So, with respect to the agent's future revision behaviour we can act *as if* its past inputs were in the form of a logical chain. The proposition deals only with a single revision step, but from Proposition 2.24 we know that the beliefs after any sequence of revision steps can be characterised by a single revision.

Proposition 2.27. *For any epistemic state* $[\rho, \blacktriangle]$*, there exists an epistemic state* $[\sigma, \blacktriangle]$*, such that* σ *is a logical chain and for all* φ*:*

$$Bel([\rho, \blacktriangle] * \varphi) = Bel([\sigma, \blacktriangle] * \varphi)$$

To give an example, the epistemic state $[(p \vee q, q, p \wedge q), \top]$ will yield the same beliefs given any sequence of future revision inputs as $[(p, q), \top]$. Note that $(p \vee q, q, p \wedge q)$ is a logical chain whereas (p, q) is not. For the remainder of this thesis, Proposition 2.24 is of extreme importance as it allows us to view a sequence of revisions as a set of single revision steps in

the same initial state. We exploit this property of the belief revision framework for reasoning about the agent. Proposition 2.27 is of technical importance for a number of proofs as it allows us to restrict our attention to epistemic states $[\rho, \blacktriangle]$ where ρ is a logical chain.

$[\rho, \blacktriangle]$, $*$ and total preorders on worlds We have now described the belief revision framework the observed agent $\mathcal{A}$ is assumed to employ. As mentioned before, the epistemic state consisting of a core belief and a sequence of formulae can be seen as a syntactic representation for a total preorder on worlds. In the remainder of this section, we will illustrate this connection in more detail. We will make extensive use of results presented in [8].

First, we want to position the framework with respect to the AGM-postulates which we gave in Section 1.2. An epistemic state $[\rho, \blacktriangle]$ with an inconsistent core belief is not interesting as then the beliefs will be inconsistent for all sequences of revision inputs that may be received by $\mathcal{A}$. So assume $\blacktriangle$ to be consistent. When restricting our attention to revision inputs which are consistent with the agent's core belief $\blacktriangle$, we can show that all AGM-postulates are satisfied by $*$. By Definitions 2.17 and 2.16 closure is trivial. To see that success holds recall that an input is accepted if and only if it is consistent with the core belief. If the input is inconsistent with the current belief set then inclusion is trivial, otherwise Proposition 2.10 yields that revision amounts to adding the revision input and closing the beliefs under logical consequence. Hence, inclusion and preservation are satisfied. Under the current assumptions the revision input cannot be inconsistent and as the core belief is consistent the belief set will always be consistent — yielding consistency. Proposition 2.4 yields equivalence and a variant of Proposition 2.12 can be used to prove that superexpansion and subexpansion hold as well.

If the revision input φ is inconsistent with $\blacktriangle$ then Proposition 2.9 yields that we can act as if that input has not been received at all. Assuming $[\rho, \blacktriangle]$ to be the agent's epistemic state, let $\alpha = \blacktriangle$, $\beta = \varphi$ and σ be *any* sequence of revision inputs received after φ. In particular, after having received such a formula φ the belief set remains unchanged.

This method of using an AGM-revision operator in case the revision input is consistent with the core beliefs and leaving the belief set unchanged otherwise is what [8] calls a regular revision operator. And for each such operator there is a corresponding total preorder on worlds. Minimal worlds in that preorder correspond to the belief set and minimal φ-worlds (φ being consistent with the core beliefs) to the beliefs after revision by φ.

Note that $*$ is a core-invariant revision operator (Definition 2 in [8]), i.e., core beliefs before and after any revision are exactly the same. Together with the general property of core beliefs being a subset of the agent's belief set, this yields that the total preorder always strictly prefers worlds satisfying the core beliefs over worlds that do not (Proposition 1 in

[8]). As revision inputs contradicting the core ▲ are always rejected, we basically deal with a total preorder that contains all ▲-worlds but no ¬▲-world.

We showed in Proposition 2.26 that $*$ satisfies the property (C5') from [8]. Proposition 5 in that paper yields that in this case $*$ also satisfies the Darwiche and Pearl (DP) postulates (C3) and (C4) which state that if ψ is believed ($\neg\psi$ is not believed) after revision by φ then ψ is believed ($\neg\psi$ is not believed) after first revising by ψ and then revising by φ. The first two DP postulates cannot be satisfied in general. However, $*$ satisfies the following two variants (C1') and (C2') suggested in [8].[3] The following list contains the properties in terms of our assumed belief revision framework.

(C1') If $\blacktriangle \not\vdash \neg\varphi$ and $\varphi \vdash \psi$ then $Bel([\rho, \blacktriangle] * \psi * \varphi) = Bel([\rho, \blacktriangle] * \varphi)$

(C2') If $\blacktriangle \not\vdash \neg\varphi$ and $\varphi \vdash \neg\psi$ then $Bel([\rho, \blacktriangle] * \psi * \varphi) = Bel([\rho, \blacktriangle] * \varphi)$

(C3) If $Bel([\rho, \blacktriangle] * \varphi) \vdash \psi$ then $Bel([\rho, \blacktriangle] * \psi * \varphi) \vdash \psi$

(C4) If $Bel([\rho, \blacktriangle] * \varphi) \not\vdash \neg\psi$ then $Bel([\rho, \blacktriangle] * \psi * \varphi) \not\vdash \neg\psi$

(C5') If $\blacktriangle \not\vdash (\varphi_2 \rightarrow \neg\varphi_1)$ then $Bel([\rho, \blacktriangle] * \varphi_1 * \varphi_2) \vdash \varphi_1$

Given these properties Propositions 3 and 4 from [8] allow us to make the following statements about how the total preorder on worlds corresponding to an epistemic state $[\rho, \blacktriangle]$ changes upon revising it by a formula φ. We already argued that only ▲-worlds are present in the preorder and that if φ is not consistent with ▲ then nothing changes. So let $\blacktriangle \wedge \varphi$ be consistent. The order among φ-worlds remains untouched. The order among $\neg\varphi$-worlds remains unchanged, as well. However, after the revision every φ-worlds will be strictly preferred to any $\neg\varphi$-world. In other words, modulo the treatment of inputs contradicting the core belief which do not change the preorder at all, the assumed revision operation corresponds to Nayak's lexicographic revision [67] which we illustrated in Section 1.2.

Viewing the total preorder on worlds as a ranking of these worlds, iterated revision will refine the ranking, splitting up ranks and thus yielding more and more pairs of worlds where one is strictly more plausible than the other. Note that this is another way to illustrate Proposition 2.20. The splitting up of ranks of worlds cannot be undone in this framework. Worlds that were once placed into different ranks will never again meet in the same one, no matter what revision inputs will be received. So, the agent can never again be agnostic about the status of an input φ once received. As the input φ splits every rank into φ- and $\neg\varphi$-worlds, the lowest rank will henceforth contain either only φ-worlds or only $\neg\varphi$-worlds.

The epistemic state $[(), \blacktriangle]$ corresponds to the total preorder which contains only ▲-worlds and all of them are equally plausible, i.e., they are all in the lowest rank. The belief set

[3] The key to the proof, which looks very much like that for Proposition 2.9, is that since $\blacktriangle \not\vdash \neg\varphi$ and $\varphi \vdash (\neg)\psi$ we have that $f((\psi, \varphi) \cdot \blacktriangle) = f(\varphi \cdot \blacktriangle) = \varphi \wedge \blacktriangle$.

$Bel([(), \blacktriangle] * \varphi)$ is $Cn(\blacktriangle \wedge \varphi)$ if φ is consistent with $\blacktriangle$ and $Cn(\blacktriangle)$ if it is not. So all $\blacktriangle$-worlds must be equally plausible as otherwise the beliefs after revision would be stronger for some input φ. This gives us a method for transforming a state $[(\varphi_1, \ldots, \varphi_n), \blacktriangle]$ into a total preorder on worlds. We simply start with the preorder for $[(), \blacktriangle]$ and revise it with φ_1 using the procedure described above. The resulting preorder is then revised by φ_2 and so on.

If a formula φ is already believed, which means that all minimal worlds satisfy φ, revision by that formula does not modify the lowest rank of worlds. This corresponds to Proposition 2.22. However, the *other* ranks may change, i.e., although the belief set is not affected by the revision, the epistemic state may well be. Analogously, a single revision may suffice to get the same lowest rank as after a sequence of revisions and hence yield the same *beliefs* (Proposition 2.24), but in general it is not possible to get the same resulting total preorder on worlds. This is easy to see when viewing the preorder as a ranking of worlds. A single revision step can only double the number of (non-empty) ranks, two revision steps however may quadruple that number.

In the framework we use, ρ can be interpreted as the actual revision sequence bringing $\mathcal{A}$ from the preorder where there is only one rank containing all $\blacktriangle$-worlds to its current one. However, there are many (infinitely many, to be precise) revision sequences doing that. The result in Proposition 2.27 can be illustrated as follows. Assume $[\rho, \blacktriangle]$ is a state whose corresponding total preorder on worlds has n ranks (as we assume a finite language there can only be a finite number of ranks). The agent could have arrived at the same preorder (in retrospect) by starting in the epistemic state $[(), \blacktriangle]$ first revising by a formula that corresponds exactly to the worlds contained in the lowest $n-1$ ranks,[4] then by one corresponding to the lowest $n-2$ ranks and so on. Due to the subset relation among sets of worlds, the corresponding formulae will become increasingly stronger. In the end, the agent's revision history will be a logical chain and the preorder on worlds will be the same as for ρ. Hence, future revisions will have exactly the same effect.

The same method can be used for finding an epistemic state $[\sigma, \blacktriangle]$ for any total preorder on worlds. $\blacktriangle$ is the formula corresponding to all worlds present in the preorder. Then $[\sigma, \blacktriangle]$ is constructed starting in $[(), \blacktriangle]$ proceeding as described above. Note that each epistemic state $[\rho, \blacktriangle]$ gives rise to a unique total preorder on worlds but that the converse never holds. When we talk about two epistemic states having the same (future) belief revision behaviour this is what is meant. They give rise to the *same* total preorder on worlds which in turn completely determines the beliefs for any sequence of revision inputs.

An agent may receive an enormous number of revision inputs which is reflected in the length

[4] Interpreting a world as a conjunction of literals, the formula corresponding to a set of worlds is equivalent to the disjunction of the corresponding conjunctions.

of the sequence ρ of its epistemic state $[\rho, \blacktriangle]$. Obviously, we are unable to deal with infinite sequences in the current setting, but we still can say something about the convergence towards an infinite sequence. The above considerations about the interpretation of the revision framework in terms of a total preorder on worlds have several implications. As each revision input potentially refines the order, assuming a finite language, at some point the agent's epistemic state may be represented by a total order on worlds which means that from that point on the lowest rank will contain a single world — either the one the agent already believes or the most plausible one consistent with the last accepted revision input.

The agent's beliefs need not converge towards a belief held for all times, as can be easily seen from a sequence of revision inputs like $\rho = (p, \neg p, p, \neg p, \dots)$ the core belief entailing neither p nor $\neg p$. The beliefs need not even repeat themselves periodically, as the inputs p and $\neg p$ in this example might take turns in a non-periodic way. So, even considering a language with only one propositional variable, the evolution of beliefs of the agent can be non-trivial.

In the remainder of Chapter 2 and in Chapter 3, we will assume that the observed agent $\mathcal{A}$ employs the framework presented in this section for representing its epistemic state and for revising and calculating its beliefs. In Chapter 4, we will also look at some slightly different frameworks which we will introduce then.

2.3 Observations and their explanations

After having described the observed agent's assumed belief revision framework, we now turn to the specific information we receive about a particular agent $\mathcal{A}$ — some observation o on its belief revision behaviour. In this section, we want to define this term precisely and formally describe when an initial epistemic state explains o. We will use the term *initial state* to talk about the agent's state immediately before the observation started, not the state at the beginning of its life. An observation contains information about revision inputs $\mathcal{A}$ received, what it believed and did not believe upon receiving them. We will not deal with the question of *how* such an observation is obtained. Our motivating scenarios illustrated some settings in which information about revision inputs, beliefs and non-beliefs may be available, but we will not investigate how sensory data, communication protocols, case data, etc. can be transformed into the required format.

Definition 2.28. *An* observation $o = \langle(\varphi_1, \theta_1, D_1), \dots, (\varphi_n, \theta_n, D_n)\rangle$ *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.*

We will denote the sequence of all revision inputs $(\varphi_1, \dots, \varphi_n)$ *by* ι *and a prefix* $(\varphi_1, \dots, \varphi_i)$ *of length* i *of that sequence by* ι_i*.*

The definition of the syntax of an observation is not really useful without a meaning attached to it. The intuitive interpretation of an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ 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. 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.[5] 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 $D_i = \emptyset$.

As the agent's beliefs are closed under logical consequence (see Definition 2.17), we know that any element of $Cn(\theta_i)$ belongs to its beliefs after having received the first i revision inputs recorded in the observation. We will refer to the elements of D_i as non-beliefs[6] as these formulae do not belong to the agent's belief set. Note that any δ' such that $\delta' \vdash \delta$ for some $\delta \in D_i$ implicitly belongs to the non-beliefs as well. It is not possible to encode the non-beliefs from D_i as a single (propositional) formula. This fact is familiar from modal logic ($\neg\Box\varphi_1 \wedge \neg\Box\varphi_2$ cannot generally be rewritten to some $\neg\Box\varphi$). $\mathcal{A}$ may believe neither p nor $\neg p$. Further, not believing p and q is different from not believing $p \wedge q$ or $p \vee q$. Hence, the formulae not believed by the agent after a revision step are recorded as a set D_i of formulae.

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. As the beliefs recorded in o are correct, θ_i must be consistent with the agent's core belief $\blacktriangle$. Hence, φ_i is also consistent with $\blacktriangle$ and consequently must have been accepted. If $\theta_i \vdash \neg\varphi_i$ or $\varphi_i \vdash \delta$ for some $\delta \in D_i$ then it must have been rejected as otherwise it would have to be believed, contradicting the information in o. But if none of these conditions hold, it is not obvious whether an input has been accepted or rejected. Often, neither 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. Recall that the question of whether a revision input is accepted or rejected by the agent depends on its core belief only.

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

[5] We will relax this assumption in Chapter 3.

[6] We use the term non-belief rather than disbelief as for us the latter is attached to a stronger notion. While a formula θ may simply not be entailed by the agent's beliefs, a disbelief in θ implies that the agent actively rejects θ to be among its beliefs. The observation does not express which of the two cases applies, so we settle for the more neutral term.

of the agent's revision history. In particular, the agent will have received inputs before the observation started. In other words, 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 2.29. *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* $1 \leq i \leq n$*:*
   $Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \vdash \theta_i$ *and*
   $\forall \delta \in D_i : Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \nvdash \delta$

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 2.29 are satisfied, i.e., that $\blacktriangle$ is consistent and for all i $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \vdash \theta_i$ and $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \nvdash \delta$ for any $\delta \in D_i$.

Example 2.30. *(i)* $[\rho, \blacktriangle] = [(p \to q), r]$ *explains* $\langle(p, q, \emptyset), (q, r, \emptyset)\rangle$ *because* $f(p \to q, p, r)$ *and* $f(p \to q, p, q, r)$ *are equivalent to* $p \wedge q \wedge r$ *which entails* q *and* r.

(ii) $[(p \to q), \top]$ *does not explain* $\langle(p, q, \emptyset), (q, r, \emptyset)\rangle$ *because* $f(p \to q, p, q, \top) \equiv p \wedge q \nvdash r$.

(iii) $[(p \to q), \top]$ *does not explain* $\langle(p, \top, \{q\})\rangle$ *because* $f(p \to q, p, \top) \equiv p \wedge q \vdash q$.

The following proposition informs us about the computational complexity of that test.

Proposition 2.31. *The decision problem of whether* $[\rho, \blacktriangle]$ *explains an observation* o *is* Δ_2^P*-complete.*

A state with an inconsistent core belief satisfies the second condition in Definition 2.29 if and only if $D_i = \emptyset$ 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 2.32. *Consider the observation* $o = \langle(p, p, \emptyset), (q, q, \emptyset)\rangle$. *Intuitively, it expresses that after* $\mathcal{A}$ *received* p *and* q*, it did indeed believe those formulae, but tells us nothing about what is not believed. The following epistemic states are (a* very *small selection of the) explanations for* o*:* $[(), p \wedge q]$, $[(p \wedge q), \top]$, $[(r, s), \neg p \vee \neg q]$, $[(), \top]$.

This illustrates that generally there are many different explanations, even with different o-acceptable cores, for a single observation o. For the observation in Example 2.32 *any* epistemic state $[\rho, \blacktriangle]$ such that $\blacktriangle \nvdash \neg p$ and $\blacktriangle \nvdash \neg q$ is an explanation. Recall that an epistemic state defines a total preorder on worlds, which uniquely determines the agent's beliefs through a sequence of revisions. In general, there will be many such preorders that give rise to explaining states. Moreover, there are infinitely many epistemic states $[\rho, \blacktriangle]$ representing the same total preorder on worlds. Hence we know that whenever there is *one* explanation for o, there are *infinitely many*.

So, a first criterion for distinguishing different epistemic states is whether they are or are not an explanation for an observation. The general method for reasoning about an observed agent will be to select *one* explanation $[\rho, \blacktriangle]$ and use it to draw conclusions.[7] $\blacktriangle$ determines which formulae we predict to be accepted and rejected by the agent, and the belief trace (Definition 2.19) allows us to complete the information on beliefs and non-beliefs during the time of observation — even for an arbitrary sequence of revision inputs starting in the initial state. Let us look at the implications of choosing one of the possible explanations in Example 2.32. There are implications, because the epistemic state of an agent completely determines its future beliefs, i.e., what it will believe after an arbitrary sequence of revisions.

Selecting $[(), p \wedge q]$ as the explanation implies the claim that the agent believed p and q even before receiving them as inputs. A further claim is that $\mathcal{A}$ will never accept a revision input entailing $\neg p \vee \neg q$. The explanation $[(p \wedge q), \top]$ shares the first claim that the revision inputs did not change the belief set but does not assume the agent to reject any revision inputs. $[(r, s), \neg p \vee \neg q]$ implies that $\mathcal{A}$ believed $r \wedge s$ before the recorded inputs were received, kept believing that formula and that $\mathcal{A}$ cannot simultaneously believe both p and q.

The aim now is further to distinguish between better and worse explanations, i.e., potential initial epistemic states. Here it is less clear what the criteria should be. Basically, it depends on what the explanations are used for. If we choose an epistemic state $[\rho, \blacktriangle]$ as the explanation for an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ then we can answer the following questions.

- Which of the revision inputs recorded in o were accepted and which were rejected? Will the agent accept some other revision input φ? Here we would check whether the formulae in question are consistent with the assumed core belief. If so, we conclude they are accepted, if not, our conclusion is that they were rejected.

[7] It is a non-trivial task to find even one explanation for an observation. So we consider this to be a good first step. As we will see, reasoning about *all* explanations is possible in some cases. Quantitative considerations, such as investigating properties of the *majority* of explanations, are problematic and we will carry out no investigations in that direction.

- What did the agent believe immediately before the observation started? The answer to this question obviously is $Bel([\rho, \blacktriangle])$, i.e., $Cn(Bel_0^{[\rho,\blacktriangle]})$.

- What did the agent believe after the i^{th} revision step? Here we just need to calculate $Bel([\rho, \blacktriangle] * \varphi_1 * \ldots * \varphi_i)$ or equivalently $Cn(Bel_i^{[\rho,\blacktriangle]})$. Note that this question is not without justification, as not all beliefs are recorded in the observation. So in particular we might be interested in what else the agent believed, apart from θ_i.

- There is the question of future beliefs. What will the agent believe after receiving one or more further inputs $\psi_1, \ldots, \psi_j$? The answer is $Bel([\rho \cdot (\varphi_1, \ldots, \varphi_n), \blacktriangle] * \psi_1 * \ldots * \psi_j)$. Using the notion of a belief trace we would use $(\varphi_1, \ldots, \varphi_n, \psi_1, \ldots, \psi_j)$ as the sequence of revision inputs.

- Similarly, we might wonder what the agent's evolution of beliefs would have been when receiving a totally different sequence of revision inputs. In this case, we would look at the belief trace for $\mathcal{A}$'s assumed initial state using that sequence.[8]

The quality of these answers depends on how close our explanation $[\rho, \blacktriangle]$ gets to the actual initial epistemic state of the agent. Depending on our choice of $[\rho, \blacktriangle]$ the conclusions we draw about $\mathcal{A}$'s belief revision behaviour can vary greatly. Naturally, we would like to draw conclusions we can support using the observation. In example 2.32, a possible explanation was $[(r, s), \neg p \vee \neg q]$. If we were asked what the agent will believe after receiving the revision input $p \wedge q$, we would have to answer $Cn(r \wedge s \wedge \neg p \wedge q)$. We would be wrong with respect to all but the variable p, if we are informed that after a further revision by $p \wedge q$, $\mathcal{A}$ believed $p \wedge q$ and did not believe r or s, that is, if the observation was extended to $\langle (p, p, \emptyset), (q, q, \emptyset), (p \wedge q, p \wedge q, \{r, s\}) \rangle$.

Nothing in the original observation told us anything about r or s. Nothing indicated that p and q could not be believed simultaneously by the agent. So there is insufficient justification for choosing $[(r, s), \neg p \vee \neg q]$ as the explanation. Note that we would not have been wrong with respect to the extension of the observation had we chosen $[(), \top]$.

The following proposition tells us that if we have an explanation for some observation o then we can immediately construct one for any subobservation of o, i.e., for any o' such that $o = o_1 \cdot o' \cdot o_2$. More precisely, any o-acceptable core is also o'-acceptable. As a special case we get that the same explanation can be used for any prefix of o.

[8] This problem is interesting, e.g., in the expert scenario. We may have observed the expert reasoning in one case and want to predict what she might conclude in another one. We will discuss this scenario in Section 4.2.

Proposition 2.33. *If* $[\rho, \blacktriangle]$ *is an explanation for* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *then* $[\rho \cdot (\varphi_1, \ldots, \varphi_{j-1}), \blacktriangle]$ *explains* $o' = \langle(\varphi_j, \theta_j, D_j), \ldots, (\varphi_{j+k}, \theta_{j+k}, D_{j+k})\rangle$, $1 \leq j + k \leq n$.

The contraposition then tells us that if $\blacktriangle$ is not o'-acceptable, it cannot be o-acceptable. In particular, if $\blacktriangle$ is not $\langle(\varphi_i, \theta_i, D_i)\rangle$-acceptable for some i, it cannot be o-acceptable.

The essence of the remainder of this chapter is to give and justify criteria for comparing epistemic states as well as to give a method for calculating an epistemic state that is best with respect to these criteria. As we saw in the answers to the possible questions concerning an observed agent, the core belief and the belief trace play an important role, determining the conclusions we draw about $\mathcal{A}$. So, the criteria for comparing epistemic states will focus on these two aspects.

2.4 The weakest core belief

Given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$, finding an o-acceptable core is not trivial. Appendix B contains an example illustrating that a given observation o may have extremely few o-acceptable core beliefs. Also, there are very simple examples showing that almost nothing can be said about the o-acceptability of a formula given the status of its subformulae, i.e., whether they are o-acceptable or not.

Example 2.34. *Each cell in the following tables contains an observation* o*. A row heading* $\lambda(+)$*, where* $\lambda \in \{p, q\}$*, indicates that the formula* λ *is* o*-acceptable,* $\lambda(-)$ *that it is not* o*-acceptable. Column headings give the same information with respect to the formulae constructed from those in the row headings. Taking the last column of the middle row in the second table as an example (*$o = \langle(\neg q, \neg q, \emptyset)\rangle$*),* p *is* o*-acceptable but* q *and* $p \wedge q$ *are not. The case* $p(-), q(+)$ *is essentially the same as* $p(+), q(-)$*. We just need to interchange* p *and* q *to get the corresponding examples.*

	$\neg p(+)$	$\neg p(-)$
$p(-)$	$\langle(p, \top, \{p\})\rangle$	$\langle(q, \top, \{q\})\rangle$
$p(+)$	$\langle(q, q, \emptyset)\rangle$	$\langle(p, p, \emptyset)\rangle$

		$p \wedge q(+)$	$p \wedge q(-)$
$p(-)$	$q(-)$	$\langle(\neg p, \top, \{\neg p\}), (\neg q, \top, \{\neg q\})\rangle$	$\langle(\neg p, \neg p, \emptyset), (\neg q, \neg q, \emptyset)\rangle$
$p(+)$	$q(-)$	$\langle(\neg p, \top, \{\neg p\})\rangle$	$\langle(\neg q, \neg q, \emptyset)\rangle$
$p(+)$	$q(+)$	$\langle(r, r, \emptyset)\rangle$	$\langle(p \leftrightarrow \neg q, p \leftrightarrow \neg q, \emptyset)\rangle$

		$p \vee q(+)$	$p \vee q(-)$
$p(-)$	$q(-)$	$\langle(\neg p, \neg p, \emptyset), (\neg q, \neg q, \emptyset)\rangle$	$\langle(r, \top, \{r\})\rangle$
$p(+)$	$q(-)$	$\langle(\neg q, \neg q, \emptyset)\rangle$	$\langle(\neg p, \top, \{\neg p\})\rangle$

In the last table, the row for $p(+), q(+)$ is missing. The reason is that in this case we have a general result which is one of the key results for the work presented. It states that the disjunction of any two o-acceptable cores will also be o-acceptable.

Proposition 2.35. *If $\blacktriangle_1$ and $\blacktriangle_2$ are o-acceptable then so is $\blacktriangle_1 \vee \blacktriangle_2$.*

One consequence of this proposition is that in case at least one o-acceptable core exists, there is a logically weakest o-acceptable core belief. It can be retrieved using the following function mapping observations to formulae. Recall that we consider a finite language. This entails that there are only finitely many logically different formulae. So although there may be an infinite number of o-acceptable cores, there will be a formula that is equivalent to their disjunction.

Definition 2.36.

$$\blacktriangle_\vee(o) = \begin{cases} \bigvee\{\blacktriangle \mid \blacktriangle \textit{ is } o\textit{-acceptable}\} & \textit{if an } o\textit{-acceptable core exists} \\ \bot & \textit{otherwise} \end{cases}$$

By this definition $\blacktriangle_\vee(o)$ is inconsistent if and only if there is no o-acceptable core. So, we can read off whether an o-acceptable core exists. It is easy to see that any o-acceptable core entails $\blacktriangle_\vee(o)$, and by Proposition 2.35, $\blacktriangle_\vee(o)$ is indeed o-acceptable, provided an o-acceptable core exists at all. But why is this particular core so interesting? Recall that a core belief of an agent $\mathcal{A}$ is a belief it will commit to at all times. Coming up with a belief that is entailed by *any* possible core belief keeps us on the safe side with respect to conclusions about the agent's core belief. $\mathcal{A}$ will never accept an input contradicting the core belief, i.e., the weaker the core belief, the fewer formulae we will claim to be rejected. We are cautious and formulae we claim to be rejected will be rejected by any possible o-acceptable core.

Proposition 2.37. *The function $\blacktriangle_\vee(\cdot)$ satisfies the following properties of a function $\blacktriangle(\cdot)$ mapping observations to formulae, for any observations o, o':*

(Acceptability)	*If an o-acceptable core exists then $\blacktriangle(o)$ is o-acceptable.*
(Consistency)	*If $\blacktriangle(o) \not\equiv \bot$ then there is an o-acceptable core.*
(Right Monotony)	$\blacktriangle(o \cdot o') \vdash \blacktriangle(o)$
(Left Monotony)	$\blacktriangle(o' \cdot o) \vdash \blacktriangle(o)$

Acceptability and Consistency are minimal requirements we would expect to be satisfied by any function that calculates the core belief to be used for explaining an observation. They basically state that we are returned an o-acceptable core if there exists one and that if we are returned a consistent formula it is indeed o-acceptable. Left Monotony and Right Monotony express that we draw only safe conclusions about the core belief — conclusions that cannot

be defeated by extensions of the observation. If Right Monotony was violated, it would be possible to get $\blacktriangle$ as core for o and $\blacktriangle'$ as core for $o \cdot o'$ with $\blacktriangle' \not\vdash \blacktriangle$. But by Proposition 2.33 $\blacktriangle'$ is also o-acceptable and hence $\blacktriangle \vee \blacktriangle'$ is o-acceptable. This means that we would have attributed the agent a core belief that is stronger than necessary which is undesirable as argued above. An analogous argument exists for the violation of Left Monotony. Hence, we claim that these two properties should also be satisfied by a function yielding good o-acceptable cores.

Proposition 2.38. *Let $\blacktriangle(\cdot)$ be any function which returns a formula given any observation o. Then the following are equivalent:*
(i) $\blacktriangle(\cdot)$ satisfies Acceptability, Consistency and Right Monotony.
(ii) $\blacktriangle(\cdot)$ satisfies Acceptability, Consistency and Left Monotony.
(iii) $\blacktriangle(o) \equiv \blacktriangle_{\vee}(o)$ for all observations o.

This proposition tells us that in the presence of Acceptability and Consistency, the properties Left Monotony and Right Monotony turn out to be equivalent. More importantly, it shows that $\blacktriangle_{\vee}(\cdot)$ is uniquely characterised by those properties — any function satisfying these desirable properties will turn out to be equivalent to $\blacktriangle_{\vee}(\cdot)$.

This is a good time to comment on the assumption we made about the observations we receive. We required that the agent did not receive any inputs between φ_i and φ_{i+1}, i.e., between neighbouring revision inputs in the observation. Left Monotony and Right Monotony express that conclusions we draw about the core belief of an agent do not become invalid when the observation is extended *at the ends*, i.e., when we get additional information about what the agent received right before or right after receiving the revision inputs recorded in the original observation. Such a property cannot be found when inserting some new observation somewhere in the original one. In these cases, acceptable cores may have to be logically stronger, but they also may become weaker!

Example 2.39. *Consider $o = \langle (p, p, \emptyset), (q, \neg p, \emptyset)\rangle$ for which $\blacktriangle_{\vee}(o) = q \rightarrow \neg p$. If the core belief did not entail $q \rightarrow \neg p$, p could not be blocked from being introduced into the belief set and hence $\neg p$ could not be believed consistently (see Proposition 2.26). If we now assume the additional input $\neg p$ (with no further information) between the inputs on record, we have $o' = \langle (p, p, \emptyset), (\neg p, \top, \emptyset), (q, \neg p, \emptyset)\rangle$ and $\blacktriangle_{\vee}(o') = \top$.*

This example shows that intermediate inputs may explain effects we had to attribute to the core belief, in this case that p must be blocked. Relaxing the assumption means losing the property of drawing safe conclusions, which is why we imposed it for the current chapter. We will investigate the case where information about the inputs received is incomplete in the next chapter.

When choosing a particular epistemic state $[\rho, \blacktriangle]$ for explaining an observation o, we propose to use $\blacktriangle \equiv \blacktriangle_\vee(o)$ as the core belief. It is entailed by any o-acceptable core which ensures that any formula we predict $\mathcal{A}$ to reject will indeed be rejected. Using *any other* o-acceptable core, this would not be the case.

2.5 Conditional beliefs and rational closure

As Example 2.32 indicated, generally there are many possible explanations $[\rho, \blacktriangle]$ for an observation o. Whereas the last section dealt with identifying a best o-acceptable core belief $\blacktriangle$, this section will talk about ρ. In order to identify a suitable sequence, we will make use of work dealing with conditionals. As noted before, the epistemic state of an agent corresponds to a total preorder on worlds which in turn has been shown to define a rational consequence relation. We will formally prove this property of the revision framework now and show a way for translating an observation into a partial description of a rational consequence relation. Completing this information allows us to come up with a potential initial state of the agent.

Definition 2.40. $\varphi \Rightarrow_{[\rho,\blacktriangle]} \theta$ *is a* conditional belief *(satisfied) in the epistemic state* $[\rho, \blacktriangle]$ *if and only if* $\blacktriangle \vdash \neg\varphi$ *or* $Bel([\rho, \blacktriangle] * \varphi) \vdash \theta$.

The intuition behind a conditional belief $\varphi \Rightarrow \theta$ in the current setting is as follows. If the agent were to receive a revision input φ and accept it, then it would believe θ (as well). With this interpretation of a conditional belief in mind, the second condition for a conditional belief to hold in an epistemic state is clear. The first one is justified by arguing that as the input φ will not be accepted since it contradicts the core belief, we need not care what the agent would believe after accepting it. The following result tells us that our assumed revision framework gives rise to a rational consequence relation whose properties we recalled in Section 1.2.2.

Proposition 2.41. $\Rightarrow_{[\rho,\blacktriangle]}$ *is a rational consequence relation.*

Recall Proposition 2.24 which stated that as far as beliefs go, any sequence of revision steps with inputs $\varphi_1, \ldots, \varphi_i$ can be interpreted as a single revision by $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ where $\blacktriangle$ is the agent's core belief. In other words, the agent's beliefs, in particular also the future ones, are *completely* characterised by $\mathcal{A}$'s core belief and the conditional beliefs in its initial epistemic state. However, the epistemic state $[\rho \cdot \varphi_1, \ldots, \varphi_i, \blacktriangle]$ resulting from a sequence of revisions in a state $[\rho, \blacktriangle]$ will generally be different from the one resulting from a single revision by $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$. This can also be seen from their respective conditional beliefs.

Consider $[(), \top]$, $\varphi_1 = p$, $\varphi_2 = q$. The beliefs after the sequence of revisions and the corresponding single revision by $f(p, q, \top) = p \wedge q$ are both represented by $p \wedge q$. But whereas $\neg q \Rightarrow_{[(p,q),\top]} p$ we have $\neg q \not\Rightarrow_{[(p \wedge q),\top]} p$.

We can now combine the above results: If the agent's core belief $\blacktriangle$ is known, an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ can be transformed into a partial description of the agent's conditional beliefs in its state just before the observation started. We can read off conditionals that should be satisfied, the set of which we will denote with $\mathcal{C}_{\blacktriangle}(o)$, and conditionals that should not be satisfied, the set of which we will denote with $\mathcal{N}_{\blacktriangle}(o)$. Recall that ι_i denotes $(\varphi_1, \ldots, \varphi_i)$ and ι the sequence of all revision inputs recorded in o.

Definition 2.42. *Given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and a core belief $\blacktriangle$, we define*

$$\mathcal{C}_{\blacktriangle}(o) = \{f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i \mid i = 1, \ldots, n\} \text{ and}$$
$$\mathcal{N}_{\blacktriangle}(o) = \{f(\iota_i \cdot \blacktriangle) \Rightarrow \delta \mid i = 1, \ldots, n \wedge \delta \in D_i\}.$$

In the following we will call conditionals that should be satisfied (elements of $\mathcal{C}_{\blacktriangle}(o)$) positive conditionals and those that should not be satisfied (elements of $\mathcal{N}_{\blacktriangle}(o)$) negative conditionals. Note that for consistent $\blacktriangle$ the antecedents $f(\iota_i \cdot \blacktriangle) = f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ of the conditionals thus defined are always accepted by the agent. This is because they are consistent with $\blacktriangle$ by definition. This means that the positive conditionals will all meet the non-trivial condition in Definition 2.40. Hence, if we find an epistemic state satisfying all the positive conditionals, the θ_i will indeed be believed, and if none of the negative ones is satisfied, no $\delta \in D_i$ will be believed after the corresponding revision. The case of an inconsistent core belief $\blacktriangle$ is not interesting as then $\blacktriangle$ is not o-acceptable for any observation o and hence we need not find a sequence, anyway.

In the proofs of results we will present in the following, we sometimes need to refer to particular conditionals. Each record $(\varphi_i, \theta_i, D_i)$ in an observation gives rise to one positive and a set of negative conditionals, all of which share the same antecedent $f(\iota_i \cdot \blacktriangle)$. Often it will be convenient to refer to all conditionals such a record gave rise to at the same time. To do so, we will use the number i — the position of that record in the observation — calling it *index*. This term will also be used with a different meaning in Section 4.6.

Definition 2.43. *Given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$, a core belief $\blacktriangle$ and a conditional $f(\iota_i \cdot \blacktriangle) \Rightarrow \mu$, we call the number i the* index *of that conditional.*

We call $I_{\mathcal{C}} = \{i \mid i \text{ is index of a conditional } \lambda \Rightarrow \mu \text{ in } \mathcal{C}\}$ the index set *of a set of conditionals $\mathcal{C}$ constructed from an observation.*

Having an observation and a core belief, Definition 2.42 allows us to construct a partial description of the agent's conditional beliefs in its epistemic state. We also know that the core belief and the set of all conditional beliefs in the initial state completely determine the agent's beliefs following any sequence of revision inputs. Any epistemic state defines a rational consequence relation. So, our job is now to complete the partial description in a reasonable way in order to arrive at a full description of a rational consequence relation.

2.5.1 the rational closure construction

A method for completing a set of conditionals to get a rational consequence relation, taking into account both positive and negative information as is necessary in our case, is the *rational closure* construction presented in [14] (extending the case of positive-only information studied in [54]). We will use this method to construct the sequence ρ needed to complete the agent's initial epistemic state (recall that $\blacktriangle$ is assumed to be given). In this section, we will illustrate the rational closure construction. In the next one, we will present results justifying that the sequence ρ thus constructed can be considered better than any other sequence. We deviate slightly from the notation used in [14].

Definition 2.44. *Given a set of conditionals $\mathcal{C} = \{\lambda_i \Rightarrow \mu_i \mid i = 1, \ldots, l\}$ we denote by $\tilde{\mathcal{C}} = \{\lambda_i \rightarrow \mu_i \mid i = 1, \ldots, l\}$ the set of material counterparts of all the conditionals in $\mathcal{C}$.*

A conditional $\lambda \Rightarrow \mu$ is p-exceptional *for a set of formulae U if and only if $U \vdash \neg\lambda$. $\lambda \Rightarrow \mu$ is* n-exceptional *for U if and only if $U \cup \{\lambda\} \vdash \mu$.*

Before providing the technical details, we want to give an intuition on what the construction will be about. We will do so with respect to our specific setting. The rational closure construction does not require the positive and negative conditionals to be of the particular form Definition 2.42 yields.

Assume we have constructed a suitable sequence ρ, given an observation o and a consistent core belief $\blacktriangle$; an inconsistent core is not interesting as it is not o-acceptable. So we have the initial state $[\rho, \blacktriangle]$ we were looking for. One way to think of U — one of the sets from the rational closure construction — is as the set of formulae collected from ρ when constructing $\mathcal{A}$'s belief set after revising $[\rho, \blacktriangle]$ by $f(\iota_i \cdot \blacktriangle)$ (or equivalently after iterated revision by $\varphi_1, \ldots, \varphi_i$). First of all, U must be consistent with $f(\iota_i \cdot \blacktriangle)$. Otherwise, U is not the correct set of formulae ($f(\rho \cdot f(\iota_i \cdot \blacktriangle))$ entails $f(\iota_i \cdot \blacktriangle)$ and is consistent as $\blacktriangle$ is). In other words, $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ must not be p-exceptional for U. Secondly, U must entail $f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i$, as otherwise θ_i will not be believed after the revision and hence the assumed initial state would not explain o. This is where the material counterparts of the conditionals come into play. Thirdly, $U \cup \{f(\iota_i \cdot \blacktriangle)\}$ must not entail any $\delta \in D_i$, as otherwise we would be violating the

negative information in the observation, again causing the assumed state not to explain o. This last condition amounts to demanding that $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ must not be n-exceptional for U.

Note that one such U will have to satisfy these conditions for all conditionals *with the same index* at the same time. Otherwise $[\rho, \blacktriangle]$ cannot be an explanation for o. It should be clear that in general there is not a single set U satisfying all these conditions for the conditionals with *all indexes*. So we construct different sets U_i, each (successfully) dealing with different indexes. In fact we start off with one set U_0 that tries to deal with all. For some indexes U_0 may meet all the conditions, so those do not have to be considered further. The next one, U_1, tries to deal with all those indexes for which U_0 was not successful, and so on until the conditions for all indexes have been met by some U_i. So, one could also think of U as a representation of those positive conditionals that still have to be satisfied while making sure that no negative conditional with the same index is satisfied.

Turning to the formal definition, assume we are given a set $\mathcal{C}$ of positive conditionals and a set $\mathcal{N}$ of negative ones. The sequence $\rho_R(\mathcal{C}, \mathcal{N})$ corresponding to the rational closure of $\mathcal{C}$ and $\mathcal{N}$ is determined as follows. We define two decreasing sets of conditionals $\mathcal{C}_0 \supseteq \mathcal{C}_1 \supseteq \cdots \supseteq \mathcal{C}_m$ and $\mathcal{N}_0 \supseteq \mathcal{N}_1 \supseteq \cdots \supseteq \mathcal{N}_m$ and a decreasing set of formulae $U_0 \supseteq U_1 \supseteq \cdots \supseteq U_m$ — the U_i will be defined via a least fixpoint (lfp) construction.[9]

Definition 2.45. *Let $\mathcal{C}$ be a set of positive conditionals and $\mathcal{N}$ a set of negative conditionals. Then the sequence $\rho_R(\mathcal{C}, \mathcal{N})$ corresponding to the rational closure of $\mathcal{C}$ and $\mathcal{N}$ is $\rho_R(\mathcal{C}, \mathcal{N}) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0)$ where $\bigwedge \emptyset = \top$ and*

1. *$\mathcal{C}_0 = \mathcal{C}$ and $\mathcal{N}_0 = \mathcal{N}$*
2. *$U_i = \tilde{\mathcal{C}}_i \cup lfp(\{\neg\lambda \mid \lambda \Rightarrow \mu \in \mathcal{N}_i$ and $\lambda \Rightarrow \mu$ is n-exceptional for $U_i\})$*
3. *$\mathcal{C}_{i+1}$ is the set of conditionals in $\mathcal{C}_i$ that are p-exceptional for U_i and $\mathcal{N}_{i+1}$ is the set of conditionals in $\mathcal{N}_i$ that are n-exceptional for U_i*
4. *m is minimal such that $\mathcal{C}_m = \mathcal{C}_{m+1}$ and $\mathcal{N}_m = \mathcal{N}_{m+1}$*

$\mathcal{C}_0$ and $\mathcal{N}_0$ contain all conditionals that need to be considered. According to condition 2,[10] U_i is initialised with $\tilde{\mathcal{C}}_i$ — the set of material counterparts of the positive conditionals that still need to be satisfied. Then we go through all the negative conditionals in $\mathcal{N}_i$. If there

[9] The definition captures exactly the algorithm given in [14]. However, it is a reformulation as it deals directly with formulae. It also hides the construction of the U_i using the least fixpoint notation.

[10] U_i is the smallest set which contains $\tilde{\mathcal{C}}_i$ and which is closed under the following condition: If $\lambda \Rightarrow \chi$ is in $\mathcal{N}_i$ and $\lambda \Rightarrow \chi$ is n-exceptional for U_i then $\neg\lambda \in U_i$.

is a conditional $\lambda \Rightarrow \mu$ that is n-exceptional for U_i, which means that adding λ to U_i would cause μ to become inferable, its negated antecedent $\neg\lambda$ is added to U_i. The addition of this $\neg\lambda$ may cause *other* negative conditionals in $\mathcal{N}_i$ to become n-exceptional, so we then need to check $\mathcal{N}_i$ for conditionals that are n-exceptional for the set thus obtained. This process — the least fixpoint construction for finding U_i — stops if no further negated antecedents of negative conditionals have to be added. In other words, U_i is the smallest set containing $\tilde{\mathcal{C}}_i$ and being closed under the condition that if $\lambda \Rightarrow \mu \in \mathcal{N}_i$ is n-exceptional for U_i then $\neg\lambda \in U_i$. It is obvious that we can define a total order on the negative conditionals n-exceptional for U_i corresponding to the order in which they become exceptional.[11]

In our setting, where $\mathcal{C}$ and $\mathcal{N}$ are not arbitrary sets of conditionals, the following can be observed. If *one* conditional with index j is exceptional, then *all* conditionals with that index will be exceptional as well. All conditionals with the same index j share the same antecedent $f(\iota_j \cdot \blacktriangle)$. If a positive conditional $f(\iota_j \cdot \blacktriangle) \Rightarrow \theta_j$ is p-exceptional for U_i then $U_i \vdash \neg f(\iota_j \cdot \blacktriangle)$ and hence adding $f(\iota_j \cdot \blacktriangle)$ to U_i makes *any* formula inferable, in particular the consequent δ of any negative conditional $f(\iota_j \cdot \blacktriangle) \Rightarrow \delta$ which has the same index. So any negative conditional with the same index will be n-exceptional for U_i. If a negative conditional $f(\iota_j \cdot \blacktriangle) \Rightarrow \delta$ is n-exceptional for U_i then $\neg f(\iota_j \cdot \blacktriangle)$ is added to U_i (condition 2 in Definition 2.45) which causes any conditional with that same antecedent (and hence all conditionals with the same index) to become exceptional for the modified U_i. This implies that in our setting we can equivalently define a total order on the *indexes* which represents the order in which the negative conditionals became exceptional for U_i. Again, this order need not be unique and we will later use $\prec_e$ to denote a suitable one.

Returning to the construction, $\mathcal{C}_{i+1}$ and $\mathcal{N}_{i+1}$ then contain the conditionals that still need to be considered and we can stop if these sets do not change with respect to $\mathcal{C}_i$ and $\mathcal{N}_i$. This is expressed in condition 4. We will sometimes refer to the conditionals in $\mathcal{C}_m$ and $\mathcal{N}_m$ as *ultimately exceptional* as they would continue to be propagated to the next level if we did not stop.

If $\mathcal{N} = \emptyset$ in the above process then the process simplifies to the one given in, e.g., [12, 28] which handles the case of positive conditionals only.

Writing α_i for $\bigwedge U_i$, the rational closure of $\mathcal{C}$ and $\mathcal{N}$ is then the relation $\Rightarrow_R$ given by $\lambda \Rightarrow_R \mu$ if and only if either $\alpha_m \vdash \neg\lambda$ or $\big[\alpha_j \wedge \lambda \vdash \mu$ where j is minimal such that $\alpha_j \nvdash \neg\lambda\big]$. Since $\alpha_m \dashv \cdots \dashv \alpha_0$ it is easy to check that in fact this second disjunct is equivalent to $f(\alpha_m, \ldots, \alpha_0, \lambda) \vdash \mu$. As is shown in [14], $\Rightarrow_R$ satisfies all the conditionals in $\mathcal{C}$ and and none of the ones in $\mathcal{N}$. That is $\lambda \Rightarrow_R \mu$ for all positive conditionals $(\lambda \Rightarrow \mu) \in \mathcal{C}$, while $\lambda \not\Rightarrow_R \mu$ for all negative conditionals $(\lambda \Rightarrow \mu) \in \mathcal{N}$.

[11] We want to remark that generally there is no unique order.

Definition 2.46. *For* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *and a core belief* $\blacktriangle$ *we call* $\rho_R(\mathcal{C}_\blacktriangle(o), \mathcal{N}_\blacktriangle(o))$ *the* rational prefix *of* o *with respect to* $\blacktriangle$*, and will denote it by* $\rho_R(o, \blacktriangle)$.

We will now go through a complex example which illustrates the calculation of the rational prefix of an observation with respect to a given core belief. First the sets of positive and negative conditionals are constructed and then the rational closure is applied to them.

Example 2.47. *Let* $o = \langle(p, s, \emptyset), (q, t, \{s\}), (r, \top, \{\neg(p \wedge q), u\})\rangle$. *$o$ states that the agent after receiving p believes s, after then receiving q believes t but ceases to believe s. Having received the final input r, we are only informed that the agent does not believe $\neg(p \wedge q)$ or u. Using the core belief $\top$ this translates into the following sets of conditionals.*

$$\begin{aligned}
\mathcal{C}_0 = \mathcal{C}_\blacktriangle(o) &= \{f(\iota_1 \cdot \top) \Rightarrow s, f(\iota_2 \cdot \top) \Rightarrow t, f(\iota_3 \cdot \top) \Rightarrow \top\} \\
&= \{f(p, \top) \Rightarrow s, f(p, q, \top) \Rightarrow t, f(p, q, r, \top) \Rightarrow \top\} \\
&= \{p \Rightarrow s, p \wedge q \Rightarrow t, p \wedge q \wedge r \Rightarrow \top\} \\
\mathcal{N}_0 = \mathcal{N}_\blacktriangle(o) &= \{f(\iota_2 \cdot \top) \Rightarrow s, f(\iota_3 \cdot \top) \Rightarrow \neg(p \wedge q), f(\iota_3 \cdot \top) \Rightarrow u\} \\
&= \{f(p, q, \top) \Rightarrow s, f(p, q, r, \top) \Rightarrow \neg(p \wedge q), f(p, q, r, \top) \Rightarrow u\} \\
&= \{p \wedge q \Rightarrow s, p \wedge q \wedge r \Rightarrow \neg(p \wedge q), p \wedge q \wedge r \Rightarrow u\}
\end{aligned}$$

$\tilde{\mathcal{C}}_0 = \{p \rightarrow s, p \wedge q \rightarrow t, p \wedge q \wedge r \rightarrow \top\}$. *The negative conditionals* $p \wedge q \wedge r \Rightarrow \neg(p \wedge q)$ *and* $p \wedge q \wedge r \Rightarrow u$ *are not n-exceptional for this set as* $\tilde{\mathcal{C}}_0 \cup \{p \wedge q \wedge r\}$ *does not entail* $\neg(p \wedge q)$ *or* u. *However,* $p \wedge q \Rightarrow s$ *is n-exceptional for* $\tilde{\mathcal{C}}_0$. *If* $p \wedge q$ *is added then* s *becomes inferable. So in order to arrive at the final* U_0, $\neg(p \wedge q)$ *must be added to* $\tilde{\mathcal{C}}_0$.

To see if we have reached a fixpoint, we have to check that no other negative conditionals are n-exceptional for the set constructed so far. But we see that adding $\neg(p \wedge q)$ *made* $p \wedge q \wedge r \Rightarrow \neg(p \wedge q)$ *n-exceptional, which it was not before, so now* $\neg(p \wedge q \wedge r)$ *has also to be added. We see that first the negative conditional* $f(\iota_2 \cdot \top) \Rightarrow s$ *with index* 2 *became n-exceptional and then the conditional* $f(\iota_3 \cdot \top) \Rightarrow \neg(p \wedge q)$ *with index* 3. *Due to the addition of* $\neg(p \wedge q \wedge r)$, *the negative conditional* $p \wedge q \wedge r \Rightarrow u$ *with index* 3 *automatically became n-exceptional. Adding* $\neg(p \wedge q \wedge r)$ *again to what is to become* U_0 *is not necessary as it would leave the set unchanged. The index set of the negative conditionals which are exceptional for* U_0 *can hence be totally ordered via* $2 \prec_e 3$.

We have now reached a fixpoint, as there are no further negative conditionals. $U_0 = \{p \rightarrow s, p \wedge q \rightarrow t, p \wedge q \wedge r \rightarrow \top, \neg(p \wedge q), \neg(p \wedge q \wedge r)\}$.

Of the positive conditionals in $\mathcal{C}_0$ *only* $p \Rightarrow s$ *is not p-exceptional for* U_0 *and all negative conditionals in* $\mathcal{N}_0$ *are n-exceptional for* U_0. *So we have*

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

This time $U_1 = \tilde{\mathcal{C}_1} = \{p \wedge q \rightarrow t, p \wedge q \wedge r \rightarrow \top\}$ *as adding* $p \wedge q$ *does not make* s *inferable anymore, and since* $\neg(p \wedge q)$ *need not be added,* $p \wedge q \wedge r \Rightarrow \neg(p \wedge q)$ *or* $p \wedge q \wedge r \Rightarrow u$ *do not become n-exceptional either. Further, none of the positive conditionals is p-exceptional for* U_1*, so* $\mathcal{C}_2 = \emptyset = \mathcal{C}_3$*,* $\mathcal{N}_2 = \emptyset = \mathcal{N}_3$ *and hence* $U_2 = \emptyset = U_3$*. Making use of logical equivalences, we get*

$$\rho_R(o, \blacktriangle) = (\top, p \wedge q \rightarrow t, p \rightarrow (s \wedge \neg q))\,.$$

$f(\rho_R(o, \blacktriangle) \cdot (p, \top)) \equiv p \wedge s \wedge \neg q$ *which indeed entails* s*.* $f(\rho_R(o, \blacktriangle) \cdot (p, q, \top)) \equiv p \wedge q \wedge t$ *which entails* t *and does not entail* s*.* $f(\rho_R(o, \blacktriangle) \cdot (p, q, r, \top)) \equiv p \wedge q \wedge r \wedge t$ *which entails* $\top$ *and entails neither* $\neg(p \wedge q)$ *nor* u*. So all positive conditionals are indeed satisfied and none of the negative ones is.*

2.5.2 properties of the rational prefix

In this section, we will collect properties of the rational prefix. In particular, we will show that it helps determine whether a particular core belief is o-acceptable for an observation o. We will also provide justifications for the claim that the rational prefix is the best sequence to use for explaining an observation. The first result merely confirms that the rational prefix is indeed a logical chain.

Proposition 2.48. $\rho_R(\mathcal{C}, \mathcal{N}) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0)$ *is a logical chain, that is,* $\bigwedge U_i \vdash \bigwedge U_{i+1}$ *for* $0 \leq i \leq m-1$*.*

We already know that an inconsistent core belief cannot be o-acceptable (Definition 2.29). The next two results provide a necessary and sufficient condition for a consistent core belief $\blacktriangle$ to be o-acceptable. All we have to do is calculate the rational prefix of o with respect to $\blacktriangle$. If the weakest element of the sequence calculated is a tautology then $\blacktriangle$ is o-acceptable. If it is not a tautology then $\blacktriangle$ is not o-acceptable, i.e., there is *no* sequence ρ such that $[\rho, \blacktriangle]$ explains o.

Proposition 2.49. *Given* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *and a core belief* $\blacktriangle \not\equiv \bot$ *let* $\rho_R(o, \blacktriangle) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0)$ *be the rational prefix of* o *with respect to* $\blacktriangle$*.*

If $\bigwedge U_m \equiv \top$ *then* $[\rho_R(o, \blacktriangle), \blacktriangle]$ *explains* o*.*

Proposition 2.50. *Given* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *and a core belief* $\blacktriangle \not\equiv \bot$ *let* $\rho_R(o, \blacktriangle) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0)$ *be the rational prefix of* o *with respect to* $\blacktriangle$*.*

If $\bigwedge U_m \not\equiv \top$ *then* $\blacktriangle$ *is not* o*-acceptable.*

So, one application of the rational prefix construction is for deciding whether there is some explanation using a particular core belief. Proposition 2.49 moreover says that in case ▲ is o-acceptable the epistemic state $[\rho_R(o, \blacktriangle), \blacktriangle]$ in fact explains o. That is, given *any* o-acceptable core, the rational prefix will yield a possible sequence that will complete the explanation for o. Proposition 2.51 tells us that using an inconsistent core belief ▲ the rational prefix for any observation is always $(\top)$.

Proposition 2.51. *Given* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *and a core belief* $\blacktriangle \equiv \bot$. *Then* $\rho_R(o, \blacktriangle) = (\top)$.

The following result tells us that from the point of view of computational complexity, checking the o-acceptability of a core belief is not harder than testing whether some epistemic state explains o. The proof yields that testing the o-acceptability of a core using the rational closure construction is computationally optimal.

Proposition 2.52. *The decision problem of whether a given core belief* ▲ *is* o*-acceptable is* Δ_2^P*-complete.*

We will now go on to argue that the rational prefix will also yield the *best* explanation for o — with respect to intuitive criteria — when using a particular o-acceptable core ▲. We consider an explanation better than another if it yields weaker beliefs. Assume that one explanation is telling us that the agent believes p, another that it believes $p \wedge q$ and yet another that the agent believes $p \wedge \neg q$. Further assume that the agent's true initial epistemic state is among those explanations. Which should we choose? We believe it to be reasonable to choose the first explanation. There is no guarantee that it is the correct one, but at least we are correct in concluding that the agent believes p. Conclusions that $\mathcal{A}$ believes q or $\neg q$ could well be wrong. It will turn out that — when fixing a core belief — the beliefs the rational prefix assigns to the agent's initial state are entailed by the beliefs assigned to that state by *any* explanation.

We will now make this more formal. The observation o to be explained is given and *some* o-acceptable core belief is fixed (the best core belief has been dealt with in Section 2.4). In the following definition we will capture our notion of one sequence being better than another. We will not only compare the agent's beliefs in the initial state but possibly continue with later beliefs if the two sequences yield equivalent initial beliefs. The sequences will be compared using the belief traces they give rise to. A belief trace is strictly preferred to another if the two contain equivalent beliefs up to some point and then the first one contains a strictly weaker belief. It is more cautious about the beliefs it predicts an agent to hold.

Definition 2.53. *Given an observation* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *and two possible belief traces* $(\beta_0, \ldots, \beta_n)$ *and* $(\gamma_0, \ldots, \gamma_n)$*, we define* $(\beta_0, \ldots, \beta_n) \leq_{\text{lex}} (\gamma_0, \ldots, \gamma_n)$ *iff, for all* $i = 0, \ldots, n$*,*

$$\beta_j \equiv \gamma_j \text{ for all } j < i \text{ implies } \gamma_i \vdash \beta_i.$$

Given two epistemic states $[\rho, \blacktriangle]$ *and* $[\sigma, \blacktriangle]$ *sharing the same core belief* $\blacktriangle$*,*

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

$\rho \preceq_2 \sigma$ *iff for all* λ*:* $Bel([\sigma, \blacktriangle] * \varphi_1 * \cdots * \varphi_n * \lambda) \vdash Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_n * \lambda)$*.*

$\preceq_1$ expresses giving preference to a sequence that yields weak beliefs early in the belief trace. ρ is preferred to σ if the initial beliefs $Bel([\rho, \blacktriangle])$ are logically weaker than $Bel([\sigma, \blacktriangle])$. In case they happen to be equivalent, ρ is preferred if the beliefs after having received the first revision input recorded in the observation are logically weaker than those yielded by σ. And so on. ρ and σ are considered equally preferred by $\preceq_1$ if both yield the same beliefs initially and throughout the sequence of revisions recorded in the observation.

$\preceq_2$ expresses giving preference to weak beliefs after any additional revision step. So ρ is preferred to σ if no matter which input is received after the ones recorded in the observation, ρ will yield a logically weaker (at most equivalent) belief than σ. As a sequence of revisions can be transformed into a single revision in the assumed framework, this is equivalent to ρ yielding a weaker belief after any sequence of further revision inputs.

Example 2.54. *Consider the observation* $o = \langle(p, q, \emptyset), (p \wedge \neg q, \top, \emptyset)\rangle$*.* $\top$ *is an* o*-acceptable core and* $\rho_1 = (q \wedge r)$*,* $\rho_2 = (q \wedge s)$*,* $\rho_3 = (p \wedge \neg q \wedge r, p \rightarrow q)$ *and* $\rho_4 = (p \rightarrow q)$ *are all sequences such that* $[\rho_i, \top]$ *explains* o*. The corresponding belief traces are:*

	$(Bel_0^{\rho_i}$,	$Bel_1^{\rho_i}$,	$Bel_2^{\rho_i})$
for ρ_1 :	$(q \wedge r$,	$p \wedge q \wedge r$,	$p \wedge \neg q)$
for ρ_2 :	$(q \wedge s$,	$p \wedge q \wedge s$,	$p \wedge \neg q)$
for ρ_3 :	$(p \rightarrow q$,	$p \wedge q$,	$p \wedge \neg q \wedge r)$
for ρ_4 :	$(p \rightarrow q$,	$p \wedge q$,	$p \wedge \neg q)$

ρ_1 and ρ_2 are incomparable with respect to $\preceq_1$ as neither $q \wedge r \vdash q \wedge s$ nor $q \wedge s \vdash q \wedge r$. Both are less preferred than ρ_3 and ρ_4 as $q \vdash p \rightarrow q$. $\rho_4 \preceq_1 \rho_3$ because the belief traces contain equivalent formulae except for the last element and $p \wedge \neg q \wedge r \vdash p \wedge \neg q$. In fact, ρ_4 is (equivalent to) the rational prefix for o with respect to the core belief $\top$.

The last two propositions of this section clarify the relation between the rational prefix and other possible sequences explaining the observation o assuming a common o-acceptable core belief. They provide the justification for considering the rational prefix to be the best sequence.

Proposition 2.55. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation and $\blacktriangle$ be an o-acceptable core.*

$$\textit{If } [\sigma, \blacktriangle] \textit{ explains } o \textit{ then } \rho_R(o, \blacktriangle) \preceq_1 \sigma.$$

With respect to $\preceq_1$, $\rho_R(o, \blacktriangle)$ is a best element among all sequences σ such that $[\sigma, \blacktriangle]$ explains the observation. Given that core belief, no other sequence yields weaker beliefs early in the belief trace. However, there will be several sequences yielding exactly the same beliefs throughout the belief trace.[12] Consequently, $\preceq_1$ will have several minimal elements. The next proposition tells us that with respect to $\preceq_2$ the rational prefix is a best one among those minimal elements. The rational prefix will lead to beliefs that are not logically stronger (compared to the beliefs yielded by any other $\preceq_1$-minimal sequence) no matter what further inputs may be received after those recorded in the given observation.

Proposition 2.56. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation and $\blacktriangle$ be an o-acceptable core.*

$$\textit{If } [\sigma, \blacktriangle] \textit{ explains } o \textit{ and } \sigma \preceq_1 \rho_R(o, \blacktriangle) \textit{ then } \rho_R(o, \blacktriangle) \preceq_2 \sigma.$$

Note that the criteria for comparing sequences assume the same core belief to be used. However, when trying to compare an arbitrary explanation for o with the agent's actual initial state, they might be useless as there is no guarantee that the assumed core belief and the actual one are equivalent. In case they are, the rational prefix will predict correct initial beliefs, i.e., every formula entailed by $Bel_0^{\rho_R(o,\blacktriangle)}$ will indeed be believed by the agent. But it will generally be the case that not all the agent's initial beliefs will be entailed. This means that Proposition 2.55 is no help in ensuring that predicted beliefs $Bel_i^{\rho_R(o,\blacktriangle)}$ from later points in the belief trace are really held by the agent. Similarly, Proposition 2.56 cannot be applied to guarantee that conclusions about beliefs held after future revision steps are safe. In this sense, the optimality criterion of the rational prefix being the $\preceq_2$-best among the $\preceq_1$-minimal sequences is quite weak. We will elaborate on this point in Section 2.8.

2.6 The rational explanation algorithm

We claim that for explaining an observation o the core belief $\blacktriangle_\vee(o)$ should be used. It is the weakest possible core belief. Any o-acceptable core entails it, i.e., no matter what the agent's real core belief is, any formula we predict to be rejected by the agent will indeed be rejected. No other o-acceptable core is safe in this sense. Further, no matter how the observation

[12] In fact, there are infinitely many sequences giving rise to the same total preorder on worlds and hence to equivalent epistemic states.

may be extended at both ends, conclusions about which formulae will be rejected need not be withdrawn (Left Monotony and Right Monotony). Assuming the agent to have $\blacktriangle_\vee(o)$ as core belief, we can construct conditional beliefs held in its initial epistemic state from o, obtaining a partial description of the rational consequence relation describing that state. We then use the rational prefix construction to calculate the sequence needed to complete the epistemic state that explains o.

This sequence has nice properties, as well. Compared to any other sequence ρ such that $[\rho, \blacktriangle_\vee(o)]$ explains o, it yields the weakest beliefs starting where the observation o started; and if the two sequences yield equal beliefs for the entire time of the observation, then $\rho_R(o, \blacktriangle_\vee(o))$ will yield equivalent or weaker beliefs for any further revision input. So, the message of the last two sections is: If o has an explanation then $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ is the best explanation for o.

Definition 2.57. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation s.t. there is an o-acceptable core belief. Then we call $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ the* rational explanation *for o.*

The problem is that in order to calculate the conditional beliefs of $\mathcal{A}$, we already have to know the core belief, so we have to know $\blacktriangle_\vee(o)$. Up to now, we have not given a feasible way to *construct* $\blacktriangle_\vee(o)$ — in order to take the disjunction of all o-acceptable cores, we would have to know those. Of course we could test any potential core belief (Propositions 2.49 and 2.50) but proceeding this way seems unreasonable. In this section we will introduce an algorithm that uses a more systematic way to calculate the rational explanation in case there is an o-acceptable core or tell us that o cannot be explained. In order to prove the correctness of this algorithm, some further results are useful. We start by recalling some notation and properties.

The rational closure was defined for arbitrary sets of positive and negative conditionals. However, in our setting all of them have a particular form due to their construction from the observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. Positive conditionals, i.e., those going into $\mathcal{C}_j$ in the rational prefix construction, have the form $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$. Negative conditionals, i.e., those going into $\mathcal{N}_j$, have the form $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ for some $\delta \in D_i$. The antecedents $f(\iota_i \cdot \blacktriangle)$ of both the positive and the negative conditionals are the same if they are constructed from the same prefix $\langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i, D_i)\rangle$ of o. Conditionals with an antecedent $f(\iota_i \cdot \blacktriangle)$ are called conditionals with index i.

In the rational prefix construction, conditionals may be exceptional for a set of formulae U. For positive conditionals $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ this is the case if $U \vdash \neg f(\iota_i \cdot \blacktriangle)$, for a negative conditional $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ if $U \cup \{f(\iota_i \cdot \blacktriangle)\} \vdash \delta$, so in particular if $U \vdash \neg f(\iota_i \cdot \blacktriangle)$. Now, in the rational prefix construction there are conditionals that satisfy a special property. The

conditionals in $\mathcal{C}_m$ and $\mathcal{N}_m$ are exceptional for *every* U_j including U_m. We will call them ultimately exceptional conditionals, as they would remain exceptional even if we continued constructing further levels $m + l$.

We showed in Proposition 2.50 that if ultimately exceptional conditionals exist, that is, $U_m \not\equiv \top$,[13] then the core belief used is not o-acceptable. For finding an o-acceptable core, we will iteratively strengthen it. We will show that if $\blacktriangle$, which is not o-acceptable, is not strengthened in a certain way then the resulting formula $\blacktriangle'$ cannot be o-acceptable, either. Intuitively this is because the conditionals with indexes that have not been dealt with by strengthening the core will remain ultimately exceptional.

Before we present the formal result we will give a necessary and sufficient condition for the existence of ultimately exceptional conditionals. Basically, it reflects the least fixpoint construction of the U_i in the calculation of the rational prefix.

Proposition 2.58. *Given an observation* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *and a core belief* $\blacktriangle$*. Ultimately exceptional conditionals in the rational prefix construction of* o *with respect to* $\blacktriangle$ *exist if and only if there are subsets* $I_\mathcal{C} \neq \emptyset$ *and* $I_\mathcal{N}$ *of* $\{1, \ldots, n\}$ *and a total order* $\prec_e$ *on* $I_\mathcal{N}$ *such that*

1. $\bigwedge\limits_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \ \wedge \bigwedge\limits_{j \in I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle) \vdash \bigwedge\limits_{i \in I_\mathcal{C}} \neg f(\iota_i \cdot \blacktriangle)$
2. $\forall j \in I_\mathcal{N} \exists \delta \in D_j : \bigwedge\limits_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \ \wedge \bigwedge\limits_{k \in I_\mathcal{N} \wedge k \prec_e j} \neg f(\iota_k \cdot \blacktriangle) \wedge f(\iota_j \cdot \blacktriangle) \vdash \delta$

Recall that if one conditional with index i is n-exceptional for U_j and hence $\neg f(\iota_i \cdot \blacktriangle)$ is added to U_j then automatically all conditionals with index i — positive and negative — become exceptional as then $U_j \vdash \neg f(\iota_i \cdot \blacktriangle)$ and $U_j \cup \{f(\iota_i \cdot \blacktriangle)\}$ entails any formula. Further, for every index $k \in \{1, \ldots, n\}$ there is a positive conditional. This is because in o there is a formula θ_k for every k. However, there need not be a negative conditional for every k as D_k may be empty. So if $I_\mathcal{C}$ and $I_\mathcal{N}$ denote the index sets of positive and negative ultimately exceptional conditionals, this implies $I_\mathcal{N} \subseteq I_\mathcal{C}$ and using the first condition in Proposition 2.58 in fact we get

$$\bigwedge_{i \in I_\mathcal{C}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \ \wedge \bigwedge_{j \in I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle) \equiv \bigwedge_{i \in I_\mathcal{C}} \neg f(\iota_i \cdot \blacktriangle)$$

This is because $\neg f(\iota_i \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i$ for all i and $\bigwedge\limits_{i \in I_\mathcal{C}} \neg f(\iota_i \cdot \blacktriangle) \vdash \bigwedge\limits_{j \in I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle)$ as $I_\mathcal{N} \subseteq I_\mathcal{C}$. Consequently, $\alpha_m = \bigwedge U_m$ turns out to be equivalent to $\bigwedge\limits_{i \in I_\mathcal{C}} \neg f(\iota_i \cdot \blacktriangle)$. Hence, if $\alpha_m \not\equiv \top$ then there must have been ultimately exceptional conditionals (assume there were

[13] An exception exists in case $\blacktriangle \equiv \bot$ (c.f. Proposition 2.51). We will return to this special case later.

none, which means there were neither positive nor negative exceptional conditionals, then $\bigwedge U_m = \bigwedge \emptyset \equiv \top$). The following proposition tells us that the only case where it looks as if there were none although $\mathcal{C}_m \neq \emptyset$ is when the core belief is inconsistent.

Proposition 2.59. *Let* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *be an observation and* $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$ *the rational prefix for* o *using some core* $\blacktriangle$*.*

Then $\alpha_m \equiv \top$ *and* $\mathcal{C}_m \neq \emptyset$ *implies* $\blacktriangle \equiv \bot$*.*

We now come to one of the central results of this section. It yields the key idea for the rational explanation algorithm whose power will lie in calculating $\blacktriangle_\vee(o)$. The rational prefix of o with respect to $\blacktriangle_\vee(o)$ will then complete the rational explanation which we propose to be the best explanation for o.

Proposition 2.60. *Let* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$*,* $\blacktriangle$ *a core belief and* $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$*.*

If $\blacktriangle' \vdash \blacktriangle$ *and* $\blacktriangle' \nvdash \alpha_m$ *then* $\blacktriangle'$ *cannot be* o*-acceptable.*

Assume that *some* formula $\blacktriangle$ was used to calculate the rational prefix $\rho_R(o, \blacktriangle)$ of o and $\blacktriangle'$ entails $\blacktriangle$. Then $\blacktriangle'$ *must* entail the weakest element α_m of $\rho_R(o, \blacktriangle)$ if it is to be o-acceptable. This tells us how to modify $\blacktriangle$ if the rational prefix construction was not successful ($\alpha_m \not\equiv \top$) and hence $\blacktriangle$ is not o-acceptable. We can try again with $\blacktriangle \wedge \alpha_m$. Propositions 2.49 and 2.51 tell us that $\alpha_m \equiv \top$ and hence $\blacktriangle \wedge \alpha_m \equiv \blacktriangle$ if $\blacktriangle$ is o-acceptable or $\blacktriangle$ is inconsistent. In both cases we can stop as we have found a suitable core or it cannot be strengthened further and is not acceptable by definition.

Obviously, $\blacktriangle_\vee(o)$ entails $\top$ and by Proposition 2.60 it must entail α_m of the rational prefix of o with respect to $\top$. If α_m is a tautology then we are done, otherwise $\blacktriangle_\vee(o)$ must entail the weakest element α'_m of the rational prefix of o with respect to α_m. Consequently, $\blacktriangle_\vee(o)$ entails $\alpha_m \wedge \alpha'_m$. And so on. With this idea we immediately get the following algorithm.

Algorithm 1: calculation of the rational explanation

Input: observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$

Output: the rational explanation for o

$\blacktriangle \Leftarrow \top$

repeat

  $\rho \Leftarrow \rho_R(o, \blacktriangle)$ `/* ` $\rho = (\alpha_m, \ldots, \alpha_0)$ `*/`

  $\blacktriangle \Leftarrow \blacktriangle \wedge \alpha_m$

until $\alpha_m \equiv \top$

return $[\rho, \blacktriangle]$ if $\blacktriangle \not\equiv \bot$, "no explanation" otherwise

The following two propositions ensure that the algorithm does what it is supposed to do — terminate and return the rational explanation if o can be explained.

Proposition 2.61. *For all observations o Algorithm 1 terminates.*

Proposition 2.62. *Given as input an observation o, Algorithm 1 outputs the rational explanation $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ for o, if an explanation for o exists. If no explanation exists it outputs "no explanation".*

Note that Proposition 2.60 does not restrict the core belief $\blacktriangle$. That is, we are not forced to start with $\top$, we could also use some other formula ψ. Any o-acceptable core belief $\blacktriangle'$ entailing ψ will have to entail the weakest element α_m of the rational prefix of o with respect to ψ. And if α_m is not a tautology then $\blacktriangle'$ will also have to entail the weakest element of the rational prefix of o with respect to $\psi \wedge \alpha_m$. Obviously, when we run the algorithm starting with ψ and not with $\top$, we will not necessarily get $\blacktriangle_\vee(o)$ as that core need not entail ψ and hence Proposition 2.60 is not applicable. However, we will get the weakest o-acceptable core belief entailing ψ or find out that there is no such core. In Section 2.8, we will see that this is a useful variation of the algorithm.

We want to conclude this section with a complex example illustrating the rational explanation algorithm and what its result lets us conclude about the observed agent.

Example 2.63. *Let $o = \langle (p, s, \emptyset), (q, \top, \{s\}), (r, \neg q, \emptyset)\rangle$. o expresses that after receiving p the agent believes s, but after then receiving q ceases to believe s. Finally, after receiving r it believes $\neg q$. Starting with $\blacktriangle = \top$ as Algorithm 1 tells us to, this translates into the following sets of positive and negative conditionals:*

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

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

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

This time $U_1 = \tilde{\mathcal{C}}_1 = \{p \wedge q \rightarrow \top, p \wedge q \wedge r \rightarrow \neg q\}$ as adding $p \wedge q$ does not make s inferable, anymore. Only $p \wedge q \wedge r \Rightarrow \neg q$ is exceptional for U_1. As indicated above, $p \wedge q \wedge r \Rightarrow \neg q$ is

in a sense exceptional for itself *because* $p \wedge q \wedge r$ *is inconsistent with* $p \wedge q \wedge r \rightarrow \neg q$. *So we have* $\mathcal{C}_2 = \{p \wedge q \wedge r \Rightarrow \neg q\} = \mathcal{C}_3$, $\mathcal{N}_2 = \emptyset = \mathcal{N}_3$ *and* $U_2 = \{p \wedge q \wedge r \rightarrow \neg q\} = U_3$. *Here, the rational prefix calculation stops as the sets do not change and we get*

$$\rho_R(o, \blacktriangle) = (p \wedge q \wedge r \rightarrow \neg q \ , \ p \wedge q \wedge r \rightarrow \neg q \ , \ (p \rightarrow s) \wedge (p \wedge q \wedge r \rightarrow \neg q) \wedge \neg(p \wedge q)).$$

Using logical equivalences this is the same as $(p \wedge q \rightarrow \neg r \ , \ p \wedge q \rightarrow \neg r \ , \ p \rightarrow (s \wedge \neg q))$. $\alpha_m = \alpha_2 = p \wedge q \wedge \neg r \equiv \neg p \vee \neg q \vee \neg r$ *is not a tautology, so we need to modify the core to* $\blacktriangle = \neg p \vee \neg q \vee \neg r$. *Hence, we get the following sets of conditionals for the second iteration.*

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

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

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

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

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

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

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

$\tilde{\mathcal{C}_0} = \{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r \rightarrow \top, p \wedge \neg q \wedge r \rightarrow \neg q\}$, *the last two implications being tautologies.* $p \wedge q \wedge \neg r \Rightarrow s$ *is n-exceptional for that set as* $\{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r\} \vdash s$. *Consequently,* $U_0 = \{p \wedge (\neg q \vee \neg r) \rightarrow s, p \wedge q \wedge \neg r \rightarrow \top, p \wedge \neg q \wedge r \rightarrow \neg q, \neg(p \wedge q \wedge \neg r)\}$.

$p \wedge q \wedge \neg r \Rightarrow \top$ is p-exceptional for U_0 as we had to add the negated antecedent of the only negative conditional. However, the other positive conditionals are not p-exceptional for U_0 so we get $\mathcal{C}_1 = \{p \wedge q \wedge \neg r \Rightarrow \top\}$, $\mathcal{N}_1 = \{p \wedge q \wedge \neg r \Rightarrow s\}$. $\tilde{\mathcal{C}}_1$ amounts to $\{\top\}$ and none of the remaining two conditionals is exceptional for that set. Hence we get $U_1 = \{\top\}$ and $\mathcal{C}_2 = \mathcal{C}_3 = \emptyset$, $\mathcal{N}_2 = \mathcal{N}_3 = \emptyset$ and $U_2 = U_3 = \emptyset$.

As a consequence, we get $\rho_R(o, \blacktriangle) = (\top, \top, (p \wedge (\neg q \vee \neg r) \rightarrow s) \wedge \neg(p \wedge q \wedge \neg r))$. As this means $\alpha_2 = \top$, no further iteration is necessary and the rational explanation for o is $[(\top, \top, (p \wedge (\neg q \vee \neg r) \rightarrow s) \wedge \neg(p \wedge q \wedge \neg r)), \neg q \vee \neg r]$.

$((\neg q \vee \neg r) \wedge (p \rightarrow s \wedge \neg q), p \wedge \neg q \wedge s, p \wedge q \wedge \neg r, p \wedge \neg q \wedge r \wedge s)$ is the belief trace according to this explanation. The core belief must entail $\neg q \vee \neg r$ as otherwise due to Proposition 2.26 $\mathcal{A}$ would have to believe q after receiving r but the observation tells us the opposite. Further, nothing indicates that $\blacktriangle$ would have to be stronger than that. Initially, the agent also believes $p \rightarrow s \wedge \neg q$. The belief in $p \rightarrow s$ is clear as after hearing p the agent believes s and p alone does not entail s. The belief in $p \rightarrow \neg q$ is more subtle and is best explained when looking at the beliefs after the first revision input was received. The beliefs in p and s are clear, but why should $\mathcal{A}$ commit to $\neg q$? If it did not then q would be consistent with the current beliefs and revision by q would turn out to be an expansion of the belief set. However, o tells us that $\mathcal{A}$ ceases to believe s, so it cannot be an expansion by q — this also explains the belief in $p \rightarrow \neg q$ in the initial state. The belief in r, $\neg q$ and p in the final state are quite intuitive; r has just been received, the observation requires $\neg q$ to be believed and there is no reason why $\mathcal{A}$ should suddenly reject p. s is believed again, as the apparent reason not to believe s, namely q, is gone.

Due to the nature of the calculation, the rational prefix in this example contains two tautologies. For deciding if an explanation was found, the first one is essential (Propositions 2.49 and 2.50). For calculating the belief trace etc. they are irrelevant and could be eliminated from the sequence (Proposition 2.8). In the following we will often omit the tautologies from the rational prefix. This will sometimes lead to explanations of the form $[(), \blacktriangle]$ which means that the rational prefix contained only tautologies.

In Appendix B, we will summarise some results concerning the computational complexity of problems related to observations and their explanations. At this point it suffices to say that Algorithm 1 is not efficient. The construction of the rational prefix needs a polynomial number of satisfiability tests. However, the main source for its complexity is the loop in which the core belief is strengthened until it is o-acceptable. There is a simple example showing that the algorithm may need an exponential number of iterations: $o = \langle (p_1, \top, \emptyset), \ldots, (p_n, \top, \emptyset), (p_{n+1}, \bigwedge_{1 \leq i \leq n} \neg p_i, \emptyset) \rangle$.

2.7 Observations of length one and two

Before we conclude this chapter, we want to illustrate the special case of observations in which only one or two revision inputs received by $\mathcal{A}$ are recorded. It would be unfair to say that such observations are always less informative than longer ones[14], but the information they can carry is inherently limited. This is because the (positive and negative) conditional beliefs they are translated into all have the same antecedent (at most two different ones if two revision inputs are recorded). Consequently they can only talk about a very restricted part of the rational consequence relation defined by the agent's initial state.

First consider $o = \langle(\varphi, \theta, \emptyset)\rangle$, assuming θ to be consistent. What can we say about an agent if the only information we have is that after receiving φ it believes θ? It must have believed $\varphi \rightarrow \theta$ before receiving the recorded input. If its initial beliefs are inconsistent with φ this is trivially the case. Otherwise, revision by φ (when considering only the belief set) is an expansion, that is, the original beliefs together with the new input are closed under logical consequence. If $\varphi \rightarrow \theta$ was not among the original beliefs, adding φ could not make θ inferable. In case $\varphi \wedge \theta$ is inconsistent, φ must have been rejected by $\mathcal{A}$, as otherwise θ could not be consistently believed. But then the revision cannot have had an effect on the beliefs and θ must have been believed initially.

If there is information about non-beliefs, i.e., $o = \langle(\varphi, \theta, D)\rangle$ with $D \neq \emptyset$, we may have more specific information about whether φ was accepted. As above $\varphi \wedge \theta$ must be consistent but further it must not entail any element of D. Whenever $\varphi \wedge \theta \vdash \delta$ for some $\delta \in D$, the revision input must have been rejected ($\blacktriangle \vdash \neg\varphi$). However, this does not guarantee that o has an explanation as possibly also $\neg\varphi \wedge \theta \vdash \delta'$ for some $\delta' \in D$. $o = \langle(p, p \leftrightarrow q, \{q, \neg q\})\rangle$ illustrates that. It is possible that an agent believes $p \leftrightarrow q$ while believing neither q nor $\neg q$ — it just needs to be agnostic about p as well. But once p is received as a revision input, $\mathcal{A}$ cannot be agnostic about it any longer (Proposition 2.20). So o cannot have an explanation. Also for the case that $D \neq \emptyset$ it still applies that $\varphi \rightarrow \theta$ must be believed initially and θ in case φ is rejected.

For convenience we require $\bot \in D$ for the following statement. This assumption only expresses that the agent's beliefs are to be consistent. This is the case whenever its core belief is consistent — the only case we are interested in. $[(\varphi \rightarrow \theta), \top]$ explains $o = \langle(\varphi, \theta, D)\rangle$ if and only if $\varphi \wedge \theta \nvdash \delta$ for all $\delta \in D$. $[(\theta), \neg\varphi]$ explains $o = \langle(\varphi, \theta, D)\rangle$ if and only if $\neg\varphi \wedge \theta \nvdash \delta$ for all $\delta \in D$. o has no explanation if those two do not work. The given explanations exactly mirror the conclusions illustrated above and the rational explanation is equivalent to one of the two, preferring the first one if it is an explanation.

[14] $\langle(p, \top, \emptyset), (p, \top, \emptyset), \dots\rangle$ is almost completely useless. It only tells us that the agent received an input p.

Example 2.64. *The following table contains a number of example observations of length one. The explaining epistemic state is equivalent to the rational explanation which is in turn equivalent to one of the two states we gave in the above statement. The belief trace indicates which beliefs we ascribe to the agent before the observation started and after it received the revision input. In the last four cases the agent cannot have accepted the input and hence its beliefs must have remained unchanged. For the last but one observation* $\langle(p, q, \{p \wedge q\})\rangle$*, e.g., we could also have used* $[q, \neg p]$ *as epistemic state. It is equivalent to the one given because* $\neg p \rightarrow q$ *and* q *are equivalent in the presence of the core belief* $\neg p$*.*

observation	*rational explanation*	*belief trace*
$\langle(p, p, \emptyset)\rangle$	$[(), \top]$	$(\top, p)$
$\langle(p, q, \emptyset)\rangle$	$[(p \rightarrow q), \top]$	$(p \rightarrow q, p \wedge q)$
$\langle(p, \neg p, \emptyset)\rangle$	$[(), \neg p]$	$(\neg p, \neg p)$
$\langle(p, \top, \{p\})\rangle$	$[(), \neg p]$	$(\neg p, \neg p)$
$\langle(p, q, \{p \wedge q\})\rangle$	$[(\neg p \rightarrow q), \neg p]$	$(\neg p \wedge q, \neg p \wedge q)$
$\langle(p, p \leftrightarrow q, \{q\})\rangle$	$[(p \leftrightarrow q), \neg p]$	$(\neg p \wedge \neg q, \neg p \wedge \neg q)$

If the observation has more than one recorded revision input it does not suffice to look at each record individually. The reason is once more the revision framework's property expressed in Proposition 2.20. Once an input φ was received, at any point in the future, $\mathcal{A}$ will believe either φ or its negation. This means that for later inputs it does not suffice to check whether $\varphi_i \wedge \theta_i \not\vdash \delta$ for all $\delta \in D_i$. $\bigwedge_{1 \leq j \leq i} (\neg)\varphi_j \wedge \theta_i \not\vdash \delta$ must be tested with respect to this criterion. In case it is violated, the core belief needs to make sure that this combination of (negated) revision inputs cannot be believed and hence needs to entail $\neg \bigwedge_{1 \leq j \leq i} (\neg)\varphi_j$. Such an adaptation of the core belief will in general have an impact on other entries of the observation.

For observations with two recorded revision inputs, we will only illustrate the case where no information about non-beliefs is given. That is, we consider observations of the form $o = \langle(\varphi_1, \theta_1, \emptyset), (\varphi_2, \theta_2, \emptyset)\rangle$. As in the above case, φ_i will have to be rejected ($\blacktriangle \vdash \neg\varphi_i$) if $\varphi_i \wedge \theta_i$ is inconsistent, but due to Proposition 2.20 we know that after receiving φ_2, the agent must also believe either φ_1 or $\neg\varphi_1$. So we can also test whether $\varphi_1 \wedge \varphi_2 \wedge \theta_2$ is consistent. If not, then $\mathcal{A}$ cannot commit to φ_1 and φ_2 at the same time which implies that the core belief must entail $\neg\varphi_1 \vee \neg\varphi_2$ (Proposition 2.26). This will have an impact on the first recorded input as well, as $\varphi_1 \wedge (\neg\varphi_1 \vee \neg\varphi_2)$ entails $\neg\varphi_2$. So in case $\varphi_1 \wedge \neg\varphi_2 \wedge \theta_1$ is inconsistent, φ_1 must have been rejected after all.

Example 2.65. *The next table contains some example observations of length two. For each one we give an epistemic state corresponding to the rational explanation and the belief trace*

assigned to the observed agent. In the last observation, the non-belief in $r \wedge s$ allows us to conclude that the last revision input r cannot have been accepted as we know that the agent believes s in the end. Hence, the beliefs before and after receiving r must be the same. As a consequence p cannot have been accepted either as the agent does not believe $p \wedge s$ after receiving r. The core belief must entail $\neg p \wedge \neg r$ and the beliefs remain unchanged throughout the observation. Note that $\neg p \wedge \neg r \rightarrow q \wedge s$ and $q \wedge s$ are equivalent in the presence of the core belief $\neg p \wedge \neg r$. Consequently, $(q \wedge s)$ could have been used as sequence in the explanation as well.

observation	*rational explanation* *belief trace*
$\langle (p, p, \emptyset), (r, s, \emptyset) \rangle$	$[(p \wedge r \rightarrow s), \top]$ $(p \wedge r \rightarrow s, p \wedge (r \rightarrow s), p \wedge r \wedge s)$
$\langle (p, \top, \{p\}), (r, s, \emptyset) \rangle$	$[(\neg p \wedge r \rightarrow s), \neg p]$ $(\neg p \wedge (r \rightarrow s), \neg p \wedge (r \rightarrow s), \neg p \wedge r \wedge s)$
$\langle (p, \top, \emptyset), (r, s, \{r \wedge p\}) \rangle$	$[(\neg p \wedge r \rightarrow s), \neg p \vee \neg r]$ $(\neg r \vee (\neg p \wedge s), p \wedge \neg r, \neg p \wedge r \wedge s)$
$\langle (p, p, \emptyset), (r, s, \{r \wedge p, r \wedge s\}) \rangle$	$[(p \wedge \neg r \rightarrow s), \neg r]$ $(\neg r \wedge (p \rightarrow s), p \wedge \neg r \wedge s, p \wedge \neg r \wedge s)$
$\langle (p, q, \emptyset), (r, s, \{p \wedge s, r \wedge s\}) \rangle$	$[(\neg p \wedge \neg r \rightarrow q \wedge s), \neg p \wedge \neg r]$ $(\neg p \wedge \neg r \wedge q \wedge s, \neg p \wedge \neg r \wedge q \wedge s, \neg p \wedge \neg r \wedge q \wedge s)$

2.8 Concluding remarks

In this chapter, we suggested a particular agent framework to model the observed agent — describing the epistemic state of $\mathcal{A}$, the way it revises its state and extracts its belief set. We formalised the observation o to be made of $\mathcal{A}$ and the conditions for a potential initial epistemic state $[\rho, \blacktriangle]$ to be an explanation for that observation, in which case the core belief $\blacktriangle$ is called o-acceptable. The assumed revision framework allows viewing an initial state as a rational consequence relation and transforming the observation into a partial description of that relation. Using the rational closure we are able to construct an explanation for o if one exists. Such an explanation allows us to draw conclusions about the agent. The core belief determines which revision inputs are predicted to be rejected by $\mathcal{A}$ and which are predicted to be accepted. The belief trace allows conclusions about the agent's beliefs before the observation o started, about what the agent believed apart from what is recorded in o and what it may believe after receiving further revision inputs.

Each observation could have been *generated* from a vast number of different initial epistemic states, i.e., agents may be very different from one another and may still give rise to the

same observation. This means that there cannot be an algorithm that will always return the right explanation. It will not be possible to guarantee the correct core belief, correct initial, final beliefs, etc. — compromises will have to be made. We suggested one which singles out one particular explanation and uses that for reasoning about $\mathcal{A}$. In the following, we will illustrate what the rational explanation can and cannot do and propose a method for identifying conclusions that have a particularly strong support.

One of our main results is the existence of an optimal o-acceptable core $\blacktriangle_\vee(o)$, one that is entailed by any o-acceptable core. This means $\blacktriangle_\vee(o)$ yields safe conclusions about the core beliefs of the observed agent in the sense that each revision input predicted to be rejected by the agent will indeed be rejected. Formulae predicted to be accepted may in fact be rejected by $\mathcal{A}$ as its core belief could be stronger than $\blacktriangle_\vee(o)$. We then went on to show that the rational closure yields a suitable way for calculating an optimal ρ given $\blacktriangle$. Here optimality is equated with yielding weak beliefs starting at the beginning of the observation. These results were accumulated into an algorithm that calculates the optimal core and the corresponding sequence for a given observation if an explaining initial epistemic state exists or tells us that no explanation exists otherwise.

As for many optimisation problems, e.g., scheduling or planning tasks, the quality of the solution and the conclusions we can draw from it depend heavily on the quality of the data and the validity of the assumptions made. Choosing a certain road based on an old map may turn out to be far from optimal if you suddenly find yourself at the wrong end of a one way street. Seven minutes for changing trains may suffice only under the assumption that the first train actually is on time. If this assumption turns out to be wrong, the conclusion that one will make the appointment might be useless if not harmful. 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 core belief $\blacktriangle_\vee$ depends on those. The optimality of ρ 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 2.66. *(i) Let* $o = \langle(\top, p, \emptyset), (\neg p, \top, \emptyset), (r \leftrightarrow \neg p, r \vee p, \emptyset)\rangle$*. o expresses that after receiving a tautology the agent believes p which is equivalent to enforcing an initial belief in p. The agent then receives the input $\neg p$, but we get no further information about beliefs or non-beliefs. Finally, after hearing that r holds if and only if $\neg p$ holds, the agent believes $r \vee p$. The rational explanation for o is* $[(p), \top]$ *and the corresponding belief trace is* $(p, p, \neg p, r \wedge \neg p)$*.*

If the agent's actual initial belief state was $[(), p]$*, i.e., in particular the core belief was stronger than concluded, the belief trace in truth is* $(p, p, p, \neg r \wedge p)$*. So except for the beliefs before*

the observation started and after receiving the tautology all conclusions we draw about $\mathcal{A}'s$ beliefs are wrong! Note also that in this example the two belief traces are incomparable with respect to $\leq_{\text{lex}}$. This relation, which we used for defining the optimality of the sequence in epistemic states, guarantees comparability only if the same core belief is used.

(ii) Let $o = \langle (p, p, \emptyset), (q, q, \emptyset), (r \leftrightarrow p, \top, \emptyset) \rangle$. All this observation obviously implies is that three revision inputs were received and that the first two are accepted by the agent. This lack of information is reflected in the rational explanation for o which is $[(), \top]$. The corresponding belief trace is $(\top, p, p \wedge q, p \wedge q \wedge r)$. If the agent's actual initial belief state was $[(), q \rightarrow \neg p]$, 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.

This strong dependence on the core belief can be easily explained. As mentioned before there are two main effects due to the core belief. First, it causes revision inputs to be rejected immediately in case they are inconsistent with the core. 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 — an effect which is related to Proposition 2.26 and 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. This becomes clear when taking a close look at the criterion given in Definition 2.53. Formulae entailed by the elements of the belief trace are guaranteed to be believed by the agent only up to one step beyond the point where calculated beliefs and the agent's actual ones first fail to be equivalent. This can easily be the case already for the initial beliefs. This shows that the criterion given in Definition 2.53 may be suitable to define a notion of optimality, but it is much too weak to ensure safe conclusions about beliefs throughout the belief trace.

Consider $o = \langle (p, q, \emptyset), (r, \top, \emptyset) \rangle$ for which the rational explanation is $[(p \rightarrow q), \top]$, the corresponding belief trace being $(p \rightarrow q, p \wedge q, p \wedge q \wedge r)$. 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 be wrong about q whose negation is in fact believed after having received the input r.

Is there anything we can do in order to verify conclusions we draw about $\mathcal{A}$? In other words, is there a way to turn the rather academic optimality criteria for explanations of

an observation into a tool for drawing *safe* conclusions? As mentioned before, formulae predicted to be rejected (using $\blacktriangle_\vee(o)$) must indeed have been rejected. But what about formulae consistent with $\blacktriangle_\vee(o)$ — are they necessarily accepted by $\mathcal{A}$? And what about beliefs and non-beliefs extracted from the belief trace which we calculated from the rational explanation of an observation? Can we rely on any of them? In the following we will suggest notions of safety and present a method we call *hypothetical reasoning* which allows to test whether a conclusion is safe.

2.8.1 hypothetical reasoning

We will start with a very strong notion of safety and later sketch weaker ones. Note that throughout this section we assume the observation o to have an explanation. We further assume implicitly that the rational explanation of o actually allows us to draw the conclusion we then test.

Definition 2.67. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation made of an agent $\mathcal{A}$. We call the following statements* safe *if and only if the given condition is satisfied:*

- *"$\mathcal{A}$ rejects the revision input φ": $\blacktriangle \vdash \neg\varphi$ for every o-acceptable $\blacktriangle$*
- *"$\mathcal{A}$ accepts the revision input φ": $\blacktriangle \nvdash \neg\varphi$ for every o-acceptable $\blacktriangle$*
- *"$\mathcal{A}$ believes θ after receiving the sequence of revision inputs $(\psi_1, \ldots, \psi_i)$ in its initial epistemic state": $Bel([\rho, \blacktriangle] * \psi_1 * \cdots * \psi_i) \vdash \theta$ for every explanation $[\rho, \blacktriangle]$ of o*
- *"$\mathcal{A}$ does not believe θ after receiving the sequence of revision inputs $(\psi_1, \ldots, \psi_i)$ in its initial epistemic state": $Bel([\rho, \blacktriangle] * \psi_1 * \cdots * \psi_i) \nvdash \theta$ for every explanation $[\rho, \blacktriangle]$ of o*

Note that these statements cover the answers to all the possible questions we collected at the end of Section 2.3 and also do not place any restrictions on φ, θ and the ψ_i. We call such a statement safe if and only if it is correct for any possible explanation of the observation and thus in particular for the agent's real initial epistemic state. The rational explanation can now be seen as one (but not as the exclusive) way of *generating* conjectures about $\mathcal{A}$ which then need to be verified. This verification is done via hypothetical reasoning by which we mean modifying the given observation o with respect to the conjecture and running the rational explanation construction on the modified observation o'. The modification is done in a way such that an explanation for o' will also be an explanation for o and would further be a counterexample to the conjecture. That is, finding an explanation for o' would prove the conjecture to be wrong, not finding one would prove it correct. Proposition 2.33 already

told us that any explanation for the observation $o_1 \cdot o_2 \cdot o_3$ can be adapted to explain the observation o_2. We will now give a proposition that expresses the same for a different way of extending an observation o_2.

Proposition 2.68.
Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and $o' = \langle(\varphi_1, \theta'_1, D'_1), \ldots, (\varphi_n, \theta'_n, D'_n)\rangle$ such that for all $1 \leq i \leq n$ we have $\theta'_i \vdash \theta_i$ and $D'_i \supseteq D_i$. If $[\rho, \blacktriangle]$ explains o' then it also explains o. In particular, any o'-acceptable core is o-acceptable.

o and o' contain the same sequence of revision inputs, the beliefs recorded in o' are all equivalent or logically stronger compared to those in o and all non-beliefs recorded in o' also appear in o. In other words, information about beliefs and non-beliefs in o' is more specific than in o. This proposition immediately follows from Definition 2.29 and entails that $\blacktriangle_\vee(o') \vdash \blacktriangle_\vee(o)$ as any o-acceptable core entails $\blacktriangle_\vee(o)$. So, extending the observation in the above sense can only make the minimal core belief stronger.[15] Together with Left Monotony and Right Monotony we get that we can take an observation o, strengthen the beliefs and non-beliefs in o according to Proposition 2.68, append further observations o_1 and o_3 to the front and the end to get an observation o'. And any explanation for o' will also explain o. This is what we will make use of for confirming or refuting conjectures we have about an agent.

If we want to check whether we can safely conclude that an agent rejects the revision input φ, we simply need to calculate the rational explanation of the observation o. If $\blacktriangle_\vee(o) \nvdash \neg\varphi$ then we already have a counterexample. If $\blacktriangle_\vee(o) \vdash \neg\varphi$ then we know that every o-acceptable core entails $\blacktriangle_\vee(o)$ and hence also $\neg\varphi$, in which case the conclusion is safe.

If we want to check whether we can safely conclude that an agent accepts the revision input φ, we can check whether $o' = o \cdot (\varphi, \neg\varphi, \emptyset)$ has an explanation. The appended observation forces $\blacktriangle_\vee(o')$ to entail $\neg\varphi$ and hence to reject that input. It does not have any other effect. If there is an o'-acceptable core $\blacktriangle$, the conclusion is not safe as $\blacktriangle$ is also o-acceptable (as argued above). If there is no o'-acceptable core then every o-acceptable core is consistent with φ (otherwise it would also be o'-acceptable which is easy to see) and hence the conclusion is safe. Of course, this presupposes that o can be explained at all.

Example 2.69. *Given $o = \langle(p, p, \emptyset), (q, q, \emptyset)\rangle$ we might wonder if the agent necessarily accepts a possible input $p \wedge q$. However, $[(), \neg p \vee \neg q]$ is an explanation for the observation*

[15] A similar property does not hold if beliefs and non-beliefs are modified otherwise. Consider $\langle(p, \neg p, \emptyset)\rangle$ or $\langle(p, \top, \{p\})\rangle$ which requires the core belief to entail $\neg p$. If these observations are modified to $\langle(p, \neg p \vee q, \emptyset)\rangle$ or $\langle(p, \top, \emptyset)\rangle$, i.e., beliefs or non-beliefs are in a sense retracted from the observation, then $\top$ becomes an acceptable core.

$o' = \langle (p, p, \emptyset), (q, q, \emptyset), (p \wedge q, \neg p \vee \neg q, \emptyset)\rangle$ and hence also for o. As a consequence the conclusion that $p \wedge q$ will be accepted is not safe.

Given $o = \langle (p, p, \emptyset), (q, q, \{\neg p\})\rangle$ the same conclusion is safe as there is no explanation for $o' = \langle (p, p, \emptyset), (q, q, \{\neg p\}), (p \wedge q, \neg p \vee \neg q, \emptyset)\rangle$. Consequently, every o-acceptable core is consistent with $p \wedge q$.

This covers the first two cases from Definition 2.67. The last two cases are slightly more complicated as they do not require ψ_i to coincide with the corresponding φ_i from the observation. We will first deal with the simple cases where $(\psi_1, \ldots, \psi_i)$ is a prefix of $\iota = (\varphi_1, \ldots, \varphi_n)$ or vice versa. If $(\psi_1, \ldots, \psi_i)$ is a prefix of ι then the question is whether θ must/cannot be believed after receiving the i^{th} revision input. We can check whether θ must necessarily be believed after receiving the first i inputs by adding θ to D_i to get o'. We thus get an observation expressing the same as o but additionally that θ is not among the beliefs at that point during the observation. If there is an o'-acceptable core then it is possible that θ is not believed and hence that conclusion was not safe. If there is no o'-acceptable core then θ must necessarily be believed at that point.

We can check whether θ cannot be among the beliefs after receiving the first i inputs by strengthening θ_i in o to $\theta_i \wedge \theta$ to get o'. The observation thus obtained contains the same information but additionally that θ is among the beliefs at that point during the observation. If there is an o'-acceptable core then it is possible that θ is believed and hence that conclusion was not safe. If there is no o'-acceptable core then θ really cannot be believed at that point, as otherwise an explanation would have been found.

If ι is a prefix of $(\psi_1, \ldots, \psi_i)$, i.e. the latter sequence corresponds to $(\varphi_1, \ldots, \varphi_n, \psi_{n+1}, \ldots, \psi_i)$, the beliefs after some *further* revision inputs are considered. Here we first extend o to $o \cdot (\psi_{n+1}, \top, \emptyset), \ldots, (\psi_i, \top, \emptyset)$. From the definition of an explanation it is easy to see that an epistemic state explains either both or none of the two. Now θ can be added to $D_i = \emptyset$ or $\theta_i = \top$ as above in order to see whether θ must or cannot be believed after the agent's receiving the sequence of revision inputs $(\psi_1, \ldots, \psi_i)$ in its initial epistemic state.

Example 2.70. *Consider $o = \langle (p, p, \emptyset), (q, q, \{\neg p, \neg r\})\rangle$, that is, $\mathcal{A}$ accepts both inputs and does not believe $\neg p$ or $\neg r$ after receiving q. The rational explanation $[(), \top]$ may lead us to the conclusion that after receiving p the agent did not believe $\neg q$, but is this safe? We test whether there is an explanation for $o' = \langle (p, p \wedge \neg q, \emptyset), (q, q, \{\neg p, \neg r\})\rangle$ and indeed $[(p \rightarrow \neg q), \top]$ explains o', so not all explanations lead to this conclusion and it is not safe.*

The rational explanation predicts that after a further input r, p will still be believed. We can verify this by testing whether $o' = \langle (p, p, \emptyset), (q, q, \{\neg p, \neg r\}), (r, \top, \{p\})\rangle$ has an explanation. However, there is no o'-acceptable core, so this conclusion is indeed safe.

For the special case that $i = 0$, i.e., the agents initial beliefs are to be considered, the original observation must be modified in a slightly different way. Instead of appending new records to the end, $(\varphi_0, \theta_0, D_0) = (\top, \top, \emptyset)$ is appended to the front of o. Receiving a tautology leaves everything unchanged (Proposition 2.8). Thus we get an observation containing an entry for the initial beliefs and can modify θ_0 and D_0 as described above in order to verify conjectures about beliefs and non-beliefs of the agent before the actual observation started.

If $(\psi_1, \ldots, \psi_i)$ and $(\varphi_1, \ldots, \varphi_n)$ are not in any prefix relation as assumed up to this point, this means we are interested in the agent's beliefs after some sequence of revision inputs completely different from that recorded in o — but starting in the same initial epistemic state. Modifying o in a way described above is now impossible, intuitively because assumptions made about the observation would be violated. Instead, we will consider two observations: o and $o' = \langle(\psi_1, \top, \emptyset), \ldots, (\psi_i, \top, \emptyset)\rangle$. Note that any epistemic state explains that particular o'. The trick is now to modify o' as described above, that is, replace in o' the entry $(\psi_i, \top, \emptyset)$ by $(\psi_i, \theta, \emptyset)$ or $(\psi_i, \top, \{\theta\})$ and test whether there is an epistemic state explaining both o *and* the modified o'. Up to now we have not described how it is possible to find a single explanation for two observations. We will illustrate this in Section 4.2. At this point, it suffices to know that it can be done.

As mentioned at the beginning of this section, the notion of safety where all explanations have to verify the conclusion is a very strong one. By slightly modifying the conditions in Definition 2.67 we can get weaker notions of safety. Instead of requiring the conjecture to be verified by *every* explaining epistemic state, it might be enough that it is true for every explanation $[\rho, \blacktriangle]$ where (i) $\blacktriangle$ entails some specific formula ψ, (ii) $\blacktriangle$ does not entail some formula ψ, or (iii) $\blacktriangle$ is equivalent to some ψ. These variants make sense only if we are sure that the additional condition on the core belief is indeed met by the observed agent's real core belief.[16] We want to emphasise that *being* a conclusion is a necessary condition for a conclusion to be safe. In particular, there must be an explanation for the original observation such that the core belief satisfies the given property (entailing or not entailing a given formula).

For (i) we only need to initialise the core belief in the rational explanation construction with ψ rather than $\top$. Proposition 2.60 and and a slight variant of the proof for Proposition 2.62 ensure that then the weakest acceptable core which entails ψ is calculated. o' is constructed as above — in a way such that an explanation would be a counterexample to the conjecture.

[16] In the next section we will argue how very unlikely it is to construct the agent's real core belief from a given observation. But there are scenarios where such information may indeed be available. For example, some software agent from the same series and consequently distributed with the same initial configuration could have been observed accepting or not accepting some revision input.

If there is an o'-acceptable core entailing ψ, the conjecture is wrong, otherwise it must be correct.

For (ii) we calculate the rational explanation $[\rho, \blacktriangle]$ for the modified observation o'. If $\blacktriangle \nvdash \psi$, then we have a counterexample and the conjecture was wrong. Otherwise, any o'-acceptable core must necessarily entail ψ (or o' has no explanation at all). For (iii) we only need to check whether ψ is o'-acceptable. If so we have a counterexample, if not the conjecture must be correct.

Example 2.71. *(i) Consider $o = \langle(\neg q, \neg q, \emptyset)\rangle$ and assume we know the agent's core belief must entail $p \to q$. The conjecture we want to verify is that $\mathcal{A}$ believes $\neg p$ after receiving $\neg q$. We run the adapted rational explanation algorithm (initialising the core with $p \to q$) on $o' = \langle(\neg q, \neg q, \{\neg p\})\rangle$. However, we are then informed that this observation does not have an explanation. This means that given that the core belief entails $p \to q$, any explanation predicts a belief in $\neg p$. In this example this can easily be checked. As the recorded belief in $\neg q$ is correct and the core belief is held at every point in time, $\neg p$ is entailed by the two.*

(ii) Consider $o = \langle(p, \top, \emptyset), (q, \top, \{\neg r\}), (r, \top, \emptyset)\rangle$ and assume we know that $\mathcal{A}$'s core belief does not entail $\neg r \vee \neg p$. The rational explanation $[(), \top]$ suggests that after receiving r, the agent does not believe $\neg p$. We test this by calculating the rational explanation for $o' = \langle(p, \top, \emptyset), (q, \top, \{\neg r\}), (r, \neg p, \emptyset)\rangle$ which is $[(), \neg q \vee \neg p]$. So there is indeed an explanation for o whose core belief does not entail $\neg r \vee \neg p$ and which does in fact predict a belief in $\neg p$ after the agent's receiving r. Consequently, our conjecture was wrong.

Leaving the question of safe conclusions, we want to remark two things. Firstly, the rational explanation algorithm can be seen as applying hypothetical reasoning itself. It starts by conjecturing the core belief of the agent to be a tautology and checks whether this leads to a contradiction. If so, the core belief is adapted accordingly, i.e., a new conjecture is formed and tested. Secondly, Propositions 2.33 and 2.68 also allow for some algorithmic optimisation when extending a given observation o in the specified ways.[17] If the rational explanation for o has already been calculated, we do not need to start from scratch for o'. We can start by initialising the core belief with $\blacktriangle_\vee(o)$ rather than $\top$. The correctness of this method follows from Proposition 2.60 using a slight variant of the proof of Proposition 2.62.

For the case that the observation is not extended by appending further entries to the front of o, we need not even recalculate all the positive and negative conditionals in the first run

[17]This can be useful for hypothetical reasoning, but also if *additional* information about the agent is received and the observation can thus be refined. Additional here means that the information in the original observation is correct but some beliefs turned out to be logically stronger than initially given, we are informed about further formulae not being believed at some points during the observation, or earlier or later revision inputs become known. The assumption that no input was received between the ones recorded is still essential!

(provided they are still at our disposal from the rational explanation calculation for o). As the core belief ($\blacktriangle_\vee(o)$) is the same and the revision inputs were not changed, the antecedents of the conditional beliefs will be the same as well. So if a belief was modified, i.e., $\theta_i' \not\equiv \theta_i$, we simply use $f(\iota_i \cdot \blacktriangle_\vee(o)) \Rightarrow \theta_i'$ instead of $f(\iota_i \cdot \blacktriangle_\vee(o)) \Rightarrow \theta_i$. As we are dealing with rational consequence relations it does not matter whether we use both $\lambda \Rightarrow \mu_1$ and $\lambda \Rightarrow \mu_2$ or $\lambda \Rightarrow \mu_1 \wedge \mu_2$. So, if we want to express that after receiving φ_i, the agent believed ψ in addition to θ_i, we could also simply *add* the positive conditional $f(\iota_i \cdot \blacktriangle_\vee(o)) \Rightarrow \psi$. If a non-belief δ was added to D_i, we just add $f(\iota_i \cdot \blacktriangle_\vee(o)) \Rightarrow \delta$ to the negative conditionals.

2.8.2 impossibility results

It may be nice to be able to verify conclusions drawn from the rational explanation (or conjectures obtained otherwise), 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. Even if we know that the agent's core belief comprises only variables that appear in the observation o, there usually is more than one o-acceptable core.

In a different context which we will deal with in the next chapter, Sébastien Konieczny[18] suggested the additional assumption that the last belief θ_n recorded in the observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ is in fact *complete*, i.e., the agent's beliefs at that point are $Cn(\theta_n)$. This assumption gives us an upper bound on the actual core belief $\blacktriangle$ as then $\theta_n \vdash \blacktriangle$ must hold.

We can use the methodology described above in order to narrow down the actual core belief. Starting off with calculating the rational explanation $[\rho, \blacktriangle]$ of o, we check which beliefs this implies after receiving the last input. These are represented by $f(\rho \cdot \iota \cdot \blacktriangle)$. Clearly, this formula entails θ_n, which completely characterises the agent's beliefs at that point, but if the two are not equivalent then $f(\rho \cdot \iota \cdot \blacktriangle)$ must be logically stronger and is *not* believed by the agent. So, we add $f(\rho \cdot \iota \cdot \blacktriangle)$ to D_n and rerun the rational explanation algorithm. And so on, until recorded beliefs and those implied by the explanation coincide. This will give an improved lower bound for the core belief of the agent, but it still cannot guarantee uniqueness of the core.

Consider the observation $o = \langle(p, q, \emptyset), (p \wedge r, s, \emptyset), (\neg p, \neg p, \emptyset)\rangle$. $\blacktriangle_\vee(o) = \top$ but, e.g., $p \rightarrow q$ or $p \wedge r \rightarrow \neg q$ are also o-acceptable. Note that $\neg p$ entails these formulae. In this example, there is no information gain by the assumption that $\neg p$ is the only formula believed after

[18] Personal communication.

receiving it. It is inconsistent with all the other recorded revision inputs, so it cannot help in determining which further interactions the real core belief causes among those earlier inputs.

Going yet a step further, 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 $o = \langle (p, p, \emptyset), (q, p \wedge q, \emptyset), (r, p \wedge q \wedge r, \emptyset) \rangle$ for which $[(), \top]$ is the rational explanation. However, p is also o-acceptable, so the conclusion that an input $\neg p$ will be accepted by the agent may turn out to be wrong. All this should not come as a surprise: as long as there are several o-acceptable cores, we cannot be sure that $\blacktriangle_\vee(o)$ — or any other core found by some alternative method — is the right one.

We have assumed finite observations so far, in fact we defined them to be finite. We cannot deal with an infinite one as it might give rise to infinite sets of conditionals so that the rational prefix construction might not terminate. But we can investigate the convergence behaviour of the rational explanations for observations $o_i = \langle (\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i, D_i) \rangle$ from a sequence of prefixes $(o_1, o_2, \ldots)$ of an infinite observation $o = \langle (\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n), \ldots \rangle$. Note that the D_i are still assumed to be finite. o might be the result of continuous observation of an agent — reasoning about $\mathcal{A}$ while further extending o.

We already stated that the beliefs of the agent need not converge, but the question is whether the rational explanation converges towards the actual initial epistemic state $[\rho, \blacktriangle]$ of the agent, yielding the correct core belief and a sequence behaving exactly like ρ. By Right Monotony we know that $\blacktriangle_\vee(o_i) \vdash \blacktriangle_\vee(o_{i-1})$ for all i which means that using the rational explanation construction we get closer and closer to the core that gave rise to o. However, it is not clear whether $\blacktriangle$ will have been found at some point. And even if it has, is there an i such that $\blacktriangle_\vee(o_i) \equiv \blacktriangle$ and the observer also *knows* that the right core belief has been found? If the language we are dealing with is finite and fixed then there is hope, otherwise it might happen that a unique core has been found, i.e., hypothetical reasoning tells us that $\blacktriangle_\vee(o_i)$ cannot be strengthened using any formula made up of variables that have appeared so far, and o_{i+1} contains a new variable, again enlarging the set of o_{i+1}-acceptable cores.

But even if the language is fixed, finite and known, an infinite observation is not guaranteed to narrow down the set of acceptable cores. It is possible that from some point on the observation is not informative any longer, that is, it gives rise to positive and negative conditional beliefs that are already entailed by $\Rightarrow_{[\rho_R(o_i, \blacktriangle_\vee(o_i)), \blacktriangle_\vee(o_i)]}$ of some prefix o_i of the observation o and these do not suffice to eliminate all but one core belief. As a trivial example consider $\langle (p, p, \emptyset), (\neg p, \neg p, \emptyset), (p, p, \emptyset), \ldots \rangle$ but the language containing another variable q. As that one never appears in the infinite observation, we cannot be sure that the agent's core belief does not talk about q. For example, it may very well be q or $p \leftrightarrow q$.

We can guarantee to find the correct core belief if the language is fixed, finite and known and for every i such that $\blacktriangle_\vee(o_i)$ is not the only o_i-acceptable core, there is a j such that $\blacktriangle_\vee(o_i) \nvdash \blacktriangle_\vee(o_{i+j})$. This trivial condition simply states that as long as we have not found a unique core (which can in principle be checked as there are only finitely many), we will eventually get further information — in the form of conditional beliefs not known up to that point — which further narrows down the weakest core belief. Determining whether the rational prefix then correctly represents the sequence ρ in the agent's actual initial epistemic state $[\rho, \blacktriangle]$ coincides with the question whether the observation allows only a single rational consequence relation. Hypothetical reasoning is a possible tool for answering that question.

Chapter 3

Beyond the Rational Explanation

3.1 Introductory notes

In the last chapter we developed an algorithm that constructs a best potential initial epistemic state that explains an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ on an agent's revision history. We assumed the agent to be employing a particular framework for carrying out revision and calculating its beliefs. Having calculated an initial epistemic state, we are able to draw conclusions about beliefs other than those recorded in the observation as well as future beliefs. This is possible because in this framework the agent's future beliefs after a sequence of revisions are completely determined by its initial state. The current chapter is devoted to the application of the rational explanation in slightly more general settings.

One essential assumption for the results presented in Chapter 2 to hold was that the observation was correct and complete in the sense that exactly the inputs φ_i recorded were received by the agent. So the agent did indeed receive the input φ_i and it did not receive any inputs between φ_i and φ_{i+1}. This is a very strong assumption as it basically claims that the observer kept its eye on the agent at all times (during the observation) and was capable of registering the exact revision inputs received by the agent.

In some cases this assumption might even be justified. If, for example, the observer itself is the sole source of the revision inputs, it is in control of what inputs the agent $\mathcal{A}$ receives. However, if the communication channel is noisy, it still might not be sure about the logical content actually received by $\mathcal{A}$. The following example illustrates a case where the logical content of an input is not understood but where reasoning about the observed agent should still be possible.

Example 3.1. *Imagine, we observe a dialogue where an agent $\mathcal{A}$ receives exactly two inputs. We cannot understand the first input. However, the agent explicitly acknowledges that input,*

from which we can infer that it is believed by $\mathcal{A}$. After receiving a further input p the agent believes q and the negation of the first input — we can understand logical connectives and recognise the first input.

This cannot be dealt with using previous results as there is an unknown revision input. However, we should be able to infer that $\mathcal{A}$ believed $\neg p$ after receiving the first input. 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 all have to survive the revision process (cf. the "Vacuity" or preservation postulate from AGM) which is clearly not the case here.

There are many factors preventing the available information from meeting the strict requirements of the last chapter. Agents may leave the scene for a while, use a language not completely known to the observer, or exchange secret messages. The observing agent may be temporarily distracted, otherwise prevented from continuously observing, or incapable of precisely recording every piece of available information.

In this chapter, we want to investigate what we can say about $\mathcal{A}$ if the assumption of knowing the exact revision inputs is relaxed. We will consider two related but distinct scenarios. First, we will investigate the case where the logical content of the revision inputs received is only partially known,[1] i.e., the observation is still assumed to contain an entry $\langle(\varphi, \theta, D)\rangle$ for every input received. Then we will go on to the case where not all revision inputs are recorded in the observation. It should be clear that the less complete and reliable the information in the observation is, the less informative the conclusions about the agent will be. In fact, we will provide formal results only for the core belief that can be assigned to the agent. For reasoning about the beliefs and non-beliefs during the time of observation we propose to use hypothetical reasoning.

3.2 Dealing with unknown logical content

3.2.1 modelling unknown logical content

Throughout Section 3.2, we will keep the assumption that the observation contains an entry for every single revision input received by the agent. However, the information about the exact input may be partial. We will model this by allowing formulae appearing in the

[1] In fact, we allow this to be the case for all components of an observation: revision inputs, beliefs and non-beliefs. Note that such a scenario has similarities to the situation of a non-native speaker. She might not know the meaning of all words but has a good knowledge of the grammar of the language. This allows for extracting information from the context of unknown words.

observation to contain unknown subformulae which are represented by n placeholders χ_j. $\lambda(\chi_1, \ldots, \chi_n)[(\chi_i/\phi)_i]$ denotes the result of replacing in λ every occurrence of χ_i by ϕ_i.

Definition 3.2. *Let L be a propositional language and $\chi_1, \ldots, \chi_n$ be placeholders not belonging to L.*

An object $\lambda(\chi_1, \ldots, \chi_n)$ 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. So in order to model (even more) partial knowledge, formally an observation does not really contain formulae φ_i, θ_i and $\delta \in D_i$ but parametrised formulae $\varphi_i(\chi_1, \ldots, \chi_n)$, $\theta_i(\chi_1, \ldots, \chi_n)$ and $\delta(\chi_1, \ldots, \chi_n)$ that become formulae if all χ_j are properly instantiated. We will often write λ rather than $\lambda(\chi_1, \ldots, \chi_n)$ to denote a parametrised formula in order to ease reading. Hence, we deal with parametrised observations rather than observations like in the last chapter. Note that this is more general than having partial knowledge about the inputs only. We allow unknown subformulae in the beliefs and non-beliefs of the agent as well.

The parametrised observation corresponding to Example 3.1 is $\langle(\chi, \chi, \emptyset), (p, q \wedge \neg\chi, \emptyset)\rangle$. We want to emphasise that this is not a normal observation as we do not know which formula is represented by χ. The example already illustrates how we can express formally that the logical content of a revision input is completely unknown. In this case the corresponding entry for the revision input in the observation simply is a placeholder χ_j. It also shows that unknown subformulae can be referred to several times by using them more than once in the observation. This makes sense, for example, in scenarios where parts of the language are unknown to the observing agent. The meaning of an utterance might not be understood, but one might be able to recognise it when it appears again later.

Let o be a parametrised observation, i.e., o contains parametrised formulae that contain the unknown subformulae χ_i, $1 \leq i \leq n$. $o[\chi_1/\phi_1, \ldots, \chi_n/\phi_n]$ and equivalently $o[(\chi_i/\phi_i)_i]$ denote the observation o with every occurrence of the placeholder χ_i substituted by a formula ϕ_i. 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. We will sometimes need to refer to two observations constructed from one parametrised observation o where all unknown

subformulae except for χ_i were substituted by the same formula. The first being $o[(\chi_i/\phi_i)_i]$ the second may be denoted by $o[\chi_i/\phi'_i, (\chi_j/\phi_j)_{j\neq i}]$.

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 the last chapter. 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 $L(o)$. Generally, this should not be a conceptual restriction as we can replace an unknown subformula about which we have some further knowledge by an object encoding that knowledge. The following example is to illustrate this point.

Example 3.3. *Imagine, we want to express that the first three revision inputs are unknown but we know that the third one entails the first two. We represent the first input by χ_1 and the second one by χ_2. However, we use $\chi_1 \wedge \chi_2 \wedge \chi_3$ for representing the third input (rather than just χ_3). $\chi_1 \wedge \chi_2$ makes sure that the first two inputs are entailed, χ_3 is the placeholder for whatever else might be entailed by the third input. Also, whenever we need to refer to the third input, we use $\chi_1 \wedge \chi_2 \wedge \chi_3$.*

Proposition 3.4. *Let L be a propositional language and $\phi \in L$. Let I be a set of natural numbers, for each $i \in I$ let $\lambda_i(\chi)$ be a parametrised formula based on L, $\alpha_i = \lambda_i(\chi)[\chi/\phi]$, and $\alpha'_i = \lambda_i(\chi)[\chi/x]$ where $x \notin L$, i.e., the propositional variable x is not contained in any $\lambda_i(\chi)$ or ϕ. Then for all finite $S \subseteq I$*

$$\bigwedge_{i\in S} \alpha_i \vdash \bot \text{ if and only if } (x \leftrightarrow \phi) \wedge \bigwedge_{i\in S} \alpha'_i \vdash \bot.$$

This proposition expresses that, provided x and ϕ are assigned the same truth value, i.e., $x \leftrightarrow \phi$ is satisfied, it does not matter whether the unknown subformula is replaced by ϕ or the new variable x. Of course, this is not a deep result, but it contains the idea of how to deal with the unknown subformulae contained in the observation: We replace the unknown subformulae by new variables, i.e., variables that do not yet appear in the parametrised observation. That way, we get an observation that can be handled using the rational explanation construction.

3.2.2 finding an acceptable core belief

In the following sections, we will present results on what can be said about an agent's core belief given a parametrised observation o. Once again, if an explanation exists at all, we

can give a unique weakest core belief which is entailed by any acceptable core belief ▲. 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 ▲. 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.

The next proposition formalises that given there is *some* instantiation for the unknown subformulae such that the resulting observation has an explanation, we can also replace one of them by a new variable and still know that there is an explanation. This immediately generalises to all unknown subformulae in Proposition 3.6 — with an important implication: If some explanation exists for o, i.e., there is some instantiation of the χ_i such that the resulting observation can be explained, then we can also assume all the χ_i to be new variables and an explanation still exists. Consequently, we can start off with instantiating all χ_i by new variables x_i and use the rational explanation construction. If this fails, i.e., we are returned an inconsistent core belief, then no explanation can exist using any instantiation of the unknown subformulae for that observation.

Proposition 3.5. *If* $[\rho, ▲]$ *explains* $o[(\chi_i/\phi_i)_i]$ *and* x *is a propositional variable not appearing in* o, ▲, ρ *or any* ϕ_i *then* $[\rho, ▲ \wedge (x \leftrightarrow \phi_i)]$ *explains* $o[\chi_i/x, (\chi_j/\phi_j)_{j\neq i}]$.

Proposition 3.6. *If* $[\rho, ▲]$ *explains* $o[(\chi_i/\phi_i)_i]$ *and* $x_1, \ldots, x_n$ *are propositional variables not appearing in* o, ▲, ρ *or any* ϕ_i *then* $\left[\rho, ▲ \wedge \bigwedge_{1\leq i\leq n} (x_i \leftrightarrow \phi_i)\right]$ *explains* $o[(\chi_i/x_i)_i]$.

As can be seen from these propositions, an epistemic state explains a parametrised observation with respect to a given instantiation of the unknown subformulae, i.e., in a sense the instantiation is part of the explanation. So, when comparing explanations in terms of a preference relation, this cannot be done (completely) independently of what we assume the unknown subformulae to be. We will study this problem more deeply in Section 3.2.5.

Example 3.7. *(i) Consider the parametrised observation* $o = \langle(\chi, \chi, \emptyset), (p, q \wedge \neg\chi, \emptyset)\rangle$ *capturing Example 3.1. Instantiating* χ *with* $\neg q$ *we get* $o[\chi/\neg q] = \langle(\neg q, \neg q, \emptyset), (p, q, \emptyset)\rangle$. $[(), p \rightarrow q]$ *explains* $o[\chi/\neg q]$. *The corresponding belief trace is* $(p \rightarrow q, \neg q \wedge \neg p, p \wedge q)$.

The observation constructed according to Proposition 3.5, where χ *is replaced by a new variable* x, *is* $o[\chi/x] = \langle(x, x, \emptyset), (p, q \wedge \neg x, \emptyset)\rangle$. $[(), (p \rightarrow q) \wedge (x \leftrightarrow \neg q)]$ *explains* $o[\chi/x]$, *the corresponding belief trace being* $((p \rightarrow q) \wedge (x \leftrightarrow \neg q), x \wedge \neg q \wedge \neg p, p \wedge q \wedge \neg x)$.

The rational explanation for $o[\chi/x]$ *is* $[(p \wedge \neg x \rightarrow q), p \rightarrow \neg x]$, *the corresponding belief trace being* $(p \rightarrow (\neg x \wedge q), x \wedge \neg p, p \wedge q \wedge \neg x)$.

(ii) Note that all the belief traces in the first part of this example indicate 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 $o' = \langle(\chi, \chi, \{\neg p\}), (p, q \wedge \neg\chi, \emptyset)\rangle$. According to the hypothetical reasoning methodology, $\neg p$ was added to the non-beliefs. Applying the rational explanation algorithm to $o'[\chi/x]$ yields that there is no explanation. Proposition 3.6 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. We were able to draw the conclusion we indicated in Example 3.1.

As already illustrated by the example, Proposition 3.6 does not claim to give the *weakest* possible core for $o[(\chi_i/x_i)_i]$. It just proves there is an $o[(\chi_i/x_i)_i]$-acceptable core in case there is one for some instantiation of the χ_i. By Proposition 2.62, we can use the rational explanation algorithm to get the best explanation for $o[(\chi_i/x_i)_i]$. The core belief calculated will in general be weaker than the one given in Proposition 3.6. But it 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 of the unknown subformulae.

3.2.3 finding the best core belief

Up to now we have not assumed any particular (normal) form for the formulae we are dealing with. The unknown subformulae were allowed to appear arbitrarily deep in the parametrised formulae, but for some of the proofs it will be helpful to assume the formulae to have a particular structure. Obviously, any formula can be transformed into conjunctive normal form (CNF). So without loss of generality we can assume that the additional variables x_i appear as literals in the clauses of a formula in CNF that is equivalent to $\lambda[(\chi_i/x_i)_i]$. The following result simply confirms that if the χ_i are replaced by $x_i \wedge \psi$ instead of x_i for some arbitrary ψ, logically it does not matter whether this substitution is carried out on λ itself or whether the x_i are replaced in the corresponding CNF, the result not being in CNF, of course.

Proposition 3.8. *Let $\lambda(\chi_1, \ldots, \chi_n)$ be a parametrised formula based on L, ψ a formula and $x_1, \ldots, x_n \notin L$ be n propositional variables not appearing in λ.*

If $\lambda[(\chi_i/x_i)_i] \equiv \alpha = \varphi \wedge \bigwedge_{1 \le j \le l} (\theta_j \vee \bigvee_{p \in P_j} x_p \vee \bigvee_{q \in N_j} \neg x_q)$

for a natural number l, appropriate $\varphi, \theta_j \in L$ and subsets P_j, N_j of $\{1, \ldots, n\}$, $1 \le j \le l$,

then $\lambda[(\chi_i/x_i \wedge \psi)_i] \equiv \alpha[(x_i/x_i \wedge \psi)_i] = \varphi \wedge \bigwedge_{1 \le j \le l} (\theta_j \vee \bigvee_{p \in P_j} (x_p \wedge \psi) \vee \bigvee_{q \in N_j} \neg(x_q \wedge \psi))$.

Example 3.9. *Consider the parametrised formula* $\lambda = p \wedge (\chi_1 \rightarrow \chi_2 \wedge q)$. $\lambda[(\chi_i/x_i)_i] = p \wedge (x_1 \rightarrow x_2 \wedge q)$ *is equivalent to the CNF* $p \wedge (\neg x_1 \vee x_2) \wedge (\neg x_1 \vee q)$. *This latter formula corresponds to* α *in Proposition 3.8.* φ *represents the clauses that do not contain* x_i, l *is the number of clauses containing new variables,* P_j *is the set of indexes of new variables that appear positively in the* j^{th} *such clause,* N_j *the set of indexes of new variables that appear as negative literals. Matching the elements of* α *there with our formula, we get* $\varphi = p$, $l = 2$ *(as there are two clauses containing new variables),* $\theta_1 = \bot$, $P_1 = \{2\}$, $N_1 = \{1\}$ *for the second clause* $\neg x_1 \vee x_2$, *and* $\theta_2 = q$, $P_1 = \emptyset$, $N_2 = \{1\}$ *for the third clause* $\neg x_1 \vee q$. *The proposition now states that, e.g.,* $\lambda[(\chi_i/x_i \wedge r)_i] = p \wedge (x_1 \wedge r \rightarrow x_2 \wedge r \wedge q)$ *is equivalent to* $p \wedge (\neg(x_1 \wedge r) \vee (x_2 \wedge r)) \wedge (\neg(x_1 \wedge r) \vee q)$.

The next result contains the main step towards eliminating the additional variables from the core. It formalises that it suffices to keep the part of the core that talks about the language L the parametrised observation is based on. The rest can be absorbed into the unknown subformulae and be pushed into the revision history of the agent. For later use think of σ as any sequence of revision inputs $\iota_k(o)$ the agent has received during the observation, where the unknown subformulae were instantiated by new variables. σ' is the same sequence, except that any unknown subformula was replaced by the conjunction of the corresponding new variable and the part of the core belief that also talks about the new variables.

Proposition 3.10. *Let* L *be a finitely generated propositional language.*
Let $x_1, \ldots, x_n \notin L$ *be additional propositional variables.*
Let $\blacktriangle \wedge \psi$ *be a formula such that* $\blacktriangle \in L$ *and* $Cn(\blacktriangle) = Cn(\blacktriangle \wedge \psi) \cap L$.
Let $\sigma = (\alpha_m, \ldots, \alpha_1)$ *be a sequence of formulae with*
$\alpha_i \equiv \varphi_i \wedge \bigwedge\limits_{1 \leq j \leq l_i} (\theta_{ij} \vee \bigvee\limits_{p \in P_{ij}} x_p \vee \bigvee\limits_{q \in N_{ij}} \neg x_q)$ *such that* $\varphi_i, \theta_{ij} \in L$
Let $\sigma' = (\alpha'_m, \ldots, \alpha'_1)$ *with* $\alpha'_i = \alpha_i[(x_k/x_k \wedge \psi)_k]$, *that is,*
$\alpha'_i \equiv \varphi_i \wedge \bigwedge\limits_{1 \leq j \leq l_i} (\theta_{ij} \vee \bigvee\limits_{p \in P_{ij}} (x_p \wedge \psi) \vee \bigvee\limits_{q \in N_{ij}} \neg(x_q \wedge \psi))$
Then $f(\sigma \cdot \blacktriangle \wedge \psi) \equiv f(\psi \cdot \sigma' \cdot \blacktriangle)$.

Abusing notation, we could also write $f(\sigma[(\chi_i/x_i)_i] \cdot \blacktriangle \wedge \psi) \equiv f(\psi \cdot \sigma[(\chi_i/x_i \wedge \psi)_i] \cdot \blacktriangle)$ when taking a sequence σ of parametrised formulae based on L and keeping the assumptions about $\blacktriangle$, ψ and the additional variables x_i. The core belief is weakened and the instantiations of the unknown subformulae are strengthened accordingly. The ψ appended to the front ensures the equivalence in case the $x_i \wedge \psi$ did not do the job, yet.

This result can now be used to show that for a parametrised observation there is indeed an explanation where the core belief does not contain the additional variables we used for instantiating the unknown subformulae. The idea is to replace χ_i by x_i, then calculate the

rational explanation and finally use Proposition 3.10 to weaken the core belief such that it does not talk about the additional variables.

Proposition 3.11. *If* $[\rho, \blacktriangle]$ *explains* $o[(\chi_i/x_i)_i]$ *then there exist* $\blacktriangle'$ *and* ψ *such that* $\blacktriangle'$ *does not contain any* x_i *and* $[\rho \cdot \psi, \blacktriangle']$ *explains* $o[(\chi_i/x_i \wedge \psi)_i]$.

We can even go one step further. Let $\blacktriangle'$ be the core belief returned when applying the rational explanation algorithm to $o[(\chi_i/x_i)_i]$ and $\varphi \in L(o)$ be any formula entailed by $\blacktriangle'$. Then *any* $o[(\chi_i/\phi_i)_i]$-acceptable core will entail φ. That is, $Cn(\blacktriangle') \cap L$ (which is independent of the instantiations ϕ_i of χ_i) is a safe conclusion about the core belief no matter what the instantiations of the unknown subformulae really are. Any formula inconsistent with that set will be rejected by the agent.

Proposition 3.12. *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* $Cn(\blacktriangle'') = Cn(\blacktriangle') \cap L$.

Then $\blacktriangle \vdash \blacktriangle''$.

As a consequence, analogously to the observations considered in the last chapter, we can again prove the existence of a logically weakest core that explains a (parametrised) observation. Naturally this core does not work with all instantiations of the unknown subformulae but only for some. The above results tell us how to construct that core and yield *one possible* choice for the χ_i.

Proposition 3.13. *Let* $[\rho, \blacktriangle]$ *be the rational explanation for* $o[(\chi_i/x_i)_i]$ *and* $\blacktriangle'$ *such that* $Cn(\blacktriangle') = Cn(\blacktriangle) \cap L(o)$.

Then $\blacktriangle'$ *is the unique weakest* o*-acceptable core.*

Example 3.14. *Consider again the formalisation* $o = \langle(\chi, \chi, \emptyset), (p, q \wedge \neg\chi, \emptyset)\rangle$ *of Example 3.1. We saw that the rational explanation for* $o[\chi/x]$ *is* $[(p \wedge \neg x \rightarrow q), p \rightarrow \neg x]$. o *is based on the language* L *constructed from the variables* p *and* q. $Cn(p \rightarrow \neg x) \cap L = Cn(\top)$. *Any core belief will trivially entail a tautology. To see that there really is an explanation using that core belief, note that* $o[\chi/\neg p] = \langle(\neg p, \neg p, \emptyset), (p, q \wedge p, \emptyset)\rangle$ *is explained by* $[(p \rightarrow q), \top]$. *However, these are not the observation and explanation that are used in the proof of the general result. Here we gave an example that does not even contain additional variables. This is not always possible as we will show in the next section. We can always eliminate the additional variables from the core belief (Proposition 3.13) but not necessarily from the instantiations of the unknown subformulae and the revision history.*

Although it is not quite as fundamental as Proposition 2.35, the next result shows that an analogous property also holds for parametrised observations. If there are two cores that explain an observation, possibly using different instantiations for the unknown subformulae, then we also find instantiations such that the disjunction of those cores yields an explanation.

Proposition 3.15. *If $\blacktriangle_1$ is o_1-acceptable for $o_1 = o[(\chi_i/\phi_i^1)_i]$ and $\blacktriangle_2$ is o_2-acceptable for $o_2 = o[(\chi_i/\phi_i^2)_i]$ then there are formulae $\phi'_1, \ldots, \phi'_n$ such that $\blacktriangle_1 \vee \blacktriangle_2$ is o'-acceptable for $o' = o[(\chi_i/\phi'_i)_i]$.*

3.2.4 the impact of extending the language

All the results used for proving the existence of a(n optimal) o-acceptable core belief depend on the possibility of extending the language $L(o)$. The unknown subformulae appearing in o were instantiated by formulae containing additional variables x_i not appearing in $L(o)$. Proposition 3.6 yields a necessary condition for an explanation for o to exist, which is possible only if the observation $o' = o[(\chi_i/x_i)]$ has an explanation. This can be checked using the rational explanation construction. Whether we accept this as a sufficient condition, as well, depends on the point of view.

Obviously, if there is an o'-acceptable core, then we will indeed find an explanation for o'. In this sense the condition might appear to be sufficient. But o' and hence most probably its explanation contain propositional variables that did not belong to $L = L(o)$. That is, for explaining o we used more than the language o informs us about. But does that guarantee that the χ_i might also be instantiated with elements of L and the resulting observation still has an explanation? We will show that the answer to this question is negative, and in this sense the existence of an o'-acceptable core is not sufficient. In other words, although o' may have an explanation, there may be none when restricting the instantiations of the χ_i to elements of L.

Example 3.16. *Consider $o = \langle (q, q, \emptyset), (p, p \wedge q \wedge r, \emptyset), (\neg q, \neg q, \emptyset), (\chi, \top, \emptyset), (p, p \wedge q \wedge \neg r, \emptyset)\rangle$. This parametrised observation contains one unknown subformula χ which appears only once. It represents a revision input that is completely unknown. All known revision inputs are accepted. After receiving the first p, the agent additionally believes $q \wedge r$ and after receiving the second one it believes $q \wedge \neg r$.*

$[(p \wedge q \to r), \top]$ explains $o[\chi/p \to (q \wedge \neg r)]$, i.e., instantiating the unknown subformula with $p \to (q \wedge \neg r)$ there is quite a simple explanation. Given that the core belief is to be $\top$, $p \wedge \chi$ must entail q as otherwise $f(\chi, p, \top) \not\vdash q$ and hence $\neg q$ would be admitted into the belief set, but the observation requires that after receiving the second p the agent believes q. This requires $\chi \vdash p \to q$. Further χ must entail more than just $p \to q$. Assume it did

not. Then the positive conditional corresponding to the agent's receiving the first p would be $f(q, p, \top) \Rightarrow p \wedge q \wedge r$, i.e., $p \wedge q \Rightarrow p \wedge q \wedge r$. The conditional corresponding to receiving the second p would be $f(q, p, \neg q, p \rightarrow q, p, \top) \Rightarrow p \wedge q \wedge \neg r$, i.e., $p \wedge q \Rightarrow p \wedge q \wedge \neg r$. As our framework defines a rational consequence relation (Proposition 2.41), this entails the conditional $p \wedge q \Rightarrow \bot$. However, $\mathcal{A}$ believes $\bot$ if and only if the core belief is inconsistent, which it is not — contradiction, so χ must indeed entail more than just $p \rightarrow q$. For the above explanation we chose $p \rightarrow \neg r$.

Next, consider $o' = o \cdot \langle (p \wedge r, q, \emptyset) \rangle$. $L = L(o) = L(o')$ contains only the propositional variables p, q and r. $[(p \wedge q \rightarrow r), p \rightarrow q]$ explains $o'[\chi / p \rightarrow \neg r]$, so there is an explanation that is restricted to L. Note that the core belief is not a tautology. $[(p \wedge x \rightarrow \neg r, p \wedge q \rightarrow (\neg x \wedge r)), \top]$ explains $o'[\chi / x \wedge (p \rightarrow q)]$. So there is an explanation where the core belief is a tautology, but it is not restricted to L. We would have found a similar explanation by calculating the rational explanation for $o'[\chi / x]$ and then applying Proposition 3.13 for ridding the core belief of x (the proof indicating how to instantiate χ instead). We will now show that there is no explanation restricted to L where the core belief is a tautology. Assume there is, then χ must entail $p \rightarrow q$ and something else as argued above. Further χ must be consistent with $p \wedge r$. If it was not we would have $f(\chi, p, p \wedge r) = p \wedge r \nvdash q$, so $\neg q$ would be admitted into the belief set, contradicting the information in o' that q is believed after receiving $p \wedge r$. Basically we need $f(q, p, \neg q, \chi, p) \equiv p \wedge q \wedge \chi \not\equiv p \wedge q$ as otherwise we will have the same contradicting conditionals $p \wedge q \Rightarrow p \wedge q \wedge r$ and $p \wedge q \Rightarrow p \wedge q \wedge \neg r$ as above.

Hence we look for a formula $\psi \in L$ such that $\psi \vdash p \rightarrow q$, $\psi \nvdash \neg p \vee \neg r$ implying $p \wedge q \wedge \psi \nvdash \neg r$, $p \wedge q \wedge \psi \nvdash r$ (otherwise $f(q, p, \neg q, \psi, p) \equiv p \wedge q \wedge \psi \vdash r$ but $\neg r$ must be believed after receiving the second p) and $p \wedge q \wedge \psi \not\equiv p \wedge q$. Obviously such a formula cannot exist. To be different from $p \wedge q$, $p \wedge q \wedge \psi$ would have to talk about r but the above restrictions do not allow that. Consequently, there is no instantiation of ψ from L such that $\top$ is an acceptable core.

Now, consider $o'' = o' \cdot o^{\top}$, where $o^{\top} = \langle (\varphi_1, \varphi_1, \emptyset), \ldots, (\varphi_n, \varphi_n, \emptyset) \rangle$ such that φ_i varies over all semantically different formulae from L. Basically, $o^{\top}$ requires that an o''-acceptable core belief must be consistent with any formula containing only p, q, r, as any such formula must be believed upon receiving it. It is easily checked that $[(p \wedge x \rightarrow \neg r, p \wedge q \rightarrow (\neg x \wedge r)), \top]$ also explains $o''[\chi / x \wedge (p \rightarrow q)]$. However, there cannot be an explanation that is restricted to L. Due to (the contraposition of) Proposition 2.33, $\top$ is the only core in question, as any other formula $\varphi \in L$ would be in conflict with $(\neg \varphi, \neg \varphi, \emptyset)$ which is an element of o''. But we showed above that there is no instantiation of χ from L such that $\top$ is an acceptable core.

There is an important lesson to be learned from this example. From the rational explanation for $o[(\chi_i / x_i)_i]$ (yielding an optimal core belief via Proposition 3.13) we can generally say nothing about possible explanations where the instantiations of the χ_i are from $L = L(o)$.

There may be one with the same optimal core, there may be one but the core might have to be logically stronger than the optimal one, there may be no such explanation at all.

Note that Proposition 3.6 makes no assumption about the language of the instantiations ϕ_i of the unknown subformulae. They may or may not belong to L. They may contain arbitrarily (but finitely) many propositional variables not belonging to L. However, that proposition has an interesting implication. It says that in case $o[(\chi_i/\phi_i)_i]$ has an explanation then so does $o[(\chi_i/x_i)_i]$, but $o[(\chi_i/x_i)_i]$ contains only variables from L and n additional variables x_i. As that observation has an explanation, the rational explanation construction will return one. However, that construction uses only formulae that are already present in the observation. Consequently, it does not invent new variables. So, no matter how many variables not appearing in L 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.

To understand why the number of additional variables matters, it is useful to recall that a rational consequence relation, which is another way of looking at the initial epistemic state we are after, can be represented by a total preorder on worlds [54] (see also Section 1.2). An additional variable x potentially doubles the number of available worlds, there may be one x- and one $\neg x$-variant of a former world now. Depending on our way of dealing with core beliefs, some of them may be irrelevant, e.g., the core belief $p \rightarrow x$ eliminates worlds in which p and $\neg x$ hold. Having several copies of the worlds corresponding to L allows for more structure among them. This explains why there may be an explanation when allowing an extra variable while there is none when restricting ourselves to L.

We can give a necessary and sufficient condition for the existence of an explanation for o which is restricted to $L(o)$. Unfortunately, it does not give rise to an efficient way of actually calculating such an explanation. This is because it is not obvious how to strengthen the core yielded by the rational explanation construction so that the necessary equivalence emerges and it will still be o-acceptable.

Proposition 3.17. *Let o be a parametrised observation based on L. An explanation of o restricted to L exists if and only if $\blacktriangle \wedge \bigwedge_{1\leq i\leq n} x_i \leftrightarrow \phi_i$ is $o[(\chi_i/x_i)_i]$-acceptable for some $\blacktriangle, \phi_i \in L$ and $x_i \notin L$.*

$\blacktriangle \wedge \bigwedge_{1\leq i\leq n} x_i \leftrightarrow \phi_i$ will of course entail the core belief $\blacktriangle'$ of the rational explanation of $o[(\chi_i/x_i)_i]$. But it is not guaranteed that $\blacktriangle'$ can be strengthened such that every additional variable x_i can be assumed to behave exactly as some formulae from $L(o)$. This is why in the above example there is an explanation for o'' but none that is restricted to $L(o'')$.

3.2.5 comparing explanations

The main criterion for comparing explanations should be the quality of the conclusions they allow us to draw. This is why we argued in favour of using the weakest core belief $\blacktriangle_\vee(o)$ for observations o and respectively (abusing notation) $\blacktriangle_\vee(o'[(\chi_i/x_i)_i]) \cap L(o')$ for parametrised observations o'. It guarantees safe conclusions with respect to the agent's real core belief. In Section 2.8 we showed that assuming the wrong core belief — in our case this would be a core belief that is weaker than the actual one — greatly affects the quality of the conclusions about the agent's other beliefs. And even if the core is correct, the rational prefix does not necessarily yield 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. We want to give a series of examples illustrating the difficulty of defining a suitable preference relation among explanations in this setting. Much depends not only on the core belief but also on the instantiation of the unknown subformulae. If the latter is equivalent for the two explanations in question, we can use the criteria introduced in the last chapter, yielding that the rational explanation is the best we can do. So the question is what happens if they may vary.

To put it differently, rather than looking for the best explanation of an observation, which we investigated in the last chapter, we are looking for the optimal instantiation of the unknown subformulae. Having that we could simply apply the results already presented.

Example 3.18. *Consider* $o = \langle (p, p, \emptyset), (\chi, \top, \emptyset), (q, \neg p \wedge r, \emptyset)\rangle$. *This parametrised observation tells us that the first revision input* p *is accepted. Then the agent received an input we know nothing about, neither its logical content nor what was believed or not believed after receiving it. Finally, after receiving* q *the agent believes* $\neg p \wedge r$.

It is easy to see that it is possible to explain o *using* $\top$ *as core belief. To do so* χ *must entail* $q \rightarrow \neg p$, *as otherwise the first input* p *would still be believed after receiving* q *which is not the case according to* o. *So let us assume that the unknown input was in fact* $q \rightarrow \neg p$. *This defines an observation to which we can apply the rational explanation yielding* $[q \wedge \neg p \rightarrow r, \top]$. *In particular, we would conclude that the agent initially believed* $q \wedge \neg p \rightarrow r$.

This is not necessarily the case, as $\mathcal{A}$ *may have received* $q \rightarrow (\neg p \wedge r)$ *instead of* $q \rightarrow \neg p$. *For this instantiation of* χ *the rational explanation would be* $[(), \top]$. *The corresponding belief traces would be* $(q \wedge \neg p \rightarrow r, p, p \wedge \neg q, q \wedge \neg p \wedge r)$ *for the instantiation of* χ *given first and* $(\top, p, p \wedge \neg q, q \wedge \neg p \wedge r)$ *for the second one.*

In this example, we would clearly favour the second instantiation of the unknown subformula over the first one. It yields weaker beliefs at *every* point of the observation.[2] In fact there cannot be a better explanation for o. Looking at the rational explanation $[(q \wedge x \wedge \neg p \rightarrow r), \neg x \vee \neg q \vee \neg p]$ for $o[\chi/x]$ this explanation might have been found by pushing whatever possible from both the core belief and the rational prefix into the instantiation of χ which here makes the additional variable x redundant.

Proposition 3.13 tells us which part of the core $\blacktriangle$ can be absorbed by the unknown subformulae: $Cn(\blacktriangle) \setminus L(o)$. It is the part that talks about the additional variables and was shown to be the non-essential part of the core belief.[3] Writing ψ for that part of the core, the proofs use $x_i \wedge \psi$ as the instantiation of the unknown subformulae. The following example shows that this instantiation does not yield an optimal belief trace, indicating that it is not trivial to find an optimal instantiation for the χ_i. It does not even require an optimisation with respect to the rational prefix as that is the empty sequence, anyway.

Example 3.19. *Consider* $o = \langle (p, p, \emptyset), (\chi_1, \top, \emptyset), (q, \neg p, \emptyset), (r, r, \emptyset), (\chi_2, \top, \emptyset), (s, \neg r, \emptyset)\rangle$. *$p$ is believed upon receiving it, then there is an input we have no information about and after receiving a further input q the negation of p is believed. Then the same happens again with r, another unknown input and s. The rational explanation for* $o[\chi_1/x_1, \chi_2/x_2]$ *is* $[(), \blacktriangle]$ *where* $\blacktriangle = (\neg p \vee \neg x_1 \vee \neg q) \wedge (\neg r \vee \neg x_2 \vee \neg s)$. *Note that* $Cn(\blacktriangle) \cap L(o) = Cn(\top)$, *so the entire core can be pushed into the intermediate inputs. Denoting* $\neg p \vee \neg x_1 \vee \neg q$ *by* ψ_1 *and* $\neg r \vee \neg x_2 \vee \neg s$ *by* ψ_2, *the belief trace using the rational explanation* $[(), \top]$ *of* $o[\chi_1/x_1 \wedge \blacktriangle, \chi_2/x_2 \wedge \blacktriangle]$ *is* $(\top, p, p \wedge x_1 \wedge \neg q \wedge \psi_2, \neg p \wedge x_1 \wedge q \wedge \psi_2, \neg p \wedge x_1 \wedge q \wedge \psi_2 \wedge r, \neg p \wedge x_1 \wedge q \wedge r \wedge x_2 \wedge \neg s, \neg p \wedge x_1 \wedge q \wedge \neg r \wedge x_2 \wedge s)$.

If instead we do not put the entire core $\blacktriangle$ into all unknown subformulae but only the relevant part, we would get $o[\chi_1/x_1 \wedge \psi_1, \chi_2/x_2 \wedge \psi_2]$ *as observation. The rational explanation is again* $[(), \top]$ *and the corresponding belief trace* $(\top, p, p \wedge x_1 \wedge \neg q, \neg p \wedge x_1 \wedge q, \neg p \wedge x_1 \wedge q \wedge r, \neg p \wedge x_1 \wedge q \wedge r \wedge x_2 \wedge \neg s, \neg p \wedge x_1 \wedge q \wedge \neg r \wedge x_2 \wedge s)$. *The beliefs are logically weaker (at most equivalent) at every point of the observation. The belief traces of this example will be recalled in a table below for better comparison.*

[2] Example 3.20 contains a parametrised observation illustrating that this is not always possible.

[3] As $\blacktriangle$ is believed by the agent at every single point in time, the entire core could actually be conjoined with every single input the agent ever received. The belief trace would not change. This makes it quite simple to absorb any part of the core into the unknown subformulae. For (sub)formulae from ρ in the agent's assumed initial epistemic state $[\rho, \blacktriangle]$, the case is more complicated. Different formulae from ρ might be chosen at different points of the observation. So not all the formulae chosen when the unknown input χ_i instantiated by the new variable x_i was received can be absorbed into that input.

In short, it is quite obvious what can be pushed down into the unknown subformulae in order to make the core belief as weak as possible, but it is not obvious what can be pushed up from ρ in order to optimise the belief trace, which is what the rational prefix is about.

▲ contains both additional variables x_1 and x_2. If the x_i are replaced by $x_i \wedge \blacktriangle$, the agent is assumed to learn something about the second unknown input already when receiving the first one. If the x_i are replaced by $x_i \wedge \psi_i$ instead, the agent is assumed to get information as late as possible. That way a preferable belief trace is achieved. For o in this example, it is quite obvious how to instantiate the unknown subformulae (which turn out to be unknown revision inputs) in order to get an optimal belief trace. In the general case, however, the χ_i may appear several times in the recorded revision inputs as well as in the beliefs and non-beliefs. Hence, it is not clear for arbitrary parametrised observations how to arrive at an optimal instantiation of the χ_i.

When using additional variables, we should not really be interested in all the beliefs contained in the belief trace. As we cannot say which other variables, if any, the agent has heard about, we should look only at beliefs from the language $L(o)$. Note that for the core belief we have already done so by showing that we can restrict the core to that language and still get an explanation. In the following table we compare the belief traces from the above example and the belief trace for the rational explanation for $o[\chi_1/x_1, \chi_2/x_2]$ when restricted to $L(o)$.[4]

ι	$o[\chi_1/x_1 \wedge \blacktriangle, \chi_2/x_2 \wedge \blacktriangle]$	$o[\chi_1/x_1 \wedge \psi_1, \chi_2/x_2 \wedge \psi_2]$	$o[(\chi_i/x_i)_i]\|L(o)$
	$\top$	$\top$	$\top$
p	p	p	p
χ_1	$p \wedge x_1 \wedge \neg q \wedge \psi_2$	$p \wedge x_1 \wedge \neg q$	$p \wedge \neg q$
q	$\neg p \wedge x_1 \wedge q \wedge \psi_2$	$\neg p \wedge x_1 \wedge q$	$\neg p \wedge q$
r	$\neg p \wedge x_1 \wedge q \wedge \psi_2 \wedge r$	$\neg p \wedge x_1 \wedge q \wedge r$	$\neg p \wedge q \wedge r$
χ_2	$\neg p \wedge x_1 \wedge q \wedge r \wedge x_2 \wedge \neg s$	$\neg p \wedge x_1 \wedge q \wedge r \wedge x_2 \wedge \neg s$	$\neg p \wedge q \wedge r \wedge \neg s$
s	$\neg p \wedge x_1 \wedge q \wedge \neg r \wedge x_2 \wedge s$	$\neg p \wedge x_1 \wedge q \wedge \neg r \wedge x_2 \wedge s$	$\neg p \wedge q \wedge \neg r \wedge s$

Strengthening the instantiations x_i for χ_i using the entire core belief is slightly worse than strengthening them by what is really necessary (ψ_i). However, restricting the elements of the belief trace to $L(o)$ yields exactly those elements, we are interested in. It is an open but, as we will argue, not very interesting question whether this is generally the case.

We conjecture that in fact we need not look for some optimal instantiation of the χ_i but we just have to use new variables x_i, calculate the rational explanation for $o[(\chi_i/x_i)_i]$ and restrict our conclusions to $L(o)$. The rationale is that no matter what the χ_i really are, the x_i yield a good approximation of the interactions that might be caused by the unknown subformulae (see Proposition 3.4). Using hypothetical reasoning we can verify the conclusions drawn.

[4]The rational explanation of $o[\chi_1/q \rightarrow \neg p, \chi_2/s \rightarrow \neg r]$ would yield exactly that belief trace. So this instantiation of the unknown subformulae would yield an optimal belief trace without introducing new variables. However, such instantiations cannot always be found (cf. Example 3.16).

Recall Section 2.8 where we illustrated that even if we have the correct core belief, conclusions about further beliefs of the agent might be completely wrong because the rational prefix is not close enough to the agent's actual sequence of revision inputs prior to the observation. In the current setting, the problem is even more severe. Even if we have the correct core and the right sequence, the conclusions based on the assumption of some particular instantiation may still be quite wrong.

Example 3.20. *Consider* $o = \langle(\neg p \wedge q, \neg p \wedge q, \emptyset), (\chi, \top, \emptyset), (q \leftrightarrow r, q \leftrightarrow r, \emptyset)\rangle$. *The rational explanation for* $o[\chi/x]$ *is* $[(), \top]$ *yielding as belief trace* $(\top, \neg p \wedge q, \neg p \wedge q, \neg p \wedge q \wedge r)$ *when restricting it to* $L(o)$. *We would also get this explanation and belief trace if* χ *is instantiated by* $\top$.

Now assume $\mathcal{A}$*'s initial belief state was indeed* $[(), \top]$ *but the completely unknown input was in fact* p, *so the real belief trace was* $(\top, \neg p \wedge q, p, p \wedge (q \leftrightarrow r))$. *Hence, the conclusion that* $\neg p \wedge q$ *was believed at all times during the observation and that finally* r *was believed was not safe.*

The underlying problem is similar to that of choosing the weakest core belief. There we assume that the agent accepts as many inputs as possible and when calculating the belief set from its current epistemic state $[\rho, \blacktriangle]$ adds as many important formulae from ρ as it can, rejecting one only if it is absolutely necessary. The same thing happens for the instantiation x_i of χ_i. The rational explanation tries to minimise the impact of the unknown subformulae to what is absolutely necessary. In the example, it is possible to assume that the first input $\neg p \wedge q$ remains consistent with all following inputs.

It is debatable whether such an assumption is indeed natural or justified. A negative interpretation would be to defer correcting mistakes until it has to be done, to keep sticking to the wrong story until it cannot be kept up anymore. Consider the observation $o = \langle(\neg p, \neg p, \emptyset), (\chi, \top, \emptyset), (p \leftrightarrow q_1, p \leftrightarrow q_1, \emptyset), \ldots, (p \leftrightarrow q_n, p \leftrightarrow q_n, \emptyset), (r, q_1 \wedge \cdots \wedge q_n, \emptyset)\rangle$. The rational explanation for $o[\chi/x]$ yields that a change of mind with respect to $\neg p$ (and consequently for all $\neg q_i$ as $\neg p$ is assumed to be believed until that point) comes only at the last moment caused by the agent receiving r and that $\neg r$ is believed just before. But it is also possible that the unknown input was p. This can be interpreted as correcting the small mistake of believing $\neg p$ made early on and after that everything goes smoothly. Note that in this case only the value of p changes whereas the above explanation would cause the value of all variables to change at some point. Should the latter explanation not therefore be preferred?

The flaw in this line of argument is that in order to assume that the mistake is corrected earlier, it has to be recognised as a mistake. If in the example, the last piece of information

in the observation was missing, nothing would tell us that a change in mind must have occurred. Even if it is present, nothing indicates that the unknown input must have been p. In this sense both explanations are equally (im)plausible.[5]

3.2.6 summary

In this section we showed that the calculation of an optimal core belief is still possible when assuming the existence of unknown subformulae in the observation. We motivated this scenario by claiming that it is possible that the logical content of the revision inputs is not completely known. 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 that although we can construct a core belief from $L(o)$ this does not guarantee that o can be entirely explained without extending the language. Some of the unknown subformulae may have to be instantiated by formulae containing variables not appearing in o. We claim that it is not useful to look for an optimal instantiation of the unknown subformulae. The conclusions heavily depend on the choice of the instantiation of the χ_i and even if we had the correct one, Section 2.8 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 the belief trace for the corresponding rational explanation restricted to $L(o)$. This allows 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 conclusions about other beliefs and non-beliefs implied by the explanation thus obtained.

Once more, the additional assumption that the belief corresponding to the last revision input in the (parametrised) observation is complete 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 $\langle(p \wedge \chi, \top, \emptyset), (\neg p, \neg p, \emptyset)\rangle$. This might not be a very interesting observation 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.

[5] $o = \langle(\neg p, \neg p, \emptyset), (\chi, \top, \emptyset), (p \leftrightarrow q_1, p \leftrightarrow q_1, \emptyset), \ldots, (p \leftrightarrow q_n, p \leftrightarrow q_n, \emptyset), (r, q_1 \wedge \cdots \wedge q_n, \emptyset)\rangle$ provides only very weak information, so it should not be surprising that most hypotheses based on the rational explanation turn out to be not safe.

3.3 Intermediate inputs

3.3.1 why consider 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. As mentioned before, this corresponds to the assumption of having an eye on the agent at all times during the observation. In this section, we want to investigate the case where this assumption is dropped. 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. The observing agent can never be sure which sensory data $\mathcal{A}$ might turn into revision inputs. Further, continuous surveillance is next to impossible. $\mathcal{A}$ 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 $\langle(\chi, \top, \emptyset)\rangle$; χ again represents an unknown formula. Note that this is different from $\langle(\chi, \chi, \emptyset)\rangle$ 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.

Consider the following observation without intermediate inputs: $o = \langle(p, q, \emptyset), (p, \neg q, \emptyset)\rangle$. There is no o-acceptable core. Assuming a single intermediate input $\langle(\chi, \top, \emptyset)\rangle$, there are three positions where it could have been received; (i) before the first p, (ii) in between the two p, and (iii) after the second p. Options (i) and (iii) do not really make sense. Inputs that may have been received before the observation started are already contained in the initial epistemic state, hence they are already represented in the rational prefix.[6] Inputs assumed to be received after the time of the observation o cannot help explaining it, as Right Monotony tells us that an explanation for the extended observation (with intermediate inputs assumed after o) must already be one for the prefix o. So in this example, there is only one reasonable position for the intermediate input. When we consider $o' = \langle(p, q, \emptyset), (\chi, \top, \emptyset), (p, \neg q, \emptyset)\rangle$, instantiating the unknown formula χ with $p \rightarrow \neg q$, the core belief $\top$ 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. However, if number and positions are fixed

[6] In that sense, the rational prefix can be interpreted as the sequence of intermediate inputs received before the observation started.

then we can use the results from the last section in order to calculate the weakest possible core belief. To represent the intermediate inputs we simply have to introduce *further* unknown subformulae not contained in o as generally nothing is known about the relation between recorded inputs and the intermediate ones. For example, assume we have the partial observation $o = \langle(p, q \wedge \chi_1, \emptyset), (r, \neg q, \emptyset), (p, q, \{\chi_1\})\rangle$ 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 $o' = \langle(p, q \wedge \chi_1, \emptyset), (r, \neg q, \emptyset), (\chi_2, \top, \emptyset), (\chi_3, \top, \emptyset), (p, q, \{\chi_1\})\rangle$ which now contains an entry for every input received. Hence, intermediate inputs and partial information about inputs are related but distinct cases. There is an essential difference between knowing there was (not) an input, possibly being ignorant of the exact logical content, and not knowing whether there was an input at all.

The most important result for this section is Proposition 2.33, a direct lemma of which is: If $\blacktriangle$ is not o-acceptable, then $\blacktriangle$ is not $o' \cdot o \cdot o''$-acceptable for any observations o' and o''. So if we know for some part o of the whole observation that during that time no intermediate input was received, then the core explaining the entire observation must also explain o. Left Monotony and Right Monotony imply that any $o' \cdot o \cdot o''$-acceptable core entails $\blacktriangle_\vee(o)$. In particular any $\langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$-acceptable core entails $\bigwedge \blacktriangle_\vee(\langle(\varphi_i, \theta_i, D_i)\rangle)$. This states that the core belief must block a revision input that contradicts the information about beliefs and non-beliefs after its having been received. Note that this is the part that can be read off the observation directly (cf. Section 2.7).

This directly implies that intermediate inputs cannot replace the core belief for explaining an observation. For example, $\langle(p, \neg p, \emptyset)\rangle$ or $\langle(p, \top, \{p\})\rangle$ can only be explained using core beliefs (also recall our above remark that intermediate inputs before or after an observation do not make sense). However, as mentioned above, not all observations can be explained with core beliefs only. In order to explain $\langle(p, q, \emptyset), (p, \neg q, \emptyset)\rangle$ intermediate inputs are necessary. So core beliefs cannot replace intermediate inputs. The two concepts are complementary either.[7]

3.3.2 number and positions of intermediate inputs matter

The following example is to illustrate that generally we cannot change the number and positions of intermediate inputs without an impact on the weakest explaining core. This means that it is highly relevant to our conclusions about $\mathcal{A}$ how many intermediate inputs

[7] There are still observations that cannot be explained when using both concepts. $\langle(p, \top, \{p, \neg p\})\rangle$ and $\langle(p, \top, \emptyset), (q, \top, \{p, \neg p\})\rangle$ are two examples. The reason is Proposition 2.20. In the assumed revision framework it is impossible that a revision input is ignored or forgotten. Using a different framework for modelling the observed agent, such observations might have an explanation.

we assume at which points during the observation. Note the similarity to Example 3.19. There we assumed the observation to contain two revision inputs about which nothing was known, here these are interpreted as two possible intermediate inputs.

Example 3.21. *Consider* $o = \langle (p, p, \emptyset), (q, \neg p, \emptyset), (r, r, \emptyset), (s, \neg r, \emptyset)\rangle$. *If there was no intermediate input at all then the weakest* o*-acceptable core is* $(q \rightarrow \neg p) \wedge (s \rightarrow \neg r)$.

Assuming one intermediate input to have been received, there are three sensible positions to place it. Putting the intermediate input $q \rightarrow \neg p$ *at the second position in the observation yielding* $o' = \langle (p, p, \emptyset), (q \rightarrow \neg p, \top, \emptyset), (q, \neg p, \emptyset), (r, r, \emptyset), (s, \neg r, \emptyset)\rangle$ *the weakest* o'*-acceptable core is* $s \rightarrow \neg r$. *Putting any intermediate input at the third position, due to Left Monotony and Right Monotony the weakest explaining core still needs to entail* $(q \rightarrow \neg p) \wedge (s \rightarrow \neg r)$. *Finally, assuming the agent to have received the input* $s \rightarrow \neg r$ *before receiving* s, *the core belief* $q \rightarrow \neg p$ *explains the resulting observation.*

If we think that two intermediate inputs may have been received at arbitrary positions then $[(), \top]$ *explains* $o'' = \langle (p, p, \emptyset), (q \rightarrow \neg p, \top, \emptyset), (q, \neg p, \emptyset), (r, r, \emptyset), (s \rightarrow \neg r, \top, \emptyset), (s, \neg r, \emptyset)\rangle$.

Example 3.22. *Let* o_1 *and* o_2 *be the following two observations:* $o_1 = \langle (p, p, \emptyset)(q, q, \emptyset)\rangle$ *and* $o_2 = \langle (r, \neg p, \emptyset), (s, \neg q, \emptyset), (p \wedge r \wedge t, \neg q, \emptyset), (q \wedge s \wedge \neg t, p, \emptyset)\rangle$. $[(), \top]$ *is an explanation for* $o = o_1 \cdot \langle (p \rightarrow \neg r, \top, \emptyset), (q \rightarrow \neg s, \top, \emptyset)\rangle \cdot o_2$, *i.e., when assuming those two consecutive intermediate inputs between* o_1 *and* o_2 *then* $\top$ *is an acceptable core belief.*

Now, Proposition 3.13 allows us to calculate the weakest possible core when using a single intermediate input by calculating the rational explanation $[\rho, \blacktriangle]$ *for* $o' = o_1 \cdot \langle (x, \top, \emptyset)\rangle \cdot o_2$ *and then choose* $\blacktriangle'$ *such that* $Cn(\blacktriangle') = Cn(\blacktriangle) \cap L(o_1 \cdot o_2)$. *However,* $\blacktriangle_\vee(o')$ *is equivalent to* $(\neg p \vee \neg q \vee \neg r \vee \neg s \vee \neg t) \wedge (\neg x \vee \neg q \vee \neg r \vee \neg s) \wedge (\neg x \vee \neg p \vee \neg q \vee \neg r)$. *Hence, no matter what single intermediate input is chosen, any acceptable core must entail* $\neg p \vee \neg q \vee \neg r \vee \neg s \vee \neg t$ *and so* $\top$ *will never work.*

Example 3.22 shows that it is not generally possible to join consecutive intermediate inputs into a single one without effect on the core belief.[8] This might already have been guessed from our above remark that the rational prefix is in a sense a sequence of intermediate inputs received before the observation o started. If joining consecutive intermediate inputs into a single one were possible, there would always be a sequence ρ of length one such that $[\rho, \blacktriangle_\vee(o)]$ explains o.

[8] In fact, if we extend o_2 by $\langle (p \wedge q \wedge r \wedge s \wedge t, p \wedge q \wedge r \wedge s \wedge t, \emptyset)\rangle$ we will get the same explanation when using the two intermediate inputs. However, there is no explanation at all when using a single intermediate input.

3.3.3 formal results

In this section, 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 our 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 at any time.

Any number of intermediate inputs at any time The following proposition merely looks complicated, the intuition is quite simple. Consider $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. No matter how many intermediate inputs were received at any point during the observation, exactly one intermediate input between any two inputs φ_i and φ_{i+1} suffices to explain the observation. In the proposition this input is chosen to be the conjunction of the core belief together with the known inputs from the observation and the intermediate ones collected up to φ_{i+1}.

Proposition 3.23. *Let*
$$\begin{aligned} o = \langle &(\varphi_1, \theta_1, D_1), (\psi_{11}, \top, \emptyset), \ldots, (\psi_{1m_1}, \top, \emptyset), \\ &(\varphi_2, \theta_2, D_2), (\psi_{21}, \top, \emptyset), \ldots, (\psi_{2m_2}, \top, \emptyset), \\ &\ldots, \\ &(\varphi_{n-1}, \theta_{n-1}, D_{n-1}), (\psi_{(n-1)1}, \top, \emptyset), \ldots, (\psi_{(n-1)m_{n-1}}, \top, \emptyset), \\ &(\varphi_n, \theta_n, D_n)\rangle \end{aligned}$$
and $\iota'_i(o) = (\varphi_1, \psi_{11}, \ldots, \psi_{1m_1}, \varphi_2, \ldots, \psi_{(i-1)m_{i-1}}, \varphi_i)$ *denote the prefix of* $\iota(o)$ *with* φ_i *being the last element.*

If $[\rho, \blacktriangle]$ *explains* o*, then it also explains*
$$\begin{aligned} o' = \langle &(\varphi_1, \theta_1, D_1), (f(\iota'_2(o) \cdot \blacktriangle), \top, \emptyset), \\ &(\varphi_2, \theta_2, D_2), (f(\iota'_3(o) \cdot \blacktriangle), \top, \emptyset), \\ &\ldots, \\ &(\varphi_{n-1}, \theta_{n-1}, D_{n-1}), (f(\iota'_n(o) \cdot \blacktriangle), \top, \emptyset), \\ &(\varphi_n, \theta_n, D_n)\rangle. \end{aligned}$$

Note that this result does not contradict the claim that generally two consecutive intermediate inputs cannot be joined into a single one. That claim assumed that apart from those two inputs nothing else is changed. The proposition assumes that intermediate inputs are allowed at any position. So the number of intermediate inputs might in fact increase. Applying this result to Example 3.22, the effect of the two intermediate inputs, which cannot be combined into a single one, can be simulated by putting one intermediate input before the first element of o_2 and another between the first and the second one.

This proposition gives us a simple method to calculate the weakest possible core explaining an observation o when allowing any number of intermediate inputs at any position. As it

says that one intermediate input between any two known inputs suffices, we can put an entry $\langle(x_i, \top, \emptyset)\rangle$ with a new variable x_i between the entries $\langle(\varphi_i, \theta_i, D_i)\rangle$ and $\langle(\varphi_{i+1}, \theta_{i+1}, D_{i+1})\rangle$. We apply the rational explanation construction to the observation thus obtained and get a core belief $\blacktriangle$. The weakest core belief will then be $\blacktriangle'$ such that $Cn(\blacktriangle') = Cn(\blacktriangle) \cap L(o)$. To see this, let $\blacktriangle''$ be an o''-acceptable core where o'' is an observation obtained from o by placing some arbitrary number of intermediate inputs at any position in o. Proposition 3.23 yields that $\blacktriangle''$ is also acceptable for an observation where exactly one intermediate input was placed between any two elements in o and Proposition 3.12 yields that $\blacktriangle'' \vdash \blacktriangle'$.

We know for certain that this core belief $\blacktriangle'$ must entail $\blacktriangle_\vee(\langle(\varphi_i, \theta_i, D_i)\rangle)$ for any entry in o (Left Monotony and Right Monotony). The question is whether a core entailing exactly these formulae will always explain an observation for some instantiation of the intermediate inputs. If this was the case then we could never say anything about the core belief except for what can trivially be read off the observation (often this will be the case, anyway). However, the following example proves the existence of observations where the weakest explaining core is indeed stronger than $\bigwedge \blacktriangle_\vee(\langle(\varphi_i, \theta_i, D_i)\rangle)$.

Example 3.24. *Consider $\langle(p, \top, \emptyset), (q, \top, \{p \wedge q, \neg p \wedge q\})\rangle$. For both individual entries in the observation, $\top$ is the optimal core belief. Due to Proposition 2.20 p or $\neg p$ will be believed by the agent after it received q — no matter which intermediate inputs were received at any time during the observation. However, if q is accepted, the agent will hence believe $p \wedge q$ or $\neg p \wedge q$ which the observation does not allow. Consequently, q must not be accepted by the agent and any core belief explaining this observation must entail $\neg q$. So the core belief $\top$ cannot explain the observation even when allowing intermediate inputs.*

This gives us an idea what we can say about the agent's core belief in case we put no restrictions on the number and positions of intermediate inputs. 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. This means that assuming another intermediate input could potentially weaken the explaining core belief.

Proposition 3.25. *If $x \notin L(o_1 \cdot o_2)$ and $Cn(\blacktriangle) = Cn(\blacktriangle_\vee(o_1 \cdot \langle(x, \top, \emptyset)\rangle \cdot o_2)) \cap L(o_1 \cdot o_2)$ then $\blacktriangle_\vee(o_1 \cdot o_2) \vdash \blacktriangle$.*

Fixed positions of the 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 according to Proposition 3.25. 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 that position, i.e., ignoring possible intermediate inputs appearing later. To be more precise, having and observation o with arbitrarily many intermediate inputs at some position, an o-acceptable core will also be acceptable for the observation where we assume j intermediate inputs at that position.

If this claim is correct, we can introduce into o one entry $(\chi_i, \top, \emptyset)$ 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 $o' = \langle(\varphi_1, \theta_1, D_1),\ (\varphi_2, \theta_2, D_2),\ (x_1, \top, \emptyset),\ (x_2, \top, \emptyset),\ (x_3, \top, \emptyset), (\varphi_3, \theta_3, D_3),\ (\varphi_4, \theta_4, D_4),\ (x_4, \top, \emptyset),\ (\varphi_5, \theta_5, D_5)\rangle$ and eliminating the x_i from $\blacktriangle_\vee(o')$. Again, all x_i are propositional variables not contained in $L(o)$. The claim follows almost immediately from the following proposition.

Proposition 3.26. *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 $f(\sigma \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \equiv f(\sigma' \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))$ for all $1 \leq i \leq n$.*

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.[9] Applying Proposition 3.6 then tells us that we can replace the unknown intermediate inputs with new variables and be guaranteed that an explanation is found, if one exists at all.

Note also that more intermediate inputs have to be assumed when reasoning hypothetically about *future* revision inputs. In this case the assumed block of intermediate inputs has to yield equivalent results for more than just the recorded revision inputs. For each future input an additional intermediate input (per block) is needed. However, the proposition only yields

[9]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 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.

an upper bound. It is possible that a smaller number of intermediate inputs suffices. We currently investigate this question.

Fixed number of intermediate inputs If we are given a maximal (or exact) number n of intermediate inputs we can draw conclusions about the core belief of the agent using the following method. Due to Proposition 3.25 we should indeed assume n intermediate inputs. So let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_m, \theta_m, D_m)\rangle$ be the observation containing only recorded inputs. If $n > m-2$ and there are no restrictions with respect to where the intermediate inputs may have occurred then we can apply Proposition 3.23. In this case we have enough intermediate inputs at our disposal to assume one between any two recorded inputs, which allows us to calculate the weakest acceptable core belief. Otherwise, there are not enough intermediate inputs or they are not allowed at every position in o, so that result is not applicable. In this case, we create the set of all possible observations o' where n intermediate inputs have been inserted in o: $O' = \{o_1 \cdot \langle(x_1, \top, \emptyset)\rangle \cdot o_2 \cdot \ldots \cdot \langle(x_n, \top, \emptyset)\rangle \cdot o_{n+1} \mid o = o_1 \cdot \ldots \cdot o_{n+1}\}$.[10] If we have information about the positions of the intermediate inputs we can also take this into account when constructing O'. o_j may be empty, so consecutive intermediate inputs are explicitly allowed. Now any possible core belief will entail $\bigvee\{\blacktriangle \mid Cn(\blacktriangle) = Cn(\blacktriangle_\vee(o')) \cap L(o), o' \in O'\}$. Note that this formula need not be an o'-acceptable core, i.e., it may not really explain the observation using n intermediate inputs as the following example illustrates.

Example 3.27. *Consider the observation* $o = \langle(p, p, \emptyset), (q, \neg p, \emptyset), (r, r, \emptyset), (s, \neg r, \emptyset)\rangle$ *from Example 3.21. Assuming exactly one intermediate input using the above method we obtain* $(s \to \neg r) \vee ((q \to \neg p) \wedge (s \to \neg r)) \vee (q \to \neg p) \equiv (q \to \neg p) \vee (s \to \neg r)$ *as the formula entailed by all possible core beliefs.*

However, if there is only one intermediate input then either $\langle(p, p, \emptyset), (q, \neg p, \emptyset)\rangle$ *or* $o = \langle(r, r, \emptyset), (s, \neg r, \emptyset)\rangle$ *will be a subobservation and by Left Monotony/Right Monotony the core belief must consequently either entail* $\blacktriangle_\vee(\langle(p, p, \emptyset), (q, \neg p, \emptyset)\rangle) = q \to \neg p$ *or* $\blacktriangle_\vee(\langle(r, r, \emptyset), (s, \neg r, \emptyset)\rangle) = s \to \neg r$. $(q \to \neg p) \vee (s \to \neg r)$ *entails neither, so it cannot be an acceptable core belief.*

Note that reasoning only with $\blacktriangle'' = \bigvee\{\blacktriangle \mid Cn(\blacktriangle) = Cn(\blacktriangle_\vee(o')) \cap L(o), o' \in O'\}$ leaves us with very weak (but safe) information about the agent's core belief. Any formula inconsistent with $\blacktriangle''$ will be rejected — but again not every formula consistent with $\blacktriangle''$ is necessarily accepted by the agent. $(q \to \neg p) \vee (s \to \neg r)$ is consistent with both $p \wedge q$ and $r \wedge s$. However, we saw that at least one of the two *must* be inconsistent with the agent's core belief. As a consequence reasoning about an agent while not knowing the positions of possible

[10] Applying proposition 3.26, we can further restrict this set as there need not be more intermediate inputs at one position than there are recorded inputs following that position.

intermediate inputs will require a large computational effort. All the observations in O' and their respective explanations may have to be taken into account. Conclusions about beliefs and non-beliefs can only be safe if they are safe for every observation in O'.

3.3.4 summary

In this section, we investigated the case where the assumption that the observation contains an entry for every input received is dropped. We have shown that some observations can be explained *only* when assuming that intermediate inputs have occurred. These are revision inputs received by $\mathcal{A}$ that have not been recorded in the observation. First, we investigated the case that number and positions of the intermediate inputs were fixed. In this case, the problem can be reduced to partially known inputs, so that the results of Section 3.2 apply. Then we sketched procedures for drawing conclusions about what $\mathcal{A}$'s core belief must entail if this information is not available to the observing agent. In principle, we have to consider all possible observations obtained from allowing intermediate inputs at all positions compatible with the information about $\mathcal{A}$.

One of the key methods for dealing with unknown logical content and unknown inputs in this chapter was extending the language $L(o)$ — o being a parametrised observation possibly missing some inputs. In addition to the problems of choosing an explanation for a regular observation, which we investigated in Chapter 2, here we have to take into account that the unknown subformulae could be instantiated with arbitrary formulae. As it is not feasible to consider all possible instantiations, we proposed to replace each unknown subformula by a new variable, calculate the rational explanation of the observation thus obtained and reason about the core belief and the belief trace restricted to $L(o)$. Hypotheses formed based on this method can be tested using hypothetical reasoning.

Extending the language has side effects. Example 3.16 can also be adapted to show that adding variables not contained in $L(o)$ can lead to an explanation of the observation that would not exist if the intermediate input was restricted to that language. It is open whether there is a feasible way to construct an explanation which is restricted to $L(o)$.

Example 3.21 indicated that allowing more intermediate inputs can lead to weaker explaining core beliefs. When comparing different solutions it is not obvious how to trade off between a weak core belief and few intermediate inputs. The decision will have to be based on the application setting. However, our claim is that if intermediate inputs have to be allowed, we should always assume the maximal number. This will yield the weakest core, keeping us on the safe side as to conclusions about which formulae are rejected by the observed agent. The number of intermediate inputs should be limited only if there is sufficient reason.

Chapter 4

Some Variations and Extensions

In Chapters 2 and 3, we introduced a belief revision framework an observed agent is assumed to employ and presented results on what can be said about $\mathcal{A}$ based on a given (parametrised) observation possibly missing entries for some of the inputs received. In this chapter, we want to apply the results obtained there in slightly different settings. The first sections will assume the same belief revision framework. However, they will deal with reasoning (i) about several agents at once, (ii) about a single agent taking into account several observations, (iii) about oneself and (iv) using extended versions of observations. The last two sections introduce two slight variations of the belief revision framework and investigate how the previous result can be applied to draw conclusions about agents employing those.

4.1 The multi-agent case

In general, the observing agent will share the environment with more than one agent. When reasoning about m other agents rather than a single one, we propose to apply the methods illustrated so far for each agent individually. That is, instead of one there are m observations which may or may not have been obtained independently. As the exact structure and relations of the different observations heavily depend on the application setting, we will not start a detailed discussion of this case. However, we will illustrate one example where the observations are related: a very simplistic dialogue setting.

If two agents are communicating, beliefs $\mathcal{A}_1$ expresses can be seen as revision inputs $\mathcal{A}_2$ receives. Consider the dialogue $\mathcal{A}_1 : p$, $\mathcal{A}_2 : q$, $\mathcal{A}_1 : \neg p$. The next table depicts the same dialogue in terms of which revision inputs are received by which agent.

agent	inputs received	
$\mathcal{A}_1$:	q	
$\mathcal{A}_2$:	p	$\neg p$

The two observations $o_1 = \langle(\top, p, \emptyset), (q, \neg p, \emptyset)\rangle$ and $o_2 = \langle(p, q, \emptyset), (\neg p, \top, \emptyset)\rangle$ formalise the dialogue. As it started with an utterance by $\mathcal{A}_1$, this can be seen as a belief not triggered by any input or equivalently by a tautologous one. $\mathcal{A}_1$'s expressing p is the first input $\mathcal{A}_2$ receives and it must have believed q upon receiving p as otherwise this could not have been $\mathcal{A}_2$'s reply. And so on.

The rational explanation for o_1 is $[(q \rightarrow \neg p, p \wedge \neg q), \top]$ and for o_2 it is $[(p \rightarrow q), \top]$. The results of Chapter 2 now allow us to reason about each agent individually. There is not much we can safely conclude about the two agents, the major insight being that $\mathcal{A}_1$ believed $\neg q$ initially but accepted the input q upon receiving it. More interesting, because we dealt with individual agents before, is what these explanations mean for the interaction between the two agents. How could the dialogue go on? Will they agree at some point? Is that dependent on which beliefs they openly express?[1]

For the sake of argument assume that the agents' initial epistemic states were indeed equivalent with what the rational explanation algorithm calculated. The first thing to note is that both agents will accept and believe whatever the other agent tells. This is because the core beliefs are tautologies and consequently there is no prejudice against any formula. The following table depicts the dialogue and includes the complete information about the initial epistemic state as revision inputs are appended to the sequence ρ_i in the state $[\rho_i, \blacktriangle]$ of $\mathcal{A}_i$. Note that the core belief is irrelevant here, as it is tautologous for both agents.

agent	initial ρ_i	inputs received	
$\mathcal{A}_1$:	$q \rightarrow \neg p, p \wedge \neg q$	q	
$\mathcal{A}_2$:	$p \rightarrow q$	p	$\neg p$

$\mathcal{A}_2$'s beliefs at this point of the dialogue are $Cn(\neg p)$ and $\mathcal{A}_1$'s beliefs are $Cn(q \wedge \neg p)$. Note, that we still assume that agents are sincere, that is, they will only express beliefs they hold.[2] The reader is invited to confirm that nothing $\mathcal{A}_2$ will say in the future is going to change $\mathcal{A}_1$'s mind. $\mathcal{A}_2$ can possibly expand its beliefs by q in case $\mathcal{A}_1$ expresses the belief in a formula entailing $\neg p \rightarrow q$. q has been (temporarily) forgotten by $\mathcal{A}_2$ as the reason p for its being believed is gone. So basically, the two agents have already reached an agreement.

To see that this need not always be the case, assume that the agent's initial epistemic states are in fact $[(q \rightarrow \neg p, p \wedge \neg q), \top]$ and $[(p \leftrightarrow q), \top]$. Only the epistemic state of $\mathcal{A}_2$ is slightly modified, containing an equivalence rather than an implication. These initial states are compatible with the above dialogue. Here are three possible continuations of the dialogue:

[1] This line of investigation may be a bit off-topic with respect to the thesis but it also illustrates an interesting point of future research.

[2] The assumed belief revision framework cannot deal with non-beliefs as revision inputs.

agent	initial ρ_i	inputs received						
$\mathcal{A}_1$:	$q \rightarrow \neg p, p \wedge \neg q$	q		$\neg q$		q		$\ldots$
$\mathcal{A}_2$:	$p \leftrightarrow q$	p	$\neg p$		p		$\neg p$	
$\mathcal{A}_1$:	$q \rightarrow \neg p, p \wedge \neg q$	q		$\neg q$		$\ldots$		
$\mathcal{A}_2$:	$p \leftrightarrow q$	p	$\neg p$		$\neg q \wedge p$			
$\mathcal{A}_1$:	$q \rightarrow \neg p, p \wedge \neg q$	q		$\neg p \wedge \neg q$				
$\mathcal{A}_2$:	$p \leftrightarrow q$	p	$\neg p$		$\ldots$			

In the first case, the two agents could go on forever; changing their mind, saying the same things over and over again. This is because $\mathcal{A}_1$ prefers believing $p \leftrightarrow \neg q$ unless being told otherwise and $\mathcal{A}_2$ may never do so. In the second case, both their belief sets would be $Cn(\neg q \wedge p)$ and in the third one $Cn(\neg p \wedge \neg q)$. Because both agents believe whatever the other says, the agent revealing its entire beliefs first will "win". This will be different if the core beliefs are not tautologies.

A belief expressed by an agent may be based on a formula it is not very sure about (as it appears very early in the sequence and consequently has a low priority). However, the agent receiving that belief as input treats it as the most important revision input so far. Note that in case the second agent now simply repeats the formula just received, it suddenly gets a high priority even for the first agent. Weak beliefs may get amplified by the dialogue. Unfortunately, this property is more realistic than desirable.

4.2 Observations with respect to the same initial state

One of the motivating scenarios for our work was accessing expert knowledge. An observation o as used up to now represents one sequence of revision inputs received by one agent. In the expert scenario o is interpreted as the expert $\mathcal{A}$ reasoning about one single case based on information received over time. Of course, it is very unlikely that all the expert's knowledge manifests itself in one single case. So observing $\mathcal{A}$ reasoning about a (large) number of cases seems necessary. If we assume that the expert does not learn from case to case but is independently applying its knowledge to each of them, then all observations must be explainable by the same initial epistemic state. However, this can be easily dealt with in our framework. We just have to adapt the rational explanation algorithm by processing the union of all the positive conditionals and the union of all the negative conditionals calculated from all the observations rather than those from a single observation. Each observation translates into conditional beliefs in the (same) initial epistemic state (Proposition 2.24, Definition

2.42), so many observations simply translate into even more conditional beliefs in the same state.

The assumption that the expert does not learn from case to case is debatable. However, we believe it to be a good enough approximation. Faced with a large number of *standard* cases, reflections about the correctness of the background knowledge are not necessary. A learning process, i.e., actually changing the background knowledge, is likely to occur if the expert is faced with many exceptional cases — few or even a single one may not immediately trigger changes.

The expert scenario is not the only one where several observations with respect to the same state make sense. Think of a population of agents starting in the same state. For example, we might be interested in the background knowledge encoded in a software agent. We make several copies of that agent and place it in different settings. That is, each of the identical agents receives different revision inputs. Different observations with respect to the same initial epistemic state are thus obtained. Those can be processed as described above. It goes without saying that all the observations considered in this section can be parametrised ones and that intermediate inputs may have to be taken into account.

So, assuming we are given n observations o_i with respect to the same initial state and a core belief $\blacktriangle$, we define the sets of positive and negative conditionals $\mathcal{C}_{\blacktriangle}(o)$ and $\mathcal{N}_{\blacktriangle}(o)$ by $\mathcal{C}_{\blacktriangle}(o) = \bigcup_{1 \leq i \leq n} \mathcal{C}_{\blacktriangle}(o_i)$ and $\mathcal{N}_{\blacktriangle}(o) = \bigcup_{1 \leq i \leq n} \mathcal{N}_{\blacktriangle}(o_i)$ where all $\mathcal{C}_{\blacktriangle}(o_i)$ and $\mathcal{N}_{\blacktriangle}(o_i)$ are as in Definition 2.42. The rational explanation algorithm just needs to be adapted accordingly.

In this scenario, one further limitation concerning the correctness of the belief trace applies. Given a single observation, the antecedents $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ of the corresponding conditionals are calculated from a single sequence $(\varphi_1, \ldots, \varphi_n)$ of inputs. So if we consider *any* two conditionals, one antecedent is constructed from a prefix of the sequence of inputs the other conditional is constructed from. This is a property needed to relate those antecedents using Proposition 2.15. This result in turn is used for showing that the rational prefix is preferable to other sequences with respect to $\preceq_1$ and $\preceq_2$ (Propositions 2.55 and 2.56). As a consequence, these results do not automatically carry over to the current setting. We do not consider this very problematic, as these are very weak criteria, anyway.

All other results also hold for this less restricted form of antecedents of conditionals obtained from several observations. The proofs do not depend on the particular structure of the antecedents. This also means that the more observations with respect to the same initial state we have the better the chances for calculating the right core. This is because more conditional beliefs are actually given and not inferred. Hypothetical reasoning is also still possible. Note that the question of $\mathcal{A}$ given an alternative sequence of revision inputs is particularly interesting. This question allows conclusion about what the expert would have

thought given a different case or how an agent would have reasoned being placed in a different situation.

From the point of view of computational complexity, it does not matter whether we consider a single or several observations. Testing whether an epistemic state explains a set of observations and testing whether a core belief is acceptable for all observations in a set are also Δ_2^P-complete.

Finally, we want to remark that allowing many observations allows us to represent of any information about a conditional belief base in this framework. Positive conditionals $\varphi \Rightarrow \theta$ translate into the observation $\langle(\varphi, \theta, \emptyset)\rangle$ and negative conditionals $\varphi \Rightarrow \delta$ into $\langle(\varphi, \top, \{\delta\})\rangle$. As a consequence, by using this translation we can recover the full power of the rational closure [14, 54] as we do not restrict the antecedents to the particular structure. The interpretation of the difference between the rational closure of the conditionals obtained from these observations (using the core belief $\top$) and the rational explanation for this set of observations (which may return a different core and hence modify the conditionals) is an open question.

4.3 Self-observations

The motivation of our work was to reason about an observed agent. However, we may wonder whether reasoning about *oneself* in this setting makes sense. At first sight the answer is negative. It is much more efficient with respect to both space and computational cost to simply store ones epistemic state $[\rho, \blacktriangle]$ rather than keep an observation in order to reason about earlier beliefs. It is true that for very long revision histories, an observation starting at some later point may take up less space, but the conclusions based on it will be much less reliable than those taken from an equally big suffix of ρ. From Proposition 2.3 and Definition 2.1 it follows immediately that $f(\sigma_p \cdot \sigma_s) \vdash f(\sigma_s)$. That is, conclusions about beliefs based on the suffix of a sequence are undebatable — for non-beliefs this is different of course. Further, an agent should be able to remember its own core belief. So compared to keeping an observation, simply forgetting early revision inputs is the better alternative.

But we may keep observations of ourselves in order to keep track of what *other agents* may conclude about *us*. This is reasonable in settings where we want to keep certain beliefs secret from others. Before making a response we check whether this would give away more than we want. It may also make sense when we want another agent to know θ but we cannot tell it directly. In this case, we look for a response that allows the other agent to infer θ based on its observation of us. Again, the observation (even if it is about us) is used to reason about other agents.

Of course there are a number of assumptions. The observation we keep of ourselves must correspond to the one the other agent makes of us. It must employ the same methodology for reasoning given the observation. This is a very problematic point. Will the other agent act based on a conclusion that was not verified using hypothetical reasoning? In this case, we will have to be even more careful. In a sense, the other agent must be neither smarter than us (and thus be able to draw conclusions we cannot come up with) nor must it be less smart (and thus possibly draw correct conclusions but for the wrong reason). Basically, all the assumptions we made so far must be applied in a reverse direction. The other agent must assume *us* to employ the revision framework from Section 2.2 and so on.

It should be clear that it does not suffice to keep one self-observation. We will need one for every other agent (we are interested in). This is because not all agents have access to the same information about us. Again, they may leave the scene for a while (at different times) and consequently do not record all revision inputs received by us and beliefs and non-beliefs we express. The self-observing agent $\mathcal{A}$ should not forget, though, that others tend to become suspicious if coming up with a response takes more time than expected.

4.4 Infinite observations

In Section 2.8.2, we briefly considered observations of infinite length to illustrate that even they cannot guarantee that the correct core belief can be identified. Here, we want to illustrate how observations which are infinite even with respect to beliefs and non-beliefs can be dealt with. We assume them to be correct and complete with respect to the revision inputs recorded, i.e., the observation contains exactly those inputs actually received by $\mathcal{A}$ and we do not allow intermediate inputs. Also we do not allow unknown subformulae.

Note that a belief θ can be interpreted as the conjunction of a finite set of beliefs, i.e., $\theta = \bigwedge B$ where B is a finite set of formulae with $\bigwedge \emptyset = \top$. In this sense an observation can be equivalently given as follows $o = \langle(\varphi_1, B_1, D_1), \ldots (\varphi_n, B_n, D_n)\rangle$ where each B_i and D_i is a finite set of formulae. The assumed revision framework requires the revision inputs to be single formulae.

Now assume that o is an observation (possibly of infinite length) where all B_i and D_i are allowed to be infinite sets. Of course, formulae are still assumed to be finite objects. Left Monotony, Right Monotony and Proposition 2.68 can again be used to reason about o. The first two say that making an observation longer (at the ends) can only cause the weakest acceptable core to get stronger, the last one yields the same result for making the information about beliefs and non-beliefs more specific by adding (non-)beliefs. As $B \subseteq B'$ implies $\bigwedge B' \vdash \bigwedge B$, we can use Proposition 2.68 also when talking about sets of beliefs.

The idea for handling an infinite observation $o = \langle(\varphi_1, B_1, D_1), \dots (\varphi_n, B_n, D_n), \dots \rangle$ is now straightforward. As in Section 2.8.2, we consider an infinite sequence $(o_1, o_2, \dots)$ of finite observations that approaches o. If we pose the following restrictions on that sequence, we can use any observation o_i in order to draw conclusions about the observed agent's core belief $\blacktriangle$. By Left Monotony, Right Monotony and Proposition 2.68, any formula entailed by $\blacktriangle_\vee(o_i)$ will be entailed by $\blacktriangle$.

- If $o = \langle(\varphi_1, B_1, D_1), \dots (\varphi_n, B_n, D_n)\rangle$ has finite length then, for appropriate B^i_j and D^i_j, $o_i = \langle(\varphi_1, B^i_1, D^i_1), \dots (\varphi_n, B^i_n, D^i_n)\rangle$. Otherwise $o = \langle(\varphi_1, B_1, D_1), \dots (\varphi_n, B_n, D_n), \dots \rangle$ and $o_i = \langle(\varphi_1, B^i_1, D^i_1), \dots (\varphi_n, B^i_i, D^i_i)\rangle$, for appropriate B^i_j and D^i_j.
- If B_j is finite then $B^i_j = B_j$ for all o_i; analogously for finite D_j.
- If B_j is infinite then B^i_j is finite and $B^k_j \subset B^i_j \subset B_j$ for all $k < i$; analogously for infinite D_j.

We remark that this is only *one* suitable set of conditions for $(o_1, o_2, \dots)$ that ensures applicability of the propositions. The first condition requires that we use all revision inputs in case the infinite observation o has finite length and that we keep adding revision inputs at the end otherwise. The second condition says that finite sets of beliefs and non-beliefs from o are to be used unchanged in all o_i. That is, if there is only finite information we use it completely in every step. The last condition expresses that for every o_{i+1}, the information concerning infinite sets of beliefs or non-beliefs is more specific than that in o_i. New formulae are added at each step. Note that $L(o_i)$ is finite although $L(o)$ may be infinite. As a consequence, each o_i is an observation that can be dealt with according to Chapter 2. As o is infinite we cannot expect to find an actual explanation for it. We can only approximate the core belief of $\mathcal{A}$ and reach it in the limit as $\blacktriangle_\vee(o_{i+1}) \vdash \blacktriangle_\vee(o_i)$ for all i.

4.5 Graded observations

Observations as considered up to now were assumed to be correct. In particular, we required that the recorded beliefs and non-beliefs reflect the agent's true set of beliefs at each point in time during the observation. They are all equally reliable because they are certain. In personal communication, Didier Dubois suggested incorporating reliability/confidence information into an observation. This would allow us to draw sceptical conclusions when only considering the most reliable information and more credulous ones when also considering the less reliable information. As the beliefs and non-beliefs recorded in the observation may

have been obtained more or less indirectly, attaching reliability degrees to the individual formulae is a sensible thing to do.

In this section, we will illustrate one possible way of reasoning from such observations. We will keep the assumption that the recorded revision inputs are correct. That is, we restrict ourselves to reliability information for beliefs and non-beliefs. For convenience we use the notation introduced in the last section and assume beliefs to be given as sets rather than as single formulae. This allows us to attach to a reliability label to every belief and every non-belief.

Reliability information can be interpreted and used in many ways. One fundamental question is whether less reliable information is to be interpreted as *alternative* to more reliable one or as additional information. Depending on the answer, explanations and conclusions drawn have to be interpreted differently. In the first case it is possible to express that a belief in p is very likely but there is a small chance that the agent actually believed $\neg p$. But this also means that we basically deal with several observations that may contain contradicting information and there is generally no single epistemic state explaining all of them — there may not even be a single acceptable core belief. Recall that the belief set of an agent is inconsistent if and only if its core belief is. So there cannot be an explanation where $\mathcal{A}$ believes p and $\neg p$ while having a consistent core. Reasoning about the agent would then have to be done by considering what is believed according to all belief traces constructed from the explanations of observations with a certain degree of reliability. In order to draw non-trivial conclusions it is necessary to construct from an observation containing reliability information a set of observations that can be handled using the rational explanation methodology. This is a difficult task as it is not obvious which beliefs and non-beliefs to leave out in which observation. Recall that we do not construct several observations with respect to the same initial epistemic state but observations that will have different explanations. That is, if we leave out too much the conclusions we will draw will be extremely weak. The general method will probably be to look for maximally consistent subsets of beliefs incorporating the reliability information for choosing which formulae to leave out.

Here we will consider only the second interpretation and assume less reliable information to be additional to the more reliable one. That is, we want to express that $\mathcal{A}$ believes θ with a high reliability and that with less certainty it also believes θ'. For simplicity, we consider reliability to be represented by natural numbers, 0 representing the highest reliability. Now, each belief and non-belief is labelled with a number. That is, instead of formulae θ, the sets of beliefs B_i in an observation o contain pairs of the form (θ, k). Analogously, the sets of non-beliefs D_i in o contain pairs of the form (δ, k). We will call such an o a graded observation. We assume that the labels are comparable, i.e., that k represents the same degree of reliability

everywhere in o. When reasoning about $\mathcal{A}$ we now also specify a threshold t indicating that we do not want to consider formulae whose label is greater than t but consider all formulae whose label is less or equal to t. The greater t the more formulae are taken into account and by Proposition 2.68 we know that the weakest acceptable core belief can only become stronger.

It is important to note that this notion of reliability and safe conclusions with respect to hypothetical reasoning are not identical. Using 0 as threshold does not mean that the conclusions drawn are safe. However, the two notions are not unrelated. Let o_1 be an observation obtained from a graded observation o by using a threshold t_1 and o_2 using threshold t_2 such that $t_1 < t_2$. This means o_2 may contain information that is less reliable. From Proposition 2.68 it follows that any explanation for o_2 also explains o_1. This entails that a conclusion that is not safe for o_2 is also not safe for o_1. This is because any explanation found as a counterexample to the conjecture for o_2 would be one for o_1 as well. The revision inputs in o_1 and o_2 are the same and hence the belief traces yielded by that explanation are equivalent. Only the beliefs and non-beliefs *recorded* in the two observations differ. Conversely, a safe conclusion for o_1 is also safe for o_2. If we found an explanation contradicting the conjecture with respect to o_2 it would be a counterexample for o_1 as well, but by the definition of a safe conclusion such an explanation does not exist. The following example also illustrates that conclusions that are safe when also using less reliable information need not be safe when using only more reliable information. This is what we would expect. The additional information may rule out possibilities — but not reliably.

Example 4.1. *Consider* $o = \langle (p, \{(\neg q, 0), (r, 1)\}, \emptyset), (q, \{(q, 0), (s, 2)\}, \{(p, 1)\})\rangle$. *This graded observation expresses that the observed agent received the revision inputs* p *and then* q. *We are certain that after receiving* p *it believed* $\neg q$ *and less certain that it also believed* r. q *must have been believed upon receiving it as revision input but* p *may not be among the beliefs. Further, there is remote possibility that* s *is also believed.*

Now let $t_1 = 0$ *and* $t_2 = 1$, *so* $o_1 = \langle (p, \neg q, \emptyset), (q, q, \emptyset)\rangle$ *as we consider only formulae whose label is* 0 *and* $o_2 = \langle (p, \neg q \wedge r, \emptyset), (q, q, \{p\})\rangle$ *ignoring formulae whose label is greater than* 1. *The rational explanation for* o_1 *is* $[(p \rightarrow \neg q), \top]$ *and for* o_2 *it is* $[(p \rightarrow r), p \rightarrow \neg q]$.

From the rational explanation for o_1 *we can hypothesise that* $p \rightarrow \neg q$ *is believed initially and this conclusion can be verified using hypothetical reasoning. It is also safe for* o_2. *Note that this formula has to be entailed by any core belief acceptable for* o_2.

From the rational explanation for o_2 *we can hypothesise that the revision input* p *is in fact accepted. This conclusion cannot be verified. In fact, the agent's initial state may be* $[(s, r \wedge \neg q), \neg p]$. *So this conclusion is not safe for* o_1 *either.*

Note that using o_2 it is safe to conclude that p is not believed after having received q, as it is impossible that the agent both believes and does not believe a formula at the same time. But this conclusion is not safe for o_1, the rational explanation for o_1 being a counterexample. Here the agent keeps believing p.

4.6 Revision using priority information

Whereas the last section dealt with the reliability of beliefs and non-beliefs recorded in the observation, this section is about the reliability of the revision inputs. The interpretation of an epistemic state $[\rho, \blacktriangle]$ and the revision function $*$ we used up to this point equate recency of a revision input with reliability. With the exception of the core belief, which is more important than any input ever received, inputs received later take precedence over inputs received earlier because they are appended to the end of the sequence ρ. However, this temporal interpretation is not inherent to our framework — more precisely to the method we assume the agent to use for calculating its beliefs in a given state. All the function f assumes is that there is a total order on a set of formulae. This set is processed starting with the most important one (the core belief) and then proceeding with the next most important one and so on.

That is, in fact it is (almost) irrelevant when in the past the formula was received, all that matters is its position in the sequence representing the total order. The agent still receives formulae in a certain temporal order which has to be captured in the observation o and which has to be taken into account when reasoning about the agent's beliefs at a particular point in time. After all, formulae that have not yet been received cannot be used for calculating the set of beliefs. However, the above considerations imply that we do not have to commit ourselves to a revision framework that assigns the highest priority to the formula received most recently.

Keeping the representation of an agent fixed, we can easily imagine revision frameworks where additional information allows to specify *where* in the sequence ρ the new input φ is to be inserted in order to arrive at the resulting epistemic state. If this additional information gets recorded in the observation as well and can be interpreted without ambiguity, the methods presented in Chapters 2 and 3 still allow us to reason about an observed agent. What we vaguely call additional information may include information about the source, its reliability (with respect to the topic of φ), supporting arguments etc.

To illustrate our point, we will present a very simple extension of the revision framework assumed up to now and show how the results presented before can be used to reason about an agent employing this extended framework. The epistemic state $[\rho, \blacktriangle]$ of an agent is defined

as before. The revision input comes with an *index* indicating at which point in the sequence the new formula is to be inserted.[3] Lower indexes mean higher priority, i.e., the index 1 corresponds to appending the formula to the end of the sequence. If necessary, the sequence is extended with a number of tautologies in order to allow insertion at the correct position. That is, an index i says that revision input φ has to be the i^{th} element from the back of the modified ρ.

Definition 4.2. *Let* $[(\varphi_1, \ldots, \varphi_n), \blacktriangle]$ *be an epistemic state,* φ *a formula,* $k > 0$ *and* $\lambda_j \equiv \top$ *for all* $j > 0$*. Then*

$$[(\varphi_1, \ldots, \varphi_n), \blacktriangle] *_I (\varphi, k) = \begin{cases} [(\varphi_1, \ldots, \varphi_n, \varphi), \blacktriangle] & , k = 1 \\ [(\varphi_1, \ldots, \varphi_{n-1}, \varphi, \varphi_n), \blacktriangle] & , k = 2 \\ [(\varphi_1, \ldots, \varphi_{n-k+1}, \varphi, \varphi_{n-k+2}, \ldots, \varphi_n), \blacktriangle] & , 2 < k \leq n \\ [(\varphi, \varphi_1, \ldots, \varphi_n), \blacktriangle] & , k = n+1 \\ [(\varphi, \lambda_{k-n-1}, \ldots, \lambda_1, \varphi_1, \ldots, \varphi_n), \blacktriangle] & , k > n+1 \end{cases}$$

As before $Bel([\rho, \blacktriangle]) = Cn(f(\rho \cdot \blacktriangle))$.

If the chosen index is always 1 then $*_I$ coincides with the revision operator $*$ we considered before. We assume that the index is given directly. Determining it can be seen as a pre-processing step the agent carries out before the actual revision takes place. For example, the index could be the value of a function mapping the current epistemic state, the revision input, and the additional information indicated above to a natural number.

Note that even if two revision inputs have the same index, the one received later demotes the one received earlier. This is because the framework cannot capture that two individual formulae have the same priority — the conjunction of two formulae is not the same as two separate ones. Also a revision input with a low index may have been pushed down over time so that a later input with higher index might in the end have a higher priority. This shows that the index of a formula is not an absolute measure for its reliability. There is still some temporal interpretation involved. The following example illustrates revision in the current framework.

Example 4.3.

$$[(p, p \rightarrow q), \top] *_I (\neg q, 2) = [(p, \neg q, p \rightarrow q), \top]$$
$$[(p, \neg q, p \rightarrow q), \top] *_I (r \vee q, 6) = [(r \vee q, \top, \top, p, \neg q, p \rightarrow q), \top]$$
$$[(r \vee q, \top, \top, p, \neg q, p \rightarrow q), \top] *_I (s, 1) = [(r \vee q, \top, \top, p, \neg q, p \rightarrow q, s), \top]$$

[3] In this section, the meaning of the word index is slightly different to that in the context of the rational prefix construction. There it corresponded to the point in time in the observation that gave rise to a conditional, i.e., $f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \Rightarrow \theta_i$ (respectively $f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \Rightarrow \delta$) has the index i.

For simplicity we further assume that the index is given in the observation as well. That is, instead of the revision inputs φ an observation will contain entries of the form (φ, k). The observation $\langle((p,2),q,\emptyset)\rangle$ expresses that after the agent $\mathcal{A}$ received p, placing it in the last but one position in the sequence ρ of its epistemic state $[\rho, \blacktriangle]$, $\mathcal{A}$ believed q. We also assume that the observation is complete in the sense that for the time of the observation there is a an entry for every revision input received by $\mathcal{A}$. We cannot directly apply the rational explanation construction as it assumes the revision operator to be $*$ and not $*_I$. This means that the observation cannot easily be translated into conditional beliefs in the initial epistemic state as some of the elements of ρ (which we do not know) might have to appear in the antecedents of the conditionals. How can we approach this problem?

We will make use of the ideas developed in the previous chapter, intermediate inputs in particular. We divide the sequence ρ of the initial epistemic state into two subsequences. The first one is the (largest) prefix that is not affected by the revision steps recorded in the observation o simply because the formulae in this prefix are less important than any revision input from o. This prefix ρ_p will again be calculated using the rational prefix construction. The second one ρ_s is the suffix of ρ in which the revision inputs recorded in o get inserted — the resulting suffix being ρ_s^i after the i^{th} revision input. The formulae contained in ρ_s will be *treated like* intermediate inputs although they are not intermediate inputs in the sense of Section 3.3.

Proposition 4.4. *Let (φ_i, k_i), $1 \leq i \leq n$, be a set of n revision inputs with corresponding indexes. Further let $\blacktriangle$ be a core belief and ρ_p, ρ_s and ρ_s^i, $1 \leq i \leq n$, sequences such that $[\rho_p \cdot \rho_s, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i) = \left[\rho_p \cdot \rho_s^i, \blacktriangle\right]$ for all $1 \leq i \leq n$.*

Then $Bel([\rho_p \cdot \rho_s, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i)) = Bel([\rho_p, \blacktriangle] *_I (f\left(\rho_s^i \cdot \blacktriangle\right), 1))$ for all $1 \leq i \leq n$.

$[\rho, \blacktriangle]$ is the initial epistemic state. The revision inputs φ_i are received in the order induced by the subscript i and they are inserted into the sequence of the current epistemic state according to the index k_i. Note that the ρ_s^i are uniquely determined by ρ_s and the revision inputs $\varphi_1, \ldots, \varphi_i$ and their respective indexes $k_1, \ldots, k_i$. As noted above, a prefix ρ_p of ρ may not be affected by the sequence of revisions. Changes occur only in the suffix ρ_s. The above proposition now tells us that if we know ρ_s, then, from the point of view of the agent's beliefs, the sequence of revisions can again be translated into a set of single revision steps with respect to the same state — using the original revision function $*$. This allows us to translate an observation into conditional beliefs in the state $[\rho_p, \blacktriangle]$ and consequently we can use the rational explanation construction to calculate a state satisfying the conditionals.

However, ρ_s is unknown to the observing agent — it contains a number of unknown formulae. We use the intermediate inputs idea and instantiate them with new variables. The question

is: How long do we have to assume ρ_s to be so that Proposition 4.4 is applicable, i.e., that only ρ_s is affected by the revision steps? Basically, we have to make sure that no revision step makes it necessary to extend ρ_s by tautologies (cf. Definition 4.2). We could simply use the greatest index k_j of any revision input appearing in the observation. This would mean that ρ_s has the length k_j and Proposition 4.4 would be trivially applicable. However, it would also mean to introduce k_j new variables and from previous remarks it should be clear that this should be avoided if possible. As each new variable doubles the number of worlds, their number should be kept low. Can we do better?

Each revision input recorded in o increases the length of the suffix by one. After inserting the $(i-1)^{\text{th}}$ revision input the length is at least $(i-1)$. Consequently, if $k_i - i \leq 0$ then the i^{th} revision input φ_i with index k_i can be inserted without an extension with tautologies being necessary — in other words without affecting ρ_p, which is necessary for Proposition 4.4 being applicable. If the difference is bigger, then ρ_s must have been accordingly longer.[4] This yields $\max\limits_{1\leq i\leq n}\{k_i - i\}$ as the minimal length for ρ_s given an observation $o = \langle((\varphi_1, k_1), \theta_1, D_1), \ldots, ((\varphi_n, k_n), \theta_n, D_n)\rangle$.

Example 4.5. *Consider the initial epistemic state* $[(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6), \blacktriangle]$ *and the observation* $o = \langle((\varphi_1, 2), \theta_1, D_1), ((\varphi_2, 4), \theta_2, D_2), ((\varphi_3, 1), \theta_3, D_3), ((\varphi_4, 6), \theta_4, D_4)\rangle$.

$\max\limits_{1\leq i\leq n}\{k_i - i\} = \max\{2-1, 4-2, 1-3, 6-4\} = 2$. *The epistemic state resulting from the revision steps recorded in* o *is* $[(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \varphi_4, \varphi_2, \alpha_5, \varphi_1, \alpha_6, \varphi_3), \blacktriangle]$ *and indeed only the suffix* (α_5, α_6) *of length* 2 *has been affected by the revision, the prefix* $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ *remained unchanged.*

All we have to do now is redefine the set of positive and negative conditionals given a core $\blacktriangle$ and an observation $o = \langle((\varphi_1, k_1), \theta_1, D_1), \ldots, ((\varphi_n, k_n), \theta_n, D_n)\rangle$. Let $k = \max\limits_{1\leq i\leq n}\{k_i - i\}$, that is, we will assume the suffix of the initial epistemic state to contain k unknown formulae which we instantiate with $x_1, \ldots, x_k$ according to the intermediate inputs idea. Hence we set $\rho_s = (x_1, \ldots, x_k)$. Further we define ρ_s^i such that $\left[\rho_s^i, \blacktriangle\right] = [\rho_s, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i)$ for all $1 \leq i \leq n$ and thereby get $\mathcal{C}_{\blacktriangle}(o) = \{f\left(\rho_s^i \cdot \blacktriangle\right) \Rightarrow \theta_i \mid 1 \leq i \leq n\}$ as the set of positive conditional beliefs and $\mathcal{N}_{\blacktriangle}(o) = \{f\left(\rho_s^i \cdot \blacktriangle\right) \Rightarrow \delta \mid 1 \leq i \leq n \wedge \delta \in D_i\}$ as the set of negative

[4] In fact it need not. But then the entire sequence ρ of the agent's epistemic state $[\rho, \blacktriangle]$ is affected by the insertion of revision inputs and that only the trivial prefix, i.e., the empty sequence $\rho_p = ()$, remains unchanged. For such ρ it is easy to see that for all i

$$Bel([\rho, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i)) = Bel([(\top, \ldots, \top) \cdot \rho, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i)).$$

This is because the relative position of the elements of ρ and the φ_i is not affected by the tautologies appended before ρ and these tautologies have no semantic impact on the formula collected by f. Hence, we can safely assume ρ_s to be long enough for Proposition 4.4 to be applicable. Some of the unknown formulae may simply be tautologies.

conditionals. Using these definitions for the conditionals, Algorithm 1 can be used except that the returned epistemic state must be $[\rho \cdot \rho_s, \blacktriangle]$ instead of $[\rho, \blacktriangle]$. Again, this is because we fixed the suffix of the initial epistemic state using intermediate inputs. From this initial epistemic state we can calculate the belief trace and calculate potential beliefs (in $L(o)$) at each point in time.

Although the results from Chapter 3 concerning the weakest core belief for parametrised observations cannot be applied directly to the current setting, the conclusions we draw about the core belief are correct. This is because in both cases the observation is translated into a set of conditional beliefs in the initial epistemic state. All the components (formulae from the observation, possibly some unknown formulae from ρ_s) are fixed. That is, the proofs showing the correct calculation of the only varying element, which is the core belief, would look very much the same. As in the concluding remarks to Section 3.2, we claim that hopes for the correctness of the belief trace should not be too high. They depend on knowing the correct core to begin with and we already argued that this is not likely. Further, the true instantiation of the elements of ρ_s is unknown, so we cannot be sure whether the revision inputs recorded in o are believed — even if we had the correct core.

Example 4.6. *Consider the observation* $o = \langle((p,1), q, \emptyset), ((\neg q, 3), \neg q, \emptyset), ((r,2), q, \emptyset)\rangle$. *After revising by* p *which is considered very important by the agent, it believes* q. *The agent then accepts the less prioritised input* $\neg q$, *but changes its mind again after learning* r. *We are looking for an epistemic state* $[\rho_p \cdot \rho_s, \blacktriangle]$ *explaining this observation. First we have to determine the length of* ρ_s *such that* ρ_p *is not changed by the recorded revisions. The length is* $\max\{1-1, 3-2, 2-3\}$ *which is* 1. *That is, we assume* $\rho_s = (\chi)$ *to be a sequence containing one unknown input which we will instantiate with the new variable* x. *The epistemic state after all the revision steps recorded in* o *would then be* $[\rho_p \cdot (\neg q, x, r, p), \blacktriangle]$. *Using Proposition 4.4 we can transform this observation into a set of conditional beliefs held by the agent in the initial epistemic state* $[\rho_p, \blacktriangle]$ — *assuming it to use the revision framework of Chapter 2.*

As there is no information about non-beliefs of the agent, there will only be positive conditionals. $f(x, p, \blacktriangle) \Rightarrow q$ *corresponds to the first entry in* o, $f(\neg q, x, p, \blacktriangle) \Rightarrow \neg q$ *to the second and* $f(\neg q, x, r, p, \blacktriangle) \Rightarrow q$ *to the third. We want to remark that these conditionals can also be interpreted as being obtained from several observations with respect to the same initial state. We now apply the rational explanation construction, starting with* $\blacktriangle = \top$.

$$\begin{aligned} \mathcal{C}_0 &= \{f(x, p, \top) \Rightarrow q, f(\neg q, x, p, \top) \Rightarrow \neg q, f(\neg q, x, r, p, \top) \Rightarrow q\} \\ &= \{x \wedge p \Rightarrow q, \neg q \wedge x \wedge p \Rightarrow \neg q, \neg q \wedge x \wedge r \wedge p \Rightarrow q\} \end{aligned}$$

The last two conditionals are p*-exceptional for* $U_0 = \tilde{\mathcal{C}_0}$ *as* $\neg q \wedge x \wedge p$ *is inconsistent with* $x \wedge p \rightarrow q$ *and* $\neg q \wedge x \wedge r \wedge p$ *is inconsistent with* $\neg q \wedge x \wedge r \wedge p \rightarrow q$. *The calculation continues as follows.*

$$\begin{aligned}\mathcal{C}_1 &= \{\neg q \wedge x \wedge p \Rightarrow \neg q, \neg q \wedge x \wedge r \wedge p \Rightarrow q\}\\ \mathcal{C}_2 &= \{\neg q \wedge x \wedge r \wedge p \Rightarrow q\} = \mathcal{C}_3\end{aligned}$$

As $\bigwedge U_2 = \neg p \vee \neg r \vee \neg x \vee q$ *is not a tautology, we have not found an explanation yet and will continue with the modified core belief* $\blacktriangle = \neg p \vee \neg r \vee \neg x \vee q$*. This leads to the following calculation.*

$$\begin{aligned}\mathcal{C}_0 &= \{f(x,p,\blacktriangle) \Rightarrow q, f(\neg q,x,p,\blacktriangle) \Rightarrow \neg q, f(\neg q,x,r,p,\blacktriangle) \Rightarrow q\}\\ &= \{x \wedge p \wedge (\neg r \vee q) \Rightarrow q, \neg q \wedge x \wedge p \wedge \neg r \Rightarrow \neg q, x \wedge r \wedge p \wedge q \Rightarrow q\}\\ \mathcal{C}_1 &= \{\neg q \wedge x \wedge p \wedge \neg r \Rightarrow \neg q\}\\ \mathcal{C}_2 &= \emptyset = \mathcal{C}_3\end{aligned}$$

This tells us that the core belief $\blacktriangle = \neg p \vee \neg r \vee \neg x \vee q$ *works.* $Cn(\blacktriangle) \cap L(o) = Cn(\top)$ *so the entire core belief can be absorbed into the instantiation of the unknown input, that is we use* $x \wedge \blacktriangle$ *instead of* x*. Note that the material counterparts of the last two conditionals in* $\mathcal{C}_1$ *are tautologies and that the material counterpart of the first one is equivalent to* $x \wedge p \wedge \neg r \rightarrow q$*. Using all these simplifications we get* $\rho_p = (x \wedge \blacktriangle \wedge p \wedge \neg r \rightarrow q)$*,* $\rho_s = (x \wedge \blacktriangle)$ *and* o *is thus explained by*

$$[(x \wedge \blacktriangle \wedge p \wedge \neg r \rightarrow q, x \wedge \blacktriangle), \top].$$

We will confirm this by calculating the belief trace. First note that $x \wedge \blacktriangle$ *entails* $\neg p \vee \neg r \vee q$*. So the beliefs in this assumed initial state are* $Cn(x \wedge (p \rightarrow q))$*. After having received the first recorded revision input the epistemic state is* $[(x \wedge \blacktriangle \wedge p \wedge \neg r \rightarrow q, x \wedge \blacktriangle), \top] *_I (p,1) = [(x \wedge \blacktriangle \wedge p \wedge \neg r \rightarrow q, x \wedge \blacktriangle, p), \top]$*, the beliefs in this state being* $Cn(p \wedge x \wedge q)$*.* $\mathcal{A}$*'s epistemic state after the next revision step is* $[(x \wedge \blacktriangle \wedge p \wedge \neg r \rightarrow q, \neg q, x \wedge \blacktriangle, p), \top]$*. Its beliefs are now* $Cn(p \wedge x \wedge \neg q \wedge \neg r)$*. The epistemic state after having received the final input is* $[(x \wedge \blacktriangle \wedge p \wedge \neg r \rightarrow q, \neg q, x \wedge \blacktriangle, r, p), \top]$ *and the beliefs are* $Cn(p \wedge r \wedge x \wedge q)$*. The belief trace — restricted to* $L(o)$ *which is what we care about — consequently is*

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

Finally, we will illustrate two cases for hypothetical reasoning. The belief trace indicates that p *was believed upon having been received. To see if this is necessarily the case we try to find an explanation for the modified observation* $o' = \langle((p,1), q, \{p\}), ((\neg q, 3), \neg q, \emptyset), ((r,2), q, \emptyset)\rangle$*. We will simply give an explaining epistemic state:* $[(q, r \rightarrow q), \neg p]$*. It is not the one calculated by our algorithm. The existence of an explanation* o' *indicates that it is possible that the agent did not believe* p *upon receiving it.*

A second conclusion we draw is that before receiving the final input r*, the agent believes* $\neg r$*. To see if this is necessarily the case we try to find an explanation for the observation* $o' = \langle((p,1), q, \emptyset), ((\neg q, 3), \neg q, \{\neg r\}), ((r,2), q, \emptyset)\rangle$*. However, there is none. After a number of iterations of strengthening the potential core belief, an inconsistent one is constructed. The*

intuitive reason that there cannot be an explanation is that r must have made a difference as the agent changes its mind with respect to q. However, if r was consistent with the beliefs then revising by r would have been a simple expansion of the belief set. In this case it would not have mattered which priority r has (Proposition 2.10); it could have been inserted anywhere in the sequence. So in order to cause the change from believing $\neg q$ to believing q, the agent must have believed $\neg r$ before receiving it.

Going one step further, we could also imagine a revision framework where reordering or deleting formulae from the epistemic state $[\sigma, \blacktriangle]$ are possible. This is reasonable if we think of $\mathcal{A}$ as having several sources and the position where a revision input is inserted into σ depends on a preference among those sources, possibly corresponding to their reliability. If $\mathcal{A}$ learns that one of its sources was always completely unreliable, it might want to delete (or mark as deleted) all inputs received by that source — in which case the actual epistemic state would have to be richer, of course. The meta-information about the source of a formula would have to be recorded, as well. Similarly, if the agent learns that its preference relation among sources does not correctly capture their reliability, $\mathcal{A}$ might decide to reorder the elements of σ according to its new preferences. In order to deal with these cases, we have to make sure that all changes take place in ρ_s, in which case the conditional beliefs needed for the rational explanation algorithm can still be calculated.

Again, the conditions for the optimality of the rational prefix are not satisfied and hypothetical reasoning about $\mathcal{A}$'s beliefs at each point in time is the best we can do. However, the real core belief will entail the one calculated.

We assumed that the indexes of each revision input were given. Having to reason about where the agent may have put the formula leads to a combinatorial explosion and virtually no useful cautious conclusions. We would have to calculate what is implied by all the core beliefs for every possible instantiation of indexes or do hypothetical reasoning with respect to all possible explanations. A similar problem exists if only the relative positions are given, i.e., information like the first input being less important than the second one but more important than the third one etc. The space of possible observations would be smaller than if no information was given, but still very large. We do admit that information about the indexes of the revision inputs will be available only in very few realistic settings.

4.7 Core belief revision

In [8], Booth not only investigates revision by regular inputs but also allows the core belief itself to be revised. He suggests a slightly different representation of the epistemic state of

an agent and a separate function for core belief revision. The agent's epistemic state $[\rho, \rho_\blacktriangle]$ consists of *two* sequences. ρ is analogous to the sequence in the framework we considered in the previous chapters. The second sequence $\rho_\blacktriangle$ records the inputs of core revision. Regular revision inputs are appended to ρ as before, core revision inputs are appended to $\rho_\blacktriangle$. In order to distinguish the two cases we use two revision functions $*$ for regular revision and $*_\blacktriangle$ for core belief revision. We want to emphasise that the subscript $\blacktriangle$ in $\rho_\blacktriangle$ and $*_\blacktriangle$, which we use in this section, is *not* in any way related to a particular formula. It only indicates that the sequence is one of core revision inputs and that the revision is core revision. We use the same representation for an epistemic state but calculate the beliefs in a slightly different way, allowing beliefs to be inconsistent.

Definition 4.7. *The epistemic state $[\rho, \rho_\blacktriangle]$ of an agent consists of two finite sequences of formulae ρ and $\rho_\blacktriangle$. Given an epistemic state and a formula φ, the regular revision operator $*$ is defined by $[\rho, \rho_\blacktriangle] * \varphi = [\rho \cdot \varphi, \rho_\blacktriangle]$. The core belief revision operator $*_\blacktriangle$ is defined by $[\rho, \rho_\blacktriangle] *_\blacktriangle \varphi = [\rho, \rho_\blacktriangle \cdot \varphi]$. The set of beliefs $Bel([\rho, \rho_\blacktriangle])$ in the epistemic state is $Bel([\rho, \rho_\blacktriangle]) = Cn(f(\rho \cdot \rho_\blacktriangle))$.*

Note that beliefs of an agent are inconsistent if and only if the last element of $\rho_\blacktriangle$ is a contradiction. This is the case if the last core belief revision input was inconsistent.[5] As a consequence, revising the core beliefs by any consistent formula will cause the agent's beliefs to be consistent again. The original framework from Section 2.2 is a special case of this one. It does not allow the revision function $*_\blacktriangle$ to be used and assumes $\rho_\blacktriangle$ to be a sequence of length one.[6] And if we disallow inconsistent core belief revision inputs, this framework can be seen as a special case of the one sketched in the last section. We map $[\rho, \rho_\blacktriangle]$ to $[\rho \cdot \rho_\blacktriangle, \top]$, $[\rho, \rho_\blacktriangle] *_\blacktriangle \varphi$ to $[\rho \cdot \rho_\blacktriangle, \top] *_I (\varphi, 1)$ and $[\rho, \rho_\blacktriangle] * \varphi$ to $[\rho \cdot \rho_\blacktriangle, \top] *_I (\varphi, i)$ where i is the length of $\rho_\blacktriangle$ plus one. In order to determine i, we have to know the length of the original $\rho_\blacktriangle$ and increase this number with every core belief revision step. We will later use this mapping in order to reason about an agent employing the current revision framework.

Returning to our main topic of reasoning about an observed agent, let us now assume the agent to function according to the above definition. Further, we assume the observation to contain the information whether a revision input was a regular one or a core revision input. That is, an observation is of the form $o = \langle((\varphi_1, *_1), \theta_1, D_1), \ldots, ((\varphi_n, *_n), \theta_n, D_n)\rangle$, where each $*_i$ is either $*$ or $*_\blacktriangle$. An epistemic state $[\rho, \rho_\blacktriangle]$ explains o if for all $1 \leq i \leq n$, $Bel([\rho, \rho_\blacktriangle] *_1 \varphi_1 *_2 \ldots *_i \varphi_i)$ entails θ_i and entails no $\delta \in D_i$.

[5] The original framework in [8] causes beliefs — core beliefs and regular ones — to be consistent at all times. Basically, it defines the beliefs to be $Bel([\rho, \rho_\blacktriangle]) = Cn(f(\rho \cdot \rho_\blacktriangle \cdot \top))$.

[6] More precisely, it assumes the core is given as a single formula. This is possible also by setting $\blacktriangle = f(\rho_\blacktriangle)$ for a non-empty sequence of core revision inputs $\rho_\blacktriangle$.

In the current setting we have to take special care of inconsistent core revision inputs. We always allowed the agent to have an inconsistent core belief, but whenever the rational explanation algorithm calculated one, we considered this not to be an explanation. Here we have to proceed differently. If we are actually *informed* that the observed agent received an inconsistent core revision input then we can safely use this information. The current framework yields that in this case the agent's beliefs will be inconsistent until the next consistent core revision input is received. Note that this means that there cannot be non-beliefs during that period.

Example 4.8. *Consider the observation* $o = \langle((p, *), q, \emptyset), ((\bot, *_\blacktriangle), \neg q, \emptyset), ((r, *), \top, \{\neg q\})\rangle$. *$o$ expresses that after receiving the regular revision input p, the agent believes q. It then receives an inconsistent core revision input upon which $\neg q$ is believed. After finally receiving the regular input r, the agent does not believe $\neg q$ any longer.*

This observation cannot have an explanation. Assuming the agent's initial epistemic state being $[\rho, \rho_\blacktriangle]$, the final epistemic state after having received the three recorded inputs would be $[\rho \cdot (p, r), \rho_\blacktriangle \cdot \bot]$. The beliefs in this state are obviously inconsistent as the last element of the sequence $\rho \cdot (p, r) \cdot \rho_\blacktriangle \cdot \bot$ is inconsistent. Hence it is not possible that $\neg q$ is not believed.

This illustrates that before looking for an explanation of an observation, we can check whether the observation contains a non-empty set of non-beliefs for a point in time where the beliefs are known to be inconsistent. The following proposition formalises which observations can be excluded from further consideration because they cannot have an explanation for this trivial reason.

Proposition 4.9. *Let* $o = \langle((\varphi_1, *_1), \theta_1, D_1), \ldots, ((\varphi_n, *_n), \theta_n, D_n)\rangle$ *where each* $*_i$ *is either* $*$ *or* $*_\blacktriangle$. *o cannot have an explanation if the following condition is satisfied:*
There are k, l with $k \leq l$ such that $*_k = *_\blacktriangle$, $\varphi_k \equiv \bot$, $D_l \neq \emptyset$ and for all $k \leq i \leq l$, $\varphi_i \equiv \bot$ or $*_i = *$.

The condition states that there is an inconsistent core revision input and there is no consistent core revision input recorded up to a point where there is information about the agent's not believing a certain formula. This is not a necessary condition as there are observations not satisfying this condition that still have no explanation.[7] For the remainder of this section, we consider only observations o that cannot be eliminated due to this proposition.

As noted above, for using the above mapping there must not be inconsistent core revision inputs. The reason is that the mapping turns core revision inputs into regular ones. Due

[7] To see this we can use any observation from Chapter 2 that does not have an explanation, e.g. $\langle((p, *), \top, \{p, \neg p\})\rangle$. Upon receiving p as revision input either p or $\neg p$ has to be believed.

to the index 1 they get high priority but there still is the more important tautologous *super core belief*. Proposition 2.7 yields that this tautology does not have any impact in case the latest core revision input is consistent. After an inconsistent core revision input we know the agent's beliefs to be inconsistent, however using the mapping the beliefs will always be consistent due to the tautologous super core. The following proposition tells us that we can safely eliminate the inconsistent core revision inputs by slightly modifying the observation. o and o' differ in that the entry for the inconsistent core revision input φ_i is eliminated and the beliefs after the regular revision inputs $\varphi_{i+1}, \ldots, \varphi_j$ are replaced by tautologies. That way we get an observation o' to which we can apply the mapping and which has exactly the same explanations as o.

Proposition 4.10. *Let* $*_k \in \{*, *_\blacktriangle\}$ *for all* $1 \leq k \leq n$. $[\rho, \rho_\blacktriangle]$ *explains*

$$\begin{aligned} o = \langle &((\varphi_1, *_1), \theta_1, D_1), \ldots, ((\varphi_{i-1}, *_{i-1}), \theta_{i-1}, D_{i-1}), \\ &((\bot, *_\blacktriangle), \theta_i, \emptyset), \\ &((\varphi_{i+1}, *), \theta_{i+1}, \emptyset), \ldots, ((\varphi_j, *), \theta_j, \emptyset), \\ &((\varphi_{j+1}, *_\blacktriangle), \theta_{j+1}, D_{j+1}), \\ &((\varphi_{j+2}, *_{j+2}), \theta_{j+2}, D_{j+2}), \ldots, ((\varphi_n, *_n), \theta_n, D_n)\rangle \end{aligned}$$

if and only if it explains

$$\begin{aligned} o' = \langle &((\varphi_1, *_1), \theta_1, D_1), \ldots, ((\varphi_{i-1}, *_{i-1}), \theta_{i-1}, D_{i-1}), \\ &((\varphi_{i+1}, *), \top, \emptyset), \ldots, ((\varphi_j, *), \top, \emptyset), \\ &((\varphi_{j+1}, *_\blacktriangle), \theta_{j+1}, D_{j+1}), \\ &((\varphi_{j+2}, *_{j+2}), \theta_{j+2}, D_{j+2}), \ldots, ((\varphi_n, *_n), \theta_n, D_n)\rangle \end{aligned}$$

First note that, as far as the inconsistent core revision input φ_i under consideration is concerned, o does not satisfy the condition in Proposition 4.9. All sets of recorded non-beliefs between φ_i and the following core revision input φ_{j+1} are empty. The idea of the transformation from o to o' is as follows. We can act as if the inconsistent input has not been received. For the time where the agent's beliefs are inconsistent we know they are. Once a consistent core revision input (φ_{j+1}) is received φ_i is irrelevant. This is due to Proposition 2.8 which states that inserting a contradiction anywhere but the last position has no impact. However, just leaving φ_i is not enough because the beliefs recorded between the agent's receiving φ_i and φ_{j+1} could be caused by the belief set being inconsistent. They may have no connection with the regular revision input. That is, in fact we have no information about what the agent really believes after having received the regular inputs, which can be modelled by setting all θ_k, $i < k \leq j$, to be tautologies.

We will now go on to illustrate how to continue with an observation o that has not been eliminated as not explainable by Proposition 4.9 and which has been modified according to

Proposition 4.10 such that it does not contain any inconsistent core revision input. Recall that the agent's initial epistemic state is $[\rho, \rho_{\blacktriangle}]$. If we knew $\rho_{\blacktriangle}$ we could easily translate the observation o into conditional beliefs in the epistemic state $[\rho, \top]$ (in the original interpretation). This is possible due to the above-mentioned translation into revisions using $*_I$. The antecedents of these conditionals would have the form $f(\sigma_1 \cdot \rho_{\blacktriangle} \cdot \sigma_2)$ where σ_1 is the sequence of regular revision inputs and σ_2 the sequence of core revision inputs from a prefix of o. Note that ρ is the prefix of $\rho \cdot \rho_{\blacktriangle}$ which is not affected by the revision steps. Again, the problem is we do not know $\rho_{\blacktriangle}$ and we will have to guess it. This is where the intermediate inputs idea is helpful again. We will try out several possible sequences, each containing a different number of unknown formulae, which will again be instantiated by new variables. We start by assuming $\rho_{\blacktriangle}$ to be the empty sequence.

Having calculated the conditional beliefs from o using the assumed $\rho_{\blacktriangle}$, the rational explanation construction will return an epistemic state $[\rho, \blacktriangle]$. Note that in the current setting it is not sufficient that $\blacktriangle$ is consistent. As there really is no super core belief and all interaction between core revision and regular inputs must be explained by ρ and $\rho_{\blacktriangle}$, we need that $Cn(\blacktriangle) \cap L(o) = Cn(\top)$. This will ensure that all impact the super core belief has can be absorbed by the new variables (Proposition 3.11).

In the first run where we assumed $\rho_{\blacktriangle} = ()$, if the returned core is a tautology then we are done. If it is not this could mean that not all the interactions between regular revision inputs could be handled by the core revision inputs recorded in o.[8] This hints at $\rho_{\blacktriangle}$ containing at least one formula. Next we try if a sequence of length one works. We use the intermediate input idea and instantiate the unknown formula with a new variable and calculate the corresponding conditionals. The rational explanation construction will again yield an epistemic state $[\rho', \blacktriangle']$. So if $Cn(\blacktriangle') \cap L(o) \equiv Cn(\top)$ then we are done, otherwise we will have to extend $\rho_{\blacktriangle}$ by another new variable. And so on.

The question is whether we can stop this iteration at one point, i.e., whether there is an assumed length of $\rho_{\blacktriangle}$ that allows to explain an observation no matter how long the true $\rho_{\blacktriangle}$ really is. Proposition 3.26 gives the answer. Interpreted to the current setting it says that $\rho_{\blacktriangle}$ need not be longer than the number of core belief revision inputs recorded in o plus 1. Intuitively, the extra formula is necessary for the super core belief to be included in $\rho_{\blacktriangle}$. ρ in that proposition is the sequence of core belief revision inputs received during the observation, σ is the actual sequence $\rho_{\blacktriangle}$ and σ' a sequence of the claimed length which yields equivalent formulae before processing the regular revision inputs.

[8] It could also mean that there are problems with the interaction between core belief inputs. Consider $o = \langle((p, *_{\blacktriangle}), p, \emptyset), ((q, *_{\blacktriangle}), \neg p, \emptyset)\rangle$. From our original investigations it is clear that o can only be explained using a super core belief that is not a tautology. As $p \wedge q$, the conjunction of the two latest core revision inputs, is consistent, $\neg p$ cannot consistently believed after receiving q. Hence, o does not have an explanation.

Example 4.11. *Consider* $o = \langle((p,*), p \wedge q, \emptyset),\ ((r,*_\blacktriangle), \top, \{p\}),\ ((r \wedge p,*), \top, \{r \wedge p\}),\ ((\bot,*_\blacktriangle), \top, \emptyset),\ ((r,*), \neg r, \emptyset),\ ((s,*_\blacktriangle), r \wedge p, \emptyset)\rangle$. *After receiving the regular input* p*, the agent believes* $p \wedge q$*. After receiving the core revision input* r*, it ceases to believe* p*. The regular revision input* $r \wedge p$ *does not cause this formula to be believed. The agent then receives an inconsistent core revision input and afterwards a regular one. Finally, the core revision input* s *leads to the belief in* $r \wedge p$*.*

There is an inconsistent core revision input, so we check whether this observation cannot have an explanation due to Proposition 4.9. This is not the case because all recorded non-beliefs following that input and preceding the next consistent core revision input are empty. We now modify o *according to Proposition 4.10, deleting the inconsistent core revision input and weakening the following beliefs to be tautologies. The resulting observation is*

$$o' = \langle((p,*), p \wedge q, \emptyset), ((r,*_\blacktriangle), \top, \{p\}), ((r \wedge p,*), \top, \{r \wedge p\}), ((r,*), \top, \emptyset), ((s,*_\blacktriangle), r \wedge p, \emptyset)\rangle.$$

Note that if we had not weakened the belief recorded for the regular revision input r *and left it unchanged* ($\neg r$) *then the resulting observation could not have an explanation! This is because the epistemic state after receiving that input would be* $[\rho \cdot (p, r \wedge p, r), \rho_\blacktriangle \cdot r]$ *and in this state* r *would be believed independent of what the initial epistemic state* $[\rho, \rho_\blacktriangle]$ *is.* $\neg r$ *could not be consistently believed.*

Before illustrating the construction of an explanation for o *and* o' *we want show that there is one:* $[\rho, \rho_\blacktriangle]$ *where* $\rho = (p \to q)$ *and* $\rho_\blacktriangle = (r \to (\neg p \wedge \neg s))$*. The following table shows the evolution of the epistemic state for the revision steps recorded in* o *and gives a formula characterising the beliefs of the agent. Recall that the beliefs are calculated using* f *going backwards through the concatenation of* ρ *and* $\rho_\blacktriangle$*.*

rev.	*epistemic state*	*beliefs*
	$[(p \to q), (r \to (\neg p \wedge \neg s))]$	$(\neg p \vee q) \wedge (\neg r \vee \neg p) \wedge (\neg r \vee \neg s)$
$*p$	$[(p \to q, p), (r \to (\neg p \wedge \neg s))]$	$p \wedge q \wedge \neg r$
$*_\blacktriangle r$	$[(p \to q, p), (r \to (\neg p \wedge \neg s), r)]$	$r \wedge \neg p \wedge \neg s$
$*r \wedge p$	$[(p \to q, p, r \wedge p), (r \to (\neg p \wedge \neg s), r)]$	$r \wedge \neg p \wedge \neg s$
$*_\blacktriangle \bot$	$[(p \to q, p, r \wedge p), (r \to (\neg p \wedge \neg s), r, \bot)]$	$\bot$
$*r$	$[(p \to q, p, r \wedge p, r), (r \to (\neg p \wedge \neg s), r, \bot)]$	$\bot$
$*_\blacktriangle s$	$[(p \to q, p, r \wedge p, r), (r \to (\neg p \wedge \neg s), r, \bot, s)]$	$s \wedge r \wedge p \wedge q$

There cannot be an explanation $[\rho, \rho_\blacktriangle]$ *with an empty sequence* $\rho_\blacktriangle$*. This can already be seen from the first two revision steps.* $\rho_\blacktriangle$ *being empty the resulting epistemic state is* $[\rho \cdot p, (r)]$*. It is easy to see that* $r \wedge p$ *is believed in this state, as* p *cannot be kept out of the belief set. But the observation tells us that* p *is not believed. Consequently* $\rho_\blacktriangle$ *must have a length of at least one. So let us assume there is exactly one unknown formula in that sequence. As always, we instantiate this unknown formula with a new variable* x*. The positive and negative*

conditional beliefs with respect to the epistemic state $[\rho, \top]$ *which we can now construct from* o' *are as follows.*

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

Of all those conditionals only the first positive conditional $f(p,x,\top) \Rightarrow p \wedge q$ *is not ultimately exceptional.* $\bigwedge U_m \equiv \neg p \vee \neg x \vee \neg r$ *so we strengthen the old core* $\top$ *with this formula and recalculate the conditionals using the new core belief* $\blacktriangle = \neg p \vee \neg x \vee \neg r$*.*

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

This time only the last positive conditional $\neg p \wedge x \wedge r \wedge s \Rightarrow r \wedge p$ *is ultimately exceptional. Strengthening the old core with* $\neg p \wedge x \wedge r \wedge s \rightarrow r \wedge p$ *leads to a new core belief which is equivalent to* $(\neg p \vee \neg x \vee \neg r) \wedge (\neg x \vee \neg r \vee \neg s)$*. Using this core belief none of the conditionals is exceptional even in the first iteration and the rational prefix is equivalent to* $(p \wedge x \wedge \neg r \rightarrow q)$*. We see that* $Cn((\neg p \vee \neg x \vee \neg r) \wedge (\neg x \vee \neg r \vee \neg s)) \cap L(o) = Cn(\top)$ *which tells us that we have indeed found an explanation when assuming the initial core belief sequence to have length one. Absorbing the entire (super) core belief into the unknown formula, i.e., using* $x \wedge (\neg p \vee \neg x \vee \neg r) \wedge (\neg x \vee \neg r \vee \neg s)$*, we get the following explanation for both* o *and* o'*.*

$$[(p \wedge x \wedge \neg r \rightarrow q), (x \wedge (\neg p \vee \neg r) \wedge (\neg r \vee \neg s))]$$

Restricted to $L(o)$ *and including the inconsistent core revision input from* o*, the belief trace of the agent is as follows.*

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

Now that we have a method for calculating an explanation for an observation in the current setting, what can we actually say about the agent? Our answer is again: Whatever hypothetical reasoning allows us the conclude safely. Having constructed an explanation we can make predictions about beliefs, non-beliefs and (regular) inputs being accepted or rejected and then test these hypotheses by modifying the original observation accordingly. This also

indicates that we might start by assuming $\rho_{\blacktriangle}$ to be as long as possible. By proceeding this way no explanations are missed. If an explanation exists for a shorter sequence of core beliefs the remaining unknown formulae might just be tautologies.

We want to remark that the agent's initial epistemic state might in fact be $[\rho, \rho_{\blacktriangle} \cdot \bot]$ which means that its belief set is inconsistent until it receives the first consistent core revision input recorded in the observation. The methodology sketched above cannot deal with this case immediately, intuitively because no successful calculation returns a contradiction or a sequence containing one. However, if we do not find an explanation for a given observation o, we might then try finding one for $\langle((\bot, *_{\blacktriangle}), \top, \emptyset)\rangle \cdot o$. This modified observation expresses, that the latest core revision input, which is inconsistent, was received just before the observation started. Due to Proposition 2.8, $\rho_{\blacktriangle}$ can be assumed to have a consistent last element, as inserting a contradiction in any proper prefix of a sequence has no impact. That is, if $[\rho, \rho_{\blacktriangle}]$ explains the modified observation $\langle((\bot, *_{\blacktriangle}), \top, \emptyset)\rangle \cdot o$ (in which case we will find an explanation as well), then $[\rho, \rho_{\blacktriangle} \cdot \bot]$ will explain o.

If the observation does not specify what type of revision an input triggered, we have to make guesses. The assumption in the main part of this thesis was that there are only regular revision inputs. This assumption manifested itself in the choice of belief revision framework used. In the previous section we allowed inputs to be placed at any point in the sequence and argued that if we have no information, we will have to go through all possibilities. The framework in this section limits the possibilities, as any input can only be placed at two different positions. However, this still leaves the search space quite large. In order to compare different explanations — which then have to include which type of revision a formula triggered — we would have to specify further preference criteria. They could include minimising the number of core belief revision steps, a short initial sequence of core beliefs $\rho_{\blacktriangle}$, preferences as to the position of core belief revisions, etc. With respect to the last criterion it is particularly controversial whether to prefer early or late core belief revisions. But we will not continue this line of investigation here.

We want to conclude this chapter with a general remark concerning hypothetical reasoning in the context of extending the original framework from Chapter 2. That is, we are referring to parametrised observations, intermediate inputs and the belief revision frameworks introduced in this chapter. Having drawn conclusion based on the observation o, these are tested by constructing a new observation o' such that an explanation for o' would be a counterexample to the conclusion. We want to emphasise that the potential explanation for o' has to be constructed making sure that all assumptions are correctly dealt with. That is, this explanation must be constructed as if we did not know about o. In particular, results from o need not automatically carry over. We want to give only one example. Assume o

is an observation where intermediate inputs may have occurred and we know the positions where they have occurred. We have the result that for explaining o, we need only assume j intermediate inputs at each position (j being the number of recorded inputs following that position). If we now want to draw conclusions about the agents beliefs after further inputs have been received then hypothetical reasoning tells us to extend o by these inputs in order to get o'. Note that now, there are more than j recorded inputs in o' after the positions indicated in the original observation. That is, we have to add further intermediate inputs into o'.

Chapter 5

Related Work

We are not aware of work that investigates reasoning about the evolution of an observed agent's beliefs matching our setting. In this chapter we want to relate our work to a selection of papers dealing with similar topics or using similar ideas. We will start with a number of papers that are placed in action setting before turning to papers on modal logic approaches to beliefs. The last papers we mention are only remotely related to our work.

5.1 Reasoning about other agents based on their actions

Modeling agents as qualitative decision makers [18] is very close in spirit to our work, but it is placed in an action setting. The authors propose a framework that allows to talk about agents independently of their actual implementation. They stress the importance of being capable of modelling other agents without having access to their internal structure. They assume that an agent's behaviour can be explained in terms of its mental state which is defined by beliefs, preferences and a decision criterion. Beliefs describe which worlds the agent thinks most likely to be the true one, the preferences describe which outcomes of actions it finds how desirable, and the decision criterion is used to choose an action. This structure is influenced by decision and game theory literature.

The authors suggest that reasoning about another agent is possible by interpreting its actions. In particular, they consider the belief ascription problem. In the simple (static) case, this is determining what an agent believes given its preferences, decision criterion and a single action performed by the agent. Assuming rationality of the agent, the action must have been such that it maximises the outcome with respect to the decision criterion according to its beliefs about the world. In this sense ascribed beliefs have to be an explanation for the observed action, which is analogous to our notion of an epistemic state explaining an observation.

In the complex (dynamic) case sequences of actions are considered. This involves investigating the relation between the agent's mental states at different points in time, in particular the dynamics of its beliefs. The authors give conditions under which the belief ascription problem can be solved by translating it into a set of ascriptions in the static case. Whereas [18] focuses on the presentation of the general framework, we provide a set of actual methods for reasoning about another agent.

5.2 Reasoning about the evolution of a dynamic world

There are a number of papers investigating the question of how the world evolved given (partial) information about each point in time. In this section, we will mention some of them. Even when abstracting from the difference of static vs. dynamic world, our work still does not have the same focus. Whereas those papers basically allow arbitrary changes at each point in time or allow different actions to occur, the observation in our case precisely tells us what happened. The observer knows which revision input was received at each point in time.

Belief extrapolation (or how to reason about observations and unpredicted change) [22] This is a paper that deals with completing information about beliefs over time and presents *belief extrapolation* operators. The starting point is a scenario $\langle \varphi_1, \ldots, \varphi_N \rangle$ representing that φ_i holds in a possibly changing world at time point i. Such a scenario is a partial description of how the world has evolved. Assuming that fluents (literals) tend not to change, i.e., that the world is inertial, the operator tries to identify preferred trajectories of models $\langle m_1, \ldots, m_N \rangle$ such that $m_i \models \varphi_i$.

The authors present several strategies for minimising change: counting all changes of fluents, counting changes *per* fluent, set inclusion of changing fluents, different penalties for changes of different fluents, temporal considerations like changes occurring as late as possible, etc. Each strategy gives rise to a preference relation among trajectories. Choosing one preference relation, the result of the extrapolation is a sequence of formulae representing for each point what is true according to all preferred trajectories explaining a scenario. Static laws relating different fluents are not explicitly treated, but the authors point out that these laws can simply be conjoined with every formula given in a scenario.[1] A preference relation is called inertial if all static trajectories, i.e., those where all models are identical and thus no change occurs, are equally preferred and preferred to any non-static trajectory. The rationale behind choosing an inertial preference relation is that as long as we can assume the world

[1] The paper does not investigate the case where the static laws are not known. In this case the operators assume that there are none.

not to have changed, we should do so. The authors present properties of and connections between the different preference relations and the extrapolation operators they give rise to and position extrapolation with respect to belief revision and update. Computational aspects are considered as well.

There are some essential differences between belief extrapolation and our approach. Once more, the work in [22] is focused on a first person perspective and describes what an agent *should* believe at each point in time rather than reasoning about what the agent *does* believe. The second important difference is that both approaches minimise very different things. For the sake of illustrating this, let us for now forget that we assumed the world to be static and only the agent's beliefs about it to change. Note that one possible interpretation of the formulae in a scenario $\langle \varphi_1, \ldots, \varphi_N \rangle$ is as the beliefs recorded in an observation. Given the scenario $\langle p, q, r \rangle$, an extrapolation operator based on an inertial preference relation will conclude that $p \wedge q \wedge r$ held at every point in time. This is because the conjunction of all the formulae is consistent and it can thus be assumed that nothing changed at all. We will now look at some potential translations from a scenario $\langle \varphi_1, \ldots, \varphi_N \rangle$ to observations in our setting: (i) $\langle (\varphi_1, \top, \emptyset), \ldots, (\varphi_N, \top, \emptyset) \rangle$, (ii) $\langle (\varphi_1, \varphi_1, \emptyset), \ldots, (\varphi_N, \varphi_N, \emptyset) \rangle$, (iii) $\langle (\chi_1, \varphi_1, \emptyset), \ldots, (\chi_N, \varphi_N, \emptyset) \rangle$, and (iv) $\langle (\top, \varphi_1, \emptyset), \ldots, (\top, \varphi_N, \emptyset) \rangle$. The rational explanation for *any* observation of the form (i) and (ii) will be $[(), \top]$. This is because the material counterparts for all positive conditionals will be tautologies. The belief trace for the example scenario will thus be $(\top, p, p \wedge q, p \wedge q \wedge r)$. So with respect to these translations we do not conclude that $p \wedge q \wedge r$ is believed at every point in time. The translation according to (iii) will yield a similar belief trace. Our approach tries to minimise the *beliefs* we assign to the agent and not the changes, in particular when considering hypothetical reasoning.[2] As the agent *may* consider $\neg p$ more plausible than p etc. we cannot conclude that it believed p before being informed about it. In this sense belief extrapolation is credulous, using the inertia assumption in order to come up with strong beliefs. The rational explanation (in particular in combination with hypothetical reasoning) is a (very) sceptical approach. Although the world considered may be static, the agent's information about it may be highly unreliable and hence the agent's *beliefs* may change often and dramatically.

Note that for the given example the translation according to (iv) $\langle (\top, p, \emptyset), (\top, q, \emptyset), (\top, r, \emptyset) \rangle$ will yield exactly the same conclusion (via the corresponding belief trace) as the belief extrapolation operator. However, the inputs $\top$ in fact *force* us to conclude that the agent's belief set did not change at all and every belief must have been already present in the initial state. (iv) will fail whenever $\bigwedge \varphi_i$ is inconsistent, that is, in all interesting cases. The

[2] Then we could not even conclude that p is still believed after receiving q. The core belief may in fact be $q \leftrightarrow \neg p$.

resulting observation will not have an explanation at all as the core belief is forced to be inconsistent.

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

The authors of [22] indicate how explicit information about change can be incorporated into their approach. They suggest mixed scenarios where each formula is labelled indicating whether it denotes an actual description of the state of the world or an *expected change* caused by an update. Then not all possible trajectories are considered but only those which fit the expected changes recorded in the mixed scenario. These trajectories are then compared with respect to the *unexpected* changes. This still does not cause the two approaches to collapse into the same method. The rational explanation of an observation is about making all changes expected ones. All these differences indicate that generally a translation yielding exactly the same conclusions is not possible in either direction. The task, methodology and assumptions of the two approaches are just too different.

[22] presents a number of possible extrapolation operators based on different preference relations for minimising change. We single out one possible explanation — our suggestion is to use it for generating hypotheses that can be tested via hypothetical reasoning. If asked which extrapolation operator is most similar to our approach, the answer would probably have to be the one obtained from chronological minimisation. This operator tries to delay changes as much as possible. The nature of our assumed belief revision framework is analogous. As long as a revision input can simply be added to the current beliefs, this is what is done. Once a change is necessary, the cardinality of literals that change their value is not relevant. To see this, recall the example given at the very end of Section 3.2.5 (the unknown input in the second position has been eliminated): $o = \langle (\neg p, \neg p, \emptyset), (p \leftrightarrow q_1, p \leftrightarrow q_1, \emptyset), \ldots, (p \leftrightarrow q_n, p \leftrightarrow q_n, \emptyset), (r, q_1 \wedge \cdots \wedge q_n, \emptyset) \rangle$. Retrospec-

tively, it might have been better if the agent had changed from believing $\neg p$ to believing p after having received $p \leftrightarrow q_1$. Then the agent's view of the world — or rather what we assume to have been the agent's beliefs based on the rational explanation of o — would not have to have changed so dramatically.[3]

We already noted that a tautologous revision input forces the belief set to remain unchanged. This indicates how we could define an operator *similar* to chronological (and anti-chronological) minimisation using the rational explanation. We translate a scenario $\langle \theta_1, \ldots, \theta_N \rangle$ into an observation $o = \langle (\top, \theta_1, \emptyset), \ldots, (\top, \theta_N, \emptyset) \rangle$ and check whether o can be explained. If so, there is a static trajectory for the scenario. Otherwise we allow changes which are modelled by intermediate inputs, preferring explanations for an observation where as few intermediate inputs have been assumed as late in o as possible (as early for the anti-chronological case). However, it is not the case that we will get the same conclusions as the extrapolation operator yields. Considering the example scenario $\langle a, a \vee c, b, \neg a \vee \neg b, \neg c \rangle$ from [22], we see that there cannot be a static trajectory as a, b and $\neg a \vee \neg b$ cannot be consistently believed. Consequently, at least one change must have occurred. Indeed one suffices, as the observation $\langle (\top, a, \emptyset), (\top, a \vee c, \emptyset), (\top, b, \emptyset), (\neg a \vee \neg b, \top, \emptyset), (\top, \neg a \vee \neg b, \emptyset), (\top, \neg c, \emptyset) \rangle$ has an explanation. Here we gave a particular intermediate input $(\neg a \vee \neg b)$ rather than reasoning using an unknown formula. The rational explanation yields that $a \wedge b$ was believed from the beginning but $\neg c$ was believed only after the intermediate input was received. Before that, the observation gives no indication that c cannot have been believed and again our framework tries to minimise the beliefs assigned to the observed agent rather than the number of changes.

Preferred History Semantics for Iterated Updates [6] The authors of [6] also consider a sequence σ of consistent formulae partially describing an evolving world. Their aim is to sharpen the information about the last time point, i.e., what should an agent believe having observed the world developing in a certain way, and thereby give an alternative semantics for update. The idea is again to identify trajectories of models explaining σ and the result $[\sigma]$ is the set of formulae true in the end points of all *preferred* trajectories. However, the notion of an explaining trajectory is slightly different. Unlike in [22], where scenario and trajectory had the same length, one model can be used for several time steps and there may even be models not used at all. Of course, trajectories with this last property will not be among the preferred ones.

The authors now assume that the agent has a preference relation over trajectories which is irreflexive, transitive, has no infinite descending chain and satisfies the following property: If

[3]Note that the agent may in fact have changed its mind about p already at that early point. Only, the observation does not indicate that.

one trajectory can be obtained from another by simply deleting one or more models and both explain σ then the shorter trajectory must be preferred. In fact, the authors do not consider one particular preference relation but present properties satisfied by any such operator and provide a representation result.

Most of the differences we pointed out with respect to [22] also apply here. Additionally, note that we put the focus on the initial beliefs (and the core belief) rather than the ones at the very end. The reason is that if we do not have the correct core belief then not even the beliefs with respect to the recorded revision inputs will be reliable (Proposition 2.20). Of course, our intention is to say something about *every* point in time.

Using Ranking Functions to Determine Plausible Action Histories [42] The topic of [42] is also the identification of the most plausible evolution of the world. However, the setting is different from [22] and [6]. The authors assume that a transition system is given, i.e., it is known which actions change the world in what way and changes do not happen arbitrarily. They consider an alternating sequence of actions and observations (about what holds at the current point in time), each action being followed by an observation. Now, given ranking functions expressing how likely each action and each observation is at each point in time, the task is to find the most plausible action sequence. It is not the case that the observations are reliable. This is unlike our framework where we assume the agent to really believe θ_i, here indeed the information may be incorrect. Resolving resulting conflicts is done by minimisation using the given ranking functions. This may lead to information about the world being rejected or to reconsidering the action sequence.

Not every change needs to be equally likely at each point in time. This can be encoded in the ranking. Having identified the most plausible action history, the beliefs at each step are implicitly completed if the initial situation is known. The authors do not investigate the case where the transition system is incorrect or incomplete. They note that fallible actions can be modelled by allowing non-deterministic actions.

This paper may also be seen as an action analogon to our approach. The difference in the setting is that we know which actions were performed (revisions by a known formula), but we do not know the exact effect that an action had on the beliefs. Recorded beliefs are reliable and are thus a guide to identifying the effect of the revision.

BReLS: A System for the Integration of Knowledge Bases [58] presents a unified framework for integrating information with different levels of (source) reliability taking into account the time at which this information holds. Again, the default assumption is that the world does not change. It is possible to specify penalties for changes of a literal, also for particular time points. The authors now define the notion of preferred models in terms of the

smallest Hamming distance to models of formulae with the same reliability and extend this to the case for successive time points. Again this yields a preference relation on trajectories, which are as close to the formulae specified in the knowledge base as possible[4] in order to minimise change. The notation allows to capture particular instances of revision, update and merging. By looking at what is true in all preferred models at a certain point in time, it is possible to draw conclusions beyond what is explicitly given.

As in most papers considered in this chapter, it is not possible to specify what is *not* to be believed at a given point in time. Also it is not straightforward how to model an observation in that framework. The authors suggest to model revision by giving the new formula a higher priority than the beliefs held (both referring to the same point in time). However, the beliefs recorded in the observation are definitely held whereas the inputs, although definitely received, may in fact have been rejected. So the system does not seem to be designed to model iterated non-prioritised revision.

It is possible to write down that at time t the formula θ_t should hold and that if possible also the input φ_t should be believed. But it seems impossible to specify that φ_t has to be the actual reason for a change of the beliefs from $t-1$ to t. These changes are completely determined by the minimisation strategy and thus queries about beliefs at a certain point in time are usually not related to the results expected in our framework.

5.3 Modal logic approaches

Often, modal logic is used for representing beliefs, knowledge and similar notions (even) in multi-agent settings. Expressing beliefs of one agent about beliefs of another etc. is possible in a natural way via nesting the different modal operators. However, we were not able to find papers whose focus was on *reasoning* about beliefs of another agent in a belief revision setting. So although our framework is quite restricted regarding which information about an observed agent can be represented, we believe to have made advances with respect to what we can conclude about that agent.

Belief reconstruction in cooperative dialogues [23] In [23], the authors deal with the question of determining agents' beliefs through a sequence of speech acts. The new beliefs should depend on the old ones and the input received. Old beliefs should persist if possible. The key point of the motivation is that an input should not always be accepted.

[4] In this respect, this work differs greatly from the approaches mentioned so far, including ours. Information about a point in time may be contradictory and the preferred model tries to satisfy as many of the formulae as possible taking their reliability into account. The others either do not allow contradictory information to begin with or conclude that there is no explanation.

In particular it should be rejected if the speaker is incompetent with respect to the content of the utterance.

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

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

Mutual enrichment through nested belief change [83] describes a (modal) framework for representing nested beliefs of a set of agents as well as the dynamics of these beliefs. The intention is to capture agents' beliefs and their beliefs about other agents' beliefs in a dialogue setting. The framework also incorporates agents' preferences about which source (agent) is more reliable than another as well as the nested case, i.e., beliefs about other agents' preferences.

The basic performative (speech act) is *tell* which allows one agent to let another agent know that it believes a propositional formula according to some source. Depending on its preferences among sources, the receiver may now revise its beliefs or reject the input, for example because it has contradicting information from more reliable sources. The two performatives *accept* and *deny* further allow an agent to inform another agent that the content of a speech act has been accepted or rejected. This can help the sender of a message to refine its beliefs about the receiver's beliefs and preferences among sources. In fact, in the setting introduced, the receiver is forced to inform the sender whether the input has been accepted or not.

The authors now present a number of postulates that restrict the progression of (nested) beliefs and preferences through a sequence of performatives. These state, for example, that the beliefs of (and nested beliefs about) agents not involved in a performative do not change or that the sender of a formula θ must revise its beliefs about the receiver believing $\neg\theta$ in case the receiver accepted θ.

Like the framework introduced in [23], the one presented in [83] is more sophisticated than the one we assume.[5] Propositional formulae are labelled with a source. An agent's preference relation on sources allows it to reject inputs and recency is not the dominating criterion. We can use a criterion like reliability only when considering a alternative belief revision framework like the one illustrated in Section 4.6. Whereas our framework is inherently restricted to the observer's nested beliefs about the observed agent $\mathcal{A}$, [83] handles the (nested) beliefs of many agents simultaneously. However, the authors give no explicit method for actually reasoning about other agents' beliefs. It is not clear whether it is possible to infer prior beliefs or preferences among sources when not given the initial situation. Also, we do not assume that every performative (a revision input φ can be interpreted as a *tell*) is acknowledged by *accept* or *deny*. In fact, whether a revision input is accepted or rejected is one of the questions we want to answer. Of course, the explicit knowledge of whether an input was in fact accepted or rejected makes things less ambiguous (if not easier). Note that this information can be encoded in the observation. An accepted input can be made to belong to the beliefs after revision, a rejected input to the non-beliefs.

Dynamic Epistemic Logic [93] The papers [23, 83] presented above and many other publications utilise modal logics for representing and reasoning about agents' beliefs. Often, the languages introduced allow to express the information contained in an observation. Dynamic epistemic logic [93] is another example which also handles the dynamics of such beliefs. Here, model checking is possible. That is, in case the initial state is given, it is possible to test whether it satisfies a given formula by progressing the revision inputs and checking whether the beliefs and non-beliefs calculated fit the ones expressed in the formula. However, personal communication with Hans van Ditmarsch and Wiebe van der Hoek indicates that *generation* of models is problematic. Further, if proof systems are given, they allow us to check entailment but do not generate entailed formulae. Our approach can be interpreted as doing just that. Given a formula (observation), we construct a model satisfying that formula. Using this model, we can generate hypotheses as to what might and might not be believed by the agent. Of course, not all formulae of dynamic epistemic logic can be expressed as observations. In particular, we assume that all revision inputs, beliefs and non-beliefs are objective formulae, i.e., they do not contain any modalities. But it seems

[5] More on disbeliefs can also be found, e.g., in [20, 72, 82].

that, for a subclass of DEL-formulae and the revision operator corresponding to our assumed belief revision framework, our methodology may in fact be used to construct a model for a DEL-formula. Precise connections remain to be established.

5.4 Further related papers

In this section, we present some papers that are less closely related to our work. They illustrate that similar ideas or methods appear in different settings.

Learning non-monotonic causal theories from narratives of actions [59] In [59], the author investigates a method for learning the effects of actions from the observation of a dynamic system. A *narrative* is the description of the initial situation together with a sequence of actions that have been performed and a full description of the changes that have occurred. Assuming a frame axiom, the fluents that have not changed can be inferred. The task for the learner is now to produce from a set of narratives an action theory that explains the changes recorded and that can be used for predicting the evolution of a system for arbitrary action sequences. This is done using the paradigm of Inductive Logic Programming [65]. The narratives are encoded as extended logic programs and the action effects will be returned as rules of a particular structure. The author further describes issues related to indirect effects of actions and concurrent actions.

This approach could be seen as an action analogon of our approach — given an observation, identify the rules that govern the system. However, there are some differences. In our work the immediate rules that govern the agent are fixed. The belief revision framework exactly describes how revision inputs are dealt with. The actual effects of inputs on the belief set are determined by the agent's initial epistemic state. That is, we are not after the global description of effects of "actions" but their effects with respect to some particular initial state. As a consequence, observations with respect to different initial states are of no use to us.[6] We do not assume to be given a total description of the beliefs at every point and in general we have no direct information about the initial state. However, we want to note that our task does not trivialise if initial beliefs are provided. This would be the case if the first element of the observation had the form $(\top, \theta, D)$. In Section 2.8.2, we argued that even if we are given a complete description of the agent's beliefs at every point, we are not guaranteed to identify its real initial state.

[6] A scenario where we know that a number of different agents has the same core belief we are after could be an exception. We are then given observations with respect to different initial states and try to identify a core belief that is acceptable for all of them.

Regression with Respect to Sensing Actions and Partial States [92] Many formalisms for reasoning about action and change are focused on progressing a state description through a sequence of known actions. One general problem where the inverse operation of regression is applied is planning. Given an initial state s_0 and a set of goal states, the task is to come up with a sequence of actions which, started in s_0, leads to one of the goal states. Goals are often represented as a conjunction of literals and not arbitrary formulae. One possible approach is to identify immediate predecessor states of (goal) states using regression. The two main alternatives are regression with respect to states, i.e., given a state and an action find possible predecessor states, and regression with respect to formulae, i.e., given a formula describing what holds in a state and an action find a formula describing what holds in possible predecessor states. [92] presents a method for constructing conditional plans including sensing actions. The approach works with partial state descriptions and uses the sensing actions in order to identify the value of a fluent in case it is unknown but needed. The agent thus has the capability of completing information about the current state whereas we try to infer that information retrospectively. The connection of our work to planning is thus marginal. We know the actions that have occurred as they are recorded in the observation. However, what our approach does can be seen as a combination of progressing and regressing actions through a sequence of states, completing the partial information provided by the observation. It is not exclusively one or the other. This is because additional information about one state can have an effect on both what could hold in earlier and later states.

A consistency-based approach for belief change [26] The approach to belief revision described in [26] is interesting for our work in two respects. It allows for expressing that the result of revision is not to entail a set of formulae analogous to the non-beliefs recorded in an observation, and it is based on language extension. A belief change scenario is a triple (K, R, C) where K is a set of formulae describing the agent's current beliefs, R a set of formulae that should be entailed by the result of revision, and C a set of formulae that should not be entailed. The general idea is now to make the languages of K and R disjoint by renaming the propositional variables in K obtaining a set K', taking the union of R and K' together with a maximal set of equivalences EQ such that the elements of C are not entailed. The equivalences in EQ re-establish the connection between propositional variables and their renamings. The instantiation of unknown subformulae by new variables uses a similar idea, but there the equivalence is between a variable and a complex formula (Proposition 3.5). Also, we do not make this equivalence explicit but reason without identifying the actual formula by considering the core belief and the belief trace restricted to $L(o)$.

The authors describe a class of operators based on this methodology. As there are generally several maximal sets EQ, each yielding an extension, one may be interested in the beliefs in

case a single extension is chosen or beliefs held in all extensions. The authors do not consider the problem of determining predecessor belief sets, i.e., given R, C, and the resulting set of beliefs K_* what could K have looked like. This would be the question analogous to the motivation of our work.

Chapter 6

Conclusion and Future Work

In contrast to the problem of designing an agent in the sense of specifying how it *should* change its beliefs when receiving new information, this thesis investigated a method for modelling an observed agent $\mathcal{A}$ in order to reason about its *actual* beliefs. That is, rather than treating the agent in a first person perspective we adopt a third person one. Our conclusions about $\mathcal{A}$ are based on an observation o about its belief revision behaviour. o contains information about which belief revision inputs $\mathcal{A}$ received over a certain length of time and a partial description of what it believed and did not believe upon receiving them. Aucher [3] would classify our work as imperfect external modelling approach. The observer (and modeller) is not involved in the situation[1] and the knowledge about $\mathcal{A}$ is not complete.

We are interested in drawing conclusions about $\mathcal{A}$'s unrecorded beliefs including possible future beliefs as well as which revision inputs it may or may not accept. We assumed the observed agent to employ a particular belief revision framework for iterated non-prioritised revision. It is a very simplistic framework in the sense that it only allows for representing propositional revision inputs and beliefs. Preferences, beliefs about other agents' beliefs, awareness of the observer etc. cannot be captured. These are assumed to be without impact on the propositional beliefs and on the result of revising them. In other words, we assume that $\mathcal{A}$'s true epistemic state and revision framework can be projected to the simplified one we are working with and that results we obtain carry over.

One important component of $\mathcal{A}$'s epistemic state is its core belief $\blacktriangle$ which is a formula representing a belief $\mathcal{A}$ will not give up no matter what information it will ever receive. In particular, revision inputs contradicting the core belief are rejected by the agent. When

[1] We only consider $\mathcal{A}$'s propositional beliefs. It may have beliefs about the observer's beliefs but these are simply disregarded. The observer being the source of the revision inputs does not change the setting, as the same inputs can be seen as coming from some arbitrary source. $\mathcal{A}$'s assumed belief revision framework does not distinguish different sources. So the modelling perspective is indeed external.

considering only the belief set, this framework allows for reinterpreting a sequence of revisions as a single revision in the agent's initial epistemic state — provided the core belief is known. This property was used to translate the observation into conditional beliefs in the agent's initial epistemic state. These are a partial description of the rational consequence relation that the epistemic state of an agent describes. We then used existing results to complete this relation which could immediately be used for completing the agent's initial epistemic state. As $\mathcal{A}$'s core belief is generally unknown to the observer, one of the problems to be solved was identifying acceptable ones.

We call an epistemic state an explanation for o if it gives rise to the beliefs and non-beliefs recorded in o when revised using the given inputs. In general, o will have many different explanations. We singled out a particular one which, with respect to certain criteria, minimises the beliefs we assign to the observed agent. One of the main results presented was the algorithm that actually calculates this *rational explanation* by iteratively testing and refining potential core beliefs. We proved that there is a unique weakest acceptable core belief justifying conclusions about which revision inputs must be rejected by the agent. We also showed that the beliefs and non-beliefs calculated from this explanation need not coincide with the agent's real ones and provided a tool for improving predictions. The key idea was to modify the observation o according to some conjecture such that an explanation would be a counterexample.

A number of assumptions were imposed in order to obtain these results. For the central ones we assumed that the observer has perfect knowledge of the revision inputs received by the agent, i.e, that every revision input received during the time of observation is contained in o and that the logical content of every input is known. This does not mean that the inputs are assumed to correspond to what is true in the real world — otherwise the union of the recorded inputs would have to be consistent. Consequently, $\mathcal{A}$'s beliefs at each point of time need not be complete nor correct with respect to the real world. However, the observation contains correct (partial) information about $\mathcal{A}$'s beliefs.

We went on to consider observations with incomplete or missing revision inputs and even more partial descriptions of beliefs and non-beliefs. Most importantly, this amounts to weakening the assumption of having perfect knowledge about the revision inputs received by $\mathcal{A}$. Here the main tool was to allow formulae in the observation to contain unknown subformulae. They allow the representation of unknown logical content and hence also the formalisation of less specific information about the observed agent. The unknown subformulae were dealt with by instantiating them with new variables. With this instantiation we obtained an observation in the original sense. It could thus be dealt with using the methodology introduced before.

Our work is focused on reasoning about one observed agent. We briefly illustrated the case of reasoning about several agents. We also sketched approaches to other variations of the main question, e.g., self-observations which can be used for reasoning about what other agents can conclude about oneself, several observations with respect to the same initial state which could be applied for reverse engineering software agents or accessing expert knowledge, as well as reasoning about agents using variants of the assumed belief revision framework.

This thesis also contains results concerning the computational complexity of the general problem as well as the main algorithm given. There are a number of open questions and several directions for extending this work.

We assumed one particular belief revision framework. However, the literature offers a wide variety of revision operators. The definition of an observation is very general and it would be interesting whether similar results can be obtained when assuming the observed agent to employ a different belief revision framework. In case this can be done, is there a framework which allows us to draw good conclusions no matter which revision framework $\mathcal{A}$ is *really* using? Is it possible to conclude from o which framework the agent is actually using? To start with, it would be interesting to see what can be said when *generating* observations using different frameworks and then reasoning about the agent using the methodology we presented. A first step in this direction would be an implementation of the methods presented. Conceptually, this is not a difficult task as the ingredients of the methods presented are simple formula manipulations and satisfiability tests. The pseudo-code for many of the algorithms verbally described in the thesis is given in the appendix.

We conjecture that other belief revision frameworks based on total preorders on worlds can be dealt with using a methodology similar to the one we presented. As mentioned before, one essential step is translating a sequence of revisions (each leading to a new epistemic state) into a set of revisions with respect to the same state. If this can be done, the observation can be seen as containing information that talks about a single state. A second essential step is the interpretation of the revision framework as a rational consequence relation. As long as the framework selects exactly the minimal φ-worlds for constructing the beliefs after revision by φ, this should be possible. Acceptance or rejection of an antecedent of a conditional is not captured in the rational consequence relation. Hence, we had to present a method for manipulating the conditionals by selecting a new core belief. We believe this to be the most problematic part when trying to transfer the results to other frameworks. It is not obvious that a systematic way exists for refining the conditionals until they correctly capture the observation. For frameworks not based on total preorders on worlds investigations will probably have to start from scratch, in which case we hope that this work at least provides some hints for possible directions and interesting questions. We believe that research with

our focus in a modal logic setting might be interesting. Multi-agent higher-order beliefs, which are easily expressed there, could bring this work closer to actual applications.

So far, we have discussed hypothetical reasoning — our proposed method for verifying conclusions drawn about $\mathcal{A}$ — only with respect to elementary conclusions, e.g. belief or non-belief in a single formula. However, the formal results do not prevent us from making more than one change in the original observation. Consider $o = \langle (p, \top, \emptyset), (q, \top, \emptyset), (p \vee q, p \vee q, \emptyset) \rangle$. The rational explanation $[(), \top]$ for o allows us to conclude that the agent believed p before receiving $p \vee q$. The belief in q at the same point can also be concluded. None of the two conclusions is safe as the agent's core belief may in fact be $\neg p$ or $\neg q$. However, it is safe to conclude that the agent must believe p or q at that point. It is impossible that $\mathcal{A}$ believes neither. This can be verified by considering $o' = \langle (p, \top, \emptyset), (q, \top, \emptyset), (p \vee q, p \vee q, \{p, q\}) \rangle$ which does not have an explanation. We conjecture that generally there will be few non-trivial elementary conclusions that are safe. A more thorough investigation of that question would be interesting, in particular with respect to identifying structured ways for coming up with useful complex conclusions.

When dealing with unknown subformulae and intermediate inputs, we extended the language L of the observation in order to find an explanation. We gave a necessary and sufficient condition for an explanation restricted to L to exist. However, we could not provide a way to construct one or even to decide if there is one. We argued that it is not necessary to look for a concrete instantiation for the unknown subformulae. However, it may still be an interesting problem in several ways. Is it possible to find an instantiation which belongs to L such that o can be explained if we know there exists one — other than by testing all possible formulae from L? In abduction, there is usually a limited set of possible explanations (abducibles) which are generally literals. In our case the instantiations can be arbitrary formulae. Can we objectively say that one is better than another? What if we have a set of candidate instantiations; which should we choose? Of course, these questions also involve finding, justifying and investigating reasonable preference relations.

A big field for possible future investigations is handling incorrect observations. With respect to beliefs and non-beliefs we already sketched one solution when being provided with reliability information. We believe revision inputs to be a trickier matter. Note that we assumed a static world and that hence correct revision inputs would have to be jointly consistent. This is not the incorrectness we are talking about here. We did not assume that the revision inputs correctly describe the world but that they are the ones indeed received by $\mathcal{A}$. But what if the observation is incorrect in this sense? How to reason about $\mathcal{A}$ if we are not sure which revision inputs it really received? Here again, unknown subformulae could be a useful tool, but it is also necessary to investigate general questions about which input(s)

to assume to be incorrect, how to compare explanations obtained from observations with different incorrect inputs, etc.

We assumed that the observations are used in a passive way. A next step would be to actually develop strategies for eliciting information about the observed agent, to extend the methodology to allow goal directed observation. Which revision inputs should $\mathcal{A}$ be provided with in order to gain as much information as possible? If we are allowed to ask questions about beliefs and non-beliefs what should we ask? Which cases should an expert be provided with in order to extract as much knowledge as possible?

In many settings, several agents will cooperate to achieve a common goal. So it makes sense to consider the case where a group of agents observes an agent $\mathcal{A}$ and tries to form a common picture of it. When intending to use the methodology introduced in this thesis, the main problem that needs to be solved is how to integrate information from several observations. We showed that several observations with respect to the same initial state can be dealt with, but in the scenario sketched here, we have further restrictions. All the observations have to be synchronised. The time periods of the observations may overlap so that in fact they are not with respect to the same initial state to begin with. Agents may have observed different inputs, beliefs, non-beliefs for the same point in time — the corresponding formulae may even be contradicting. With respect to missing inputs, this approach could even be helpful. One agent could have recorded inputs and beliefs another agent missed.

In Section 4.1, we briefly sketched that it may be interesting to reason about the dynamics of a dialogue from observations of the agents involved. This will also be important when actually using the conclusions drawn from observations for planning responses and coming up with strategies for the interaction with other agents.

As a last point, we want to recall an open question given in Section 4.2. We showed that being able to deal with many observations with respect to the same initial epistemic state, we can represent any conditional information about a rational consequence relation as a set of observations. The exact relationship of the rational closure of the conditionals and the rational explanation for the set of observations still needs to be determined. Note that in the rational explanation construction the conditionals may be modified by changing the core belief. The belief revision framework provides information about how the conditionals came into being, so that they can be modified in a structured way. It would be interesting to see if something similar can be done in other settings where conditional information cannot be completed in a satisfactory way.

Bibliography

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

[2] D. E. Appelt and M. E. Pollack. Weighted abduction for plan ascription. *User Modeling and User-Adapted Interaction*, 2(1-2):1–25, 1991.

[3] G. Aucher. *Perspectives on belief and change*. PhD thesis, University of Otago, Dunedin, New Zealand and Université Paul Sabatier, Toulouse, France, 2008, submitted.

[4] S. Benferhat, D. Dubois, and H. Prade. Some syntactic approaches to the handling of inconsistent knowledge bases: A comparative study part 1: The flat case. *Studia Logica*, 58(1):17–45, 1997.

[5] S. Benferhat, D. Dubois, and H. Prade. Some syntactic approaches to the handling of inconsistent knowledge bases: a comparative study part 2: the prioritized case. In E. Orłowska, editor, *Logic at Work: Essays Dedicated to the Memory of Helen Rasiowa*, volume 24, pages 437–511. Physica-Verlag, 1999.

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

[7] R. Booth. The lexicographic closure as a revision process. *Journal of Applied Non-Classical Logics*, 11(1-2):35–58, 2001.

[8] 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.

[9] R. Booth and T. Meyer. Admissible and restrained revision. *Journal of Artificial Intelligence Research*, 26:127–151, 2006.

[10] R. Booth, T. Meyer, and K. Wong. A bad day surfing is better than a good day working: How to revise a total preorder. In *Proceedings of KR'06*, pages 230–238, 2006.

[11] R. Booth and A. Nittka. Beyond the rational explanation. In *Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics*, number 05321 in Dagstuhl Seminar Proceedings, 2005.

[12] R. Booth and A. Nittka. Reconstructing an agent's epistemic state from observations. In *Proceedings of IJCAI'05*, pages 394–399, 2005.

[13] R. Booth and A. Nittka. Reconstructing an agent's epistemic state from observations about its beliefs and non-beliefs. *Journal of Logic and Computation*, 2008.

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

[15] C. Boutilier. Revision sequences and nested conditionals. In *Proceedings of IJCAI'93*, pages 519–525, 1993.

[16] C. Boutilier and V. Becher. Abduction as belief revision. *Artificial Intelligence*, 77(1):43–94, 1995.

[17] Craig Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25(3):263–305, 1996.

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

[19] G. Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *Proceedings of IJCAI'89*, pages 1043–1048, 1989.

[20] S. Chopra, J. Heidema, and T. Meyer. Some logics of belief and disbelief. In *Proceedings of NMR'02*, pages 25–32, 2002.

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

[22] 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.

[23] L. F. 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.

[24] J. P. Delgrande, D. Dubois, and J. Lang. Iterated revision as prioritized merging. In *Proceedings of KR'06*, pages 210–220, 2006.

[25] J. P. Delgrande, A. C. Nayak, and M. Pagnucco. Gricean belief change. *Studia Logica*, 79(1):97–113, 2005.

[26] J. P. Delgrande and T. Schaub. A consistency-based approach for belief change. *Artificial Intelligence*, 151(1-2):1–41, 2003.

[27] D. Dubois, J. Lang, and H. Prade. Possibilistic logic. In D. Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning*, pages 439–513. Oxford University Press, Oxford, 1994.

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

[29] N. Friedman and J. Y. Halpern. Modeling belief in dynamic systems, part I: Foundations. *Artificial Intelligence*, 95(2):257–316, 1997.

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

[31] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, 1988.

[32] M. R. Garey and D. S. Johnson. *Computers and Intractability — A Guide to the Theory of NP-Completeness*. Freeman, San Francisco, 1979.

[33] M. Gelfond and V. Lifschitz. Representing action and change by logic programs. *Journal of Logic Programming*, 17(2/3&4):301–321, 1993.

[34] M. L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, 1986.

[35] M. Goldszmidt and J. Pearl. On the relation between rational closure and system Z. In *Proceedings of NMR'90*, pages 130–140, 1990.

[36] A. Grove. Two modelings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[37] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach - part II: Explanations. In *Proceedings of IJCAI'01*, pages 27–34, 2001.

[38] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach: Part 1: Causes. In *Proceedings of UAI'01*, pages 194–202, 2001.

[39] S. O. Hansson. A survey of non-prioiritized belief revision. *Erkenntnis*, 50:413–427, 1999.

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

[41] A. Hunter. Adding modal operators to the action language A. In *Proceedings of NMR'04*, pages 219–226, 2004.

[42] A. Hunter and J. P. Delgrande. Using ranking functions to determine plausible action histories. In *Proceedings of NRAC'05*, pages 59–64, 2005.

[43] A. Hunter and J. P. Delgrande. An action description language for iterated belief change. In *Proceedings of IJCAI'07*, pages 2498–2503, 2007.

[44] Y. Jin and M. Thielscher. Representing beliefs in the fluent calculus. In *Proceedings of ECAI'04*, pages 823–827, 2004.

[45] Y. Jin and M. Thielscher. Iterated belief revision, revised. *Artificial Intelligence*, 171(1):1–18, 2007.

[46] H. Katsuno and A. O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1992.

[47] H. Katsuno and K. Satoh. A unified view of consequence relation, belief revision and conditional logic. In *Proceedings of IJCAI'91*, pages 406–412, 1991.

[48] H. Kautz. A circumscriptive theory of plan recognition. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 105–133. MIT Press, Cambridge, MA, 1990.

[49] H. A. Kautz. A formal theory of plan recognition and its implementation. In J. F. Allen, H. A. Kautz, R. Pelavin, and J. Tenenberg, editors, *Reasoning About Plans*, pages 69–125. Morgan Kaufmann Publishers, San Mateo (CA), USA, 1991.

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

[51] K. Konolige and M. E. Pollack. Ascribing plans to agents. In *Proceedings of IJCAI'89*, pages 924–930, 1989.

[52] R. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4(1):67–95, 1986.

[53] S. Kraus, D. J. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.

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

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

[56] F. Lévy and J. Quantz. Representing beliefs in a situated event calculus. In *Proceedings of ECAI'98*, pages 547–551, 1998.

[57] D. K. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, 1973.

[58] P. Liberatore and M. Schaerf. BReLS: A system for the integration of knowledge bases. In *Proceedings of KR'00*, pages 145–152, 2000.

[59] D. Lorenzo. Learning non-monotonic causal theories from narratives of actions. In *Proceedings of NMR'02*, pages 349–355, 2002.

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

[61] D. Makinson and P. Gärdenfors. Relations between the logic of theory change and nonmonotonic logic. In *Proceedings of the Workshop on The Logic of Theory Change*, pages 185–205, London, UK, 1991. Springer-Verlag.

[62] Y. Martin, I. Narasamdya, and M. Thielscher. Knowledge of other agents and communicative actions in the fluent calculus. In *Proceedings of KR'04*, pages 623–633, 2004.

[63] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

[64] T. Meyer, A. Ghose, and S. Chopra. Non-prioritised ranked belief change. In *Proceedings of TARK'01*, 2001.

[65] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19/20:629–679, 1994.

[66] A. C. Nayak. Foundational belief change. *Journal of Philosophical Logic*, 23(5):495–533, 1994.

[67] A. C. Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41:353–390, 1994.

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

[69] B. Nebel. Syntax-based approaches to belief revision. In P. Gärdenfors, editor, *Belief Revision*, volume 29, pages 52–88. Cambridge University Press, Cambridge, UK, 1992.

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

[71] B. Nebel. How hard is it to revise a belief base? In D. Dubois and H. Prade, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Volume 3: Belief Change*, pages 77–145. Kluwer Academic Publishers, Dordrecht, 1998.

[72] A. Nittka. A 3-valued approach to disbelief. Diplomarbeit, Leipzig University, 2003.

[73] A. Nittka. Reasoning about an agent based on its revision history with missing inputs. In *Proceedings of JELIA'06*, pages 373–385, 2006.

[74] A. Nittka and R. Booth. A method for reasoning about other agents' beliefs from observations. In *Formal Models of Belief Change in Rational Agents*, number 07351 in Dagstuhl Seminar Proceedings, 2007.

[75] A. Nittka and R. Booth. A method for reasoning about other agents' beliefs from observations. *Texts in Logic and Games*, 2008.

[76] C. M. Papadimitriou. *Computational complexity*. Addison-Wesley, 1994.

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

[78] G. Paul. Approaches to abductive reasoning: An overview. *Artificial Intelligence Review*, 7:109–152, 1993.

[79] J. Pearl. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In *Proceedings of TARK'90*, pages 121–135, 1990.

[80] C. S. Peirce. Abduction and induction. In J. Buchler, editor, *Philosophical Writings of Peirce*, pages 150–156. Dover Books, New York, 1955.

[81] R. Pino Pérez and C. Uzcátegui. Preferences and explanations. *Artificial Intelligence*, 149(1):1–30, 2003.

[82] L. Perrussel and J.-M. Thévenin. A logical approach for describing (dis)belief change and message processing. In *Proceedings of AAMAS'04*, pages 614–621, 2004.

[83] L. Perrussel, J.-M. Thévenin, and T. Meyer. Mutual enrichment through nested belief change. In *Proceedings of AAMAS'06*, pages 226–228, 2006.

[84] G. Pinkas and R. P. Loui. Reasoning from inconsistency: A taxonomy of principles for resolving conflict. In *Proceedings of KR'92*, pages 709–719, 1992.

[85] R. Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, pages 359–380. Academic Press, San Diego, CA, 1991.

[86] E. Sandewall. *Features and fluents (vol. 1): the representation of knowledge about dynamical systems*. Oxford University Press, Inc., New York, NY, USA, 1994.

[87] C. F. Schmidt, N. S. Sridharan, and J. L. Goodson. The plan recognition problem: An intersection of psychology and artificial intelligence. *Artificial Intelligence*, 11:45–83, 1978.

[88] S. Shapiro, M. Pagnucco, Y. Lespérance, and H. J. Levesque. Iterated belief change in the situation calculus. In *Proceedings of KR'00*, pages 527–538, 2000.

[89] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper and B. Skyrms, editors, *Causation in decision, belief change, and statistics*, volume II, pages 105–134. Kluwer Academic Publishers, 1988.

[90] M. Thielscher. Introduction to the fluent calculus. *Electronic Transactions on Artificial Intelligence*, 2:179–192, 1998.

[91] M. Thielscher. From situation calculus to fluent calculus: State update axioms as a solution to the inferential frame problem. *Artificial Intelligence*, 111(1-2):277–299, 1999.

[92] L. Tuan, C. Baral, X. Zhang, and T. C. Son. Regression with respect to sensing actions and partial states. In *Proceedings of AAAI'04*, pages 556–561, 2004.

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

[94] K. W. Wagner and G. Wechsung. *Computational Complexity*. Reidel, Dordrecht, 1986.

[95] B. Walliser, D. Zwirn, and H. Zwirn. Abductive logics in a belief revision framework. *Journal of Logic, Language and Information*, 14(1):87–117, 2005.

Appendix A

Proofs

A.1 Proofs from Chapter 2

Proposition 2.2. *$f(\beta_k, \ldots, \beta_1) \equiv \bot$ if and only if $\beta_1 \equiv \bot$.*

Proof. Induction over the length of the sequence. $f(\beta_1) = \beta_1$ (Definition 2.1) is inconsistent iff β_1 is inconsistent. Assume $f(\beta_i, \ldots, \beta_1) \equiv \bot$. This is the case iff $\beta_1 \equiv \bot$ for any sequence $(\beta_i, \ldots, \beta_1)$. By Proposition 2.3 $f(\beta_{i+1}, \ldots, \beta_1) = f(\beta_{i+1}, f(\beta_i, \ldots, \beta_1))$. If $\beta_1 \equiv \bot$ then by our assumption $f(\beta_i, \ldots, \beta_1) \equiv \bot$ and hence $f(\beta_{i+1}, \ldots, \beta_1) = f(\beta_i, \ldots, \beta_1)$ is inconsistent (by Definition 2.1). If β_1 is consistent then so is $f(\beta_i, \ldots, \beta_1)$ using the assumption. If $\beta_{i+1} \wedge f(\beta_i, \ldots, \beta_1) \nvdash \bot$ then by Definition 2.1 $f(\beta_{i+1}, \ldots, \beta_1) = \beta_{i+1} \wedge f(\beta_i, \ldots, \beta_1)$ is consistent, otherwise $f(\beta_{i+1}, \ldots, \beta_1) = f(\beta_i, \ldots, \beta_1)$ is consistent, as $f(\beta_i, \ldots, \beta_1)$ is. □

Proposition 2.3. *For all formulae β, β_i and all sequences σ and σ':*
(i) $f(\beta \cdot \sigma) = f(\beta, f(\sigma))$, implying
(ii) $f(\beta_k, \ldots, \beta_1) = f(\beta_k, f(\beta_{k-1}, f(\ldots, f(\beta_1)\ldots)))$ and
(iii) $f(\sigma \cdot \sigma') = f(\sigma \cdot f(\sigma'))$.

Proof. First note that $f(f(\sigma)) = f(\sigma)$, as $f(\sigma)$ is a single formula and thus falls under the first case in Definition 2.1.

By Definition 2.1, $f(\beta \cdot \sigma)$ is $\beta \wedge f(\sigma)$ if $\beta \wedge f(\sigma) \nvdash \bot$, $f(\sigma)$ otherwise. $f(\beta, f(\sigma))$ is $\beta \wedge f(f(\sigma)) = \beta \wedge f(\sigma)$ if $\beta \wedge f(\sigma) \nvdash \bot$, $f(f(\sigma)) = f(\sigma)$ otherwise. So the two are obviously identical. (ii) and (iii) are proved by iteratively applying (i). □

Proposition 2.4. *If $\beta \equiv \beta'$ then $f(\alpha, \beta) \equiv f(\alpha, \beta')$ and $f(\beta, \alpha) \equiv f(\beta', \alpha)$.*

Proof. By Proposition 2.3 $f(\alpha, \beta) = f(\alpha, f(\beta))$. $f(\beta) = \beta \equiv \beta' = f(\beta')$ — in the second case, we have $f(\lambda, \alpha) = f(\lambda, f(\alpha))$ for any λ — and obviously $\alpha \wedge \beta \nvdash \bot$ iff $\alpha \wedge \beta' \nvdash \bot$. □

Proposition 2.6. *$f(\beta_k, \ldots, \beta_1) \vdash \beta_i$ or $f(\beta_k, \ldots, \beta_1) \vdash \neg\beta_i$ for all $1 \leq i \leq k$.*

Proof. For inconsistent β_1 this is immediate, as then by Proposition 2.2 $f(\beta_k, \ldots, \beta_1) \equiv \bot$, so consider consistent β_1. $f(\beta_k, \ldots, \beta_1) = f((\beta_k, \ldots, \beta_{i+1}) \cdot f(\beta_i \cdot f(\beta_{i-1}, \ldots, \beta_1)))$ (by Proposition 2.3). If $\beta_i \wedge f(\beta_{i-1}, \ldots, \beta_1) \nvdash \bot$ then $f(\beta_i \cdot f(\beta_{i-1}, \ldots, \beta_1))$ will be this formula which clearly entails β_i and hence $f(\beta_k, \ldots, \beta_1) \vdash \beta_i$. If $\beta_i \wedge f(\beta_{i-1}, \ldots, \beta_1) \vdash \bot$ then $f(\beta_{i-1}, \ldots, \beta_1) \vdash \neg\beta_i$ and hence $f(\beta_k, \ldots, \beta_1) \vdash \neg\beta_i$. □

Proposition 2.7. *For consistent α: If $\alpha \vdash \beta$ then $f(\alpha, \beta) \equiv f(\alpha)$.*

Proof. If α is consistent then so is β. Hence $f(\alpha, \beta) = \alpha \wedge \beta$ and as $\alpha \vdash \beta$ we know $\alpha \wedge \beta \equiv \alpha = f(\alpha)$. □

Proposition 2.8. *For all sequences of formulae σ, ρ_1 and ρ_2 (ρ_2 being non-empty) we have $f(\rho_1 \cdot \top \cdot \rho_2 \cdot \sigma) \equiv f(\rho_1 \cdot \rho_2 \cdot \sigma)$ and $f(\rho_1 \cdot \bot \cdot \rho_2 \cdot \sigma) \equiv f(\rho_1 \cdot \rho_2 \cdot \sigma)$.*

Proof.

$$\begin{aligned} f(\rho_1 \cdot \top \cdot \rho_2 \cdot \sigma) &\equiv f(\rho_1 \cdot f(\top \cdot f(\rho_2 \cdot \sigma))) \quad \text{Proposition 2.3} \\ &\equiv f(\rho_1 \cdot f(\rho_2 \cdot \sigma)) \quad \forall \psi : \psi \wedge \top \equiv \psi \\ &\equiv f(\rho_1 \cdot \rho_2 \cdot \sigma) \quad \text{Proposition 2.3} \end{aligned}$$

$$\begin{aligned} f(\rho_1 \cdot \bot \cdot \rho_2 \cdot \sigma) &\equiv f(\rho_1 \cdot f(\bot \cdot f(\rho_2 \cdot \sigma))) \quad \text{Proposition 2.3} \\ &\equiv f(\rho_1 \cdot f(\rho_2 \cdot \sigma)) \quad \forall \psi : \psi \wedge \bot \vdash \bot \\ &\equiv f(\rho_1 \cdot \rho_2 \cdot \sigma) \quad \text{Proposition 2.3} \end{aligned}$$

□

Proposition 2.9. *If $\alpha \vdash \neg\beta$ then $f(\rho \cdot \beta \cdot \sigma \cdot \alpha) \equiv f(\rho \cdot \sigma \cdot \alpha)$ for all sequences ρ, σ.*

Proof. $f(\rho \cdot \beta \cdot \sigma \cdot \alpha) = f(\rho \cdot f(\beta \cdot \sigma \cdot \alpha))$ and $f(\rho \cdot \sigma \cdot \alpha) = f(\rho \cdot f(\sigma \cdot \alpha))$ (Proposition 2.3), so if we can show $f(\beta \cdot \sigma \cdot \alpha) = f(\sigma \cdot \alpha)$ we are done. By Definition 2.1 and $\alpha \vdash \neg\beta$ we know $f(\sigma \cdot \alpha) \vdash \neg\beta$, so $f(\beta \cdot \sigma \cdot \alpha) = f(\beta \cdot f(\sigma \cdot \alpha)) = f(\sigma \cdot \alpha)$. □

Proposition 2.10. *If $f(\sigma_1 \cdot \sigma_2) \nvdash \neg\alpha$ then $f(\sigma_1 \cdot \alpha \cdot \sigma_2) \equiv f(\sigma_1 \cdot \sigma_2) \wedge \alpha$.*

Proof. By Proposition 2.3 we know that $f(\sigma_1 \cdot \alpha \cdot \sigma_2) = f(\sigma_1 \cdot f(\alpha, f(\sigma_2)))$ and that $f(\sigma_1 \cdot \sigma_2) = f(\sigma_1 \cdot f(\sigma_2))$. Hence both collect the same formulae from σ_2. It follows from $f(\sigma_1 \cdot \sigma_2) \nvdash \neg\alpha$ that $f(\sigma_2) \nvdash \neg\alpha$ and hence $f(\alpha, f(\sigma_2)) = \alpha \wedge f(\sigma_2)$.

So, if we can show that $f(\sigma_1 \cdot \alpha \wedge f(\sigma_2))$ and $f(\sigma_1 \cdot f(\sigma_2))$ also collect the same formulae from σ_1, the proposition immediately follows. This is because the order of the elements in a conjunction does not matter.

The argument that the same formulae from σ_1 are chosen is an inductive one. Assume both have so far collected the same elements from some suffix of σ_1, their conjunction being denoted by λ. If no element has been accepted so far then $\lambda \equiv \top$. The next formula to be considered is β. Assume $f(\sigma_1 \cdot f(\sigma_2))$ rejects β, i.e., $f(\sigma_2) \wedge \lambda \vdash \neg\beta$. This implies $\alpha \wedge f(\sigma_2) \wedge \lambda \vdash \neg\beta$ and hence $f(\sigma_1 \cdot \alpha \wedge f(\sigma_2))$ also rejects β. However, if $f(\sigma_1 \cdot f(\sigma_2))$ accepts β, from $f(\sigma_1 \cdot \sigma_2) \nvdash \neg\alpha$ we know $f(\sigma_2) \wedge \lambda \wedge \beta \nvdash \neg\alpha$ and hence $\alpha \wedge f(\sigma_2) \wedge \lambda \nvdash \neg\beta$. And so, $f(\sigma_1 \cdot \alpha \wedge f(\sigma_2))$ also accepts β. □

Proposition 2.12. *If* $f(\sigma \cdot \alpha) \nvdash \neg\beta$ *then* $f(\sigma \cdot \alpha \wedge \beta) \equiv f(\sigma \cdot \alpha) \wedge \beta$

Proof. As $f(\sigma \cdot \alpha) \nvdash \neg\beta$ we know $\alpha \nvdash \neg\beta$ and hence $f(\alpha, \beta) = \alpha \wedge \beta$. Proposition 2.10 yields that $f(\sigma \cdot (\alpha, \beta)) \equiv f(\sigma \cdot \alpha) \wedge \beta$. But $f(\sigma \cdot (\alpha, \beta)) = f(\sigma \cdot f(\alpha, \beta)) = f(\sigma \cdot \alpha \wedge \beta)$. □

Proposition 2.13. *(i) If* $f(\sigma) \vdash \alpha$ *then* $f(\alpha \rightarrow \beta \cdot \sigma) \equiv f(\beta \cdot \sigma)$
(ii) If $f(\sigma) \vdash \neg\alpha$ *then* $f(\alpha \rightarrow \beta \cdot \sigma) \equiv f(\sigma)$

Proof. $f(\beta \cdot \sigma) = f(\beta \cdot f(\sigma))$. If $f(\sigma) \vdash \neg\beta$ then $f(\sigma) \vdash \alpha \wedge \neg\beta$, i.e., $f(\sigma) \vdash \neg(\alpha \rightarrow \beta)$. Hence $f(\beta \cdot \sigma) = f(\sigma) = f(\alpha \rightarrow \beta \cdot \sigma)$.

If $f(\sigma) \nvdash \neg\beta$ then $f(\sigma) \nvdash \neg(\alpha \rightarrow \beta)$. Consequently, $f(\beta \cdot \sigma) = \beta \wedge f(\sigma)$ and $f(\alpha \rightarrow \beta \cdot \sigma) = (\alpha \rightarrow \beta) \wedge f(\sigma)$. As $f(\sigma) \vdash \alpha$ this is equivalent to $(\alpha \rightarrow \beta) \wedge f(\sigma) \wedge \beta \equiv f(\sigma) \wedge \beta$ which proves (i).

If $f(\sigma) \vdash \neg\alpha$ then $f(\sigma) \nvdash \neg(\alpha \rightarrow \beta)$ and by Proposition 2.10 $f(\alpha \rightarrow \beta \cdot \sigma) = f(\sigma) \wedge (\alpha \rightarrow \beta)$, but this is equivalent to $f(\sigma)$ as the implication is already entailed. □

Proposition 2.15. *Either* $f(\sigma \cdot \rho \cdot \alpha) \vdash \neg f(\sigma \cdot \alpha)$ *or* $f(\sigma \cdot \rho \cdot \alpha) \vdash f(\sigma \cdot \alpha)$

Proof. $f(\sigma \cdot \rho \cdot \alpha) \equiv f(\sigma \cdot \alpha \cdot \rho \cdot \alpha) = f(\sigma \cdot \alpha \cdot f(\rho \cdot \alpha))$. The first equivalence is due to Proposition 2.10. α will be entailed, so adding it somewhere in the sequence will have no impact. The second is a consequence of Proposition 2.3. Now, either $f(\sigma \cdot \alpha) \wedge f(\rho \cdot \alpha)$ is consistent or it inconsistent. In the first case Proposition 2.10 tells us that $f(\sigma \cdot \alpha \cdot f(\rho \cdot \alpha)) \equiv f(\sigma \cdot \alpha) \wedge f(\rho \cdot \alpha) \vdash f(\sigma \cdot \alpha)$, while in the second case we get $f(\sigma \cdot \alpha \cdot f(\rho \cdot \alpha)) \vdash f(\rho \cdot \alpha) \vdash \neg f(\sigma \cdot \alpha)$. □

Proposition 2.20. $Bel([(\varphi_1, \ldots, \varphi_n), \blacktriangle]) \vdash \varphi_i$ *or* $Bel([(\varphi_1, \ldots, \varphi_n), \blacktriangle]) \vdash \neg\varphi_i$ *for all* $1 \leq i \leq n$.

Proof. As $Bel([(\varphi_1, \ldots, \varphi_n), \blacktriangle]) = Cn(f(\varphi_1, \ldots, \varphi_n, \blacktriangle))$ this follows immediately from Proposition 2.6. □

Proposition 2.22. *If* $Bel([\rho, \blacktriangle]) \vdash \varphi$ *then* $Bel([\rho, \blacktriangle] * \varphi) = Bel([\rho, \blacktriangle])$.

Proof. This is obvious for inconsistent $\blacktriangle$, as then in both cases the belief set is inconsistent, so consider consistent $\blacktriangle$. It suffices to show that $f(\rho \cdot (\varphi, \blacktriangle)) \equiv f(\rho \cdot \blacktriangle)$. $Bel([\rho, \blacktriangle]) \vdash \varphi$ yields that $f(\rho \cdot \blacktriangle) \vdash \varphi$. As $\blacktriangle \nvdash \bot$, we know that $f(\rho \cdot \blacktriangle) \nvdash \neg\varphi$. By Proposition 2.10 $f(\rho \cdot (\varphi, \blacktriangle)) \equiv f(\rho \cdot \blacktriangle) \wedge \varphi \equiv f(\rho \cdot \blacktriangle)$ (as the last formula already entails φ). □

Proposition 2.24. $Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = Bel([\rho, \blacktriangle] * f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$

Proof. Obvious for inconsistent $\blacktriangle$ as then in both cases the beliefs are inconsistent. So assume $\blacktriangle \nvdash \bot$, in which case

$$\begin{aligned}
Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) &= Bel([\rho \cdot (\varphi_1, \ldots, \varphi_i), \blacktriangle]) \quad \text{Definition 2.16} \\
&= Cn(f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))) \quad \text{Definition 2.17} \\
&= Cn(f(\rho \cdot f(\varphi_1, \ldots, \varphi_i, \blacktriangle))) \quad \text{Proposition 2.3} \\
&= Cn(f(\rho \cdot f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \cdot \blacktriangle)) \quad \text{Proposition 2.7} \\
&= Bel([\rho \cdot f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle]) \quad \text{Definition 2.17} \\
&= Bel([\rho, \blacktriangle] * f(\varphi_1, \ldots, \varphi_i, \blacktriangle)) \quad \text{Definition 2.16.}
\end{aligned}$$

□

Proposition 2.26. *If* $\blacktriangle \nvdash (\varphi_2 \rightarrow \neg\varphi_1)$ *then* $Bel([\rho, \blacktriangle] * \varphi_1 * \varphi_2) \vdash \varphi_1$

Proof. $Bel([\rho, \blacktriangle] * \varphi_1 * \varphi_2) = Cn(f(\rho \cdot (\varphi_1, \varphi_2, \blacktriangle)))$. $f(\rho \cdot (\varphi_1, \varphi_2, \blacktriangle)) = f(\rho \cdot f(\varphi_1, \varphi_2, \blacktriangle))$. Now, $\blacktriangle \nvdash (\varphi_2 \rightarrow \neg\varphi_1)$ so $\blacktriangle \nvdash \neg\varphi_2$ and $\blacktriangle \wedge \varphi_2 \nvdash \neg\varphi_1$. Hence $f(\varphi_1, \varphi_2, \blacktriangle) = \varphi_1 \wedge \varphi_2 \wedge \blacktriangle$ which is entailed by $f(\rho \cdot f(\varphi_1, \varphi_2, \blacktriangle))$ by definition. □

Proposition 2.27. *For any epistemic state* $[\rho, \blacktriangle]$*, there exists an epistemic state* $[\sigma, \blacktriangle]$*, such that* σ *is a logical chain and for all* φ*:*

$$Bel([\rho, \blacktriangle] * \varphi) = Bel([\sigma, \blacktriangle] * \varphi)$$

Proof. Proposition 2.41 yields that $[\rho, \blacktriangle]$ defines a rational consequence relation $\Rightarrow_{[\rho, \blacktriangle]}$. Theorem 1 in [28] tells us that for every rational consequence relation $\Rightarrow$ there is a logical chain that induces exactly the same relation. So we know that there is a logical chain σ for with $\Rightarrow_{[\rho, \blacktriangle]} = \Rightarrow_\sigma$.

$\Rightarrow_\sigma$ is defined by $\lambda \Rightarrow_\sigma \mu$ iff there is an element β in σ, such that $\beta \wedge \lambda$ is consistent, but $\beta \wedge \lambda \wedge \neg\mu$ is inconsistent (and hence $\beta \wedge \lambda \vdash \mu$). It is easy to show that $\Rightarrow_{[\sigma,\blacktriangle]} = \Rightarrow_{[\rho,\blacktriangle]}$, i.e., using the logical chain instead of ρ in the epistemic state yields the same rational consequence relation. This immediately yields $Bel([\rho, \blacktriangle] * \varphi) = Bel([\sigma, \blacktriangle] * \varphi)$ (Definition 2.40).

If $\blacktriangle \vdash \neg\varphi$, then $\varphi \Rightarrow_{[\sigma,\blacktriangle]} \mu$ and $\varphi \Rightarrow_{[\rho,\blacktriangle]} \mu$ for all μ. So let $\blacktriangle \wedge \varphi$ be consistent.

If $\varphi \Rightarrow_{[\rho,\blacktriangle]} \mu$ then $\blacktriangle \wedge \varphi \Rightarrow_{[\rho,\blacktriangle]} \mu$ (as $f(\rho \cdot (\varphi, \blacktriangle)) \equiv f(\rho \cdot (\blacktriangle \wedge \varphi, \blacktriangle))$ which means it does not matter whether the epistemic state is revised by φ or by $\blacktriangle \wedge \varphi$). $\Rightarrow_{[\rho,\blacktriangle]} = \Rightarrow_\sigma$ implies $\blacktriangle \wedge \varphi \Rightarrow_\sigma \mu$ and hence there is a β in σ such that $\beta \wedge \blacktriangle \wedge \varphi$ is consistent and entails μ. This means $f(\sigma \cdot (\blacktriangle \wedge \varphi, \blacktriangle)) \vdash \mu$ (β will definitely be selected by f and all other elements are either inconsistent with $\blacktriangle \wedge \varphi$ or entailed by β as σ is a logical chain). But this means $\blacktriangle \wedge \varphi \Rightarrow_{[\sigma,\blacktriangle]} \mu$ which implies $\varphi \Rightarrow_{[\sigma,\blacktriangle]} \mu$ ($f(\sigma \cdot (\blacktriangle \wedge \varphi, \blacktriangle)) \equiv f(\sigma \cdot (\varphi, \blacktriangle))$ same argument as above).

Now, let $\varphi \Rightarrow_{[\sigma,\blacktriangle]} \mu$. We know $f(\sigma \cdot (\varphi, \blacktriangle)) \equiv f(\sigma \cdot (\blacktriangle \wedge \varphi, \blacktriangle)) \equiv f(\sigma \cdot \blacktriangle \wedge \varphi)$. As σ is a logical chain this is equivalent to $\beta \wedge \blacktriangle \wedge \varphi$ where β is the logically strongest element of σ consistent with $\blacktriangle \wedge \varphi$. From the conditional belief we now know that $\beta \wedge \blacktriangle \wedge \varphi \vdash \mu$ and hence $\blacktriangle \wedge \varphi \Rightarrow_\sigma \mu$ which implies that $\blacktriangle \wedge \varphi \Rightarrow_{[\rho,\blacktriangle]} \mu$. And as $f(\rho \cdot (\varphi, \blacktriangle)) \equiv f(\rho \cdot (\blacktriangle \wedge \varphi, \blacktriangle))$ we get $\varphi \Rightarrow_{[\rho,\blacktriangle]} \mu$. □

Proposition 2.31. *The decision problem of whether $[\rho, \blacktriangle]$ explains an observation o is Δ_2^P-complete.*

Proof. Given die observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$, with a polynomial number of NP-oracle calls (satisfiability tests) we can calculate all $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))$. As f returns the conjunction of a subset formulae from the argument sequence, the result is bounded by the size of the input of the decision problem. Then a polynomial number of tests whether the θ_i and δ are entailed suffice. Consequently the problem is in Δ_2^P.

Nebel shows in [70, 71] that the entailment problem for linear base revision whether $(\alpha_1, \ldots, \alpha_m) *_L \alpha \vdash \beta$ is Δ_2^P-complete. The hardness-part of that proof uses a reduction such that α is guaranteed to be consistent. This allows us to extend the reduction given there (polynomially) to the question of whether $[(\alpha_1, \ldots, \alpha_m), \top]$ is an explanation for $\langle(\alpha, \beta, \emptyset)\rangle$. As in this case $(\alpha_1, \ldots, \alpha_m) *_L \alpha \equiv f(\alpha_1, \ldots, \alpha_m, \alpha) \equiv f(\alpha_1, \ldots, \alpha_m, \alpha, \top)$ we immediately get that $(\alpha_1, \ldots, \alpha_m) *_L \alpha \vdash \beta$ if and only if $[(\alpha_1, \ldots, \alpha_m), \top]$ explains $\langle(\alpha, \beta, \emptyset)\rangle$. Consequently, checking whether $[\rho, \blacktriangle]$ is an explanation for o is also Δ_2^P-complete. □

Proposition 2.33. *If $[\rho, \blacktriangle]$ is an explanation for $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ then $[\rho \cdot (\varphi_1, \ldots, \varphi_{j-1}), \blacktriangle]$ explains $o' = \langle(\varphi_j, \theta_j, D_j), \ldots, (\varphi_{j+k}, \theta_{j+k}, D_{j+k})\rangle$, $1 \leq j + k \leq n$.*

Proof. $Bel([\rho \cdot (\varphi_1, \ldots, \varphi_{j-1}), \blacktriangle] * \varphi_j * \ldots * \varphi_{j+k'}) = Bel([\rho, \blacktriangle] * \varphi_1 * \ldots * \varphi_{j+k'})$, i.e. the beliefs after having received the input $\varphi_{j+k'}$ for any $1 \leq k' \leq k$ are the same. So the proposition follows immediately from Definition 2.29. □

Proposition 2.35. *If $\blacktriangle_1$ and $\blacktriangle_2$ are o-acceptable then so is $\blacktriangle_1 \vee \blacktriangle_2$.*

Proof. Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation and $\blacktriangle_1$ and $\blacktriangle_2$ two o-acceptable cores, i.e., there are $\rho_1 = (\beta_{11}, \ldots, \beta_{1m_1})$ and $\rho_2 = (\beta_{21}, \ldots, \beta_{2m_2})$ such that $[\rho_1, \blacktriangle_1]$ and $[\rho_2, \blacktriangle_2]$ explain o. It suffices to show that there is a ρ such that $[\rho, \blacktriangle_1 \vee \blacktriangle_2]$ explains o. We will show that $\rho = (\neg\blacktriangle_1 \rightarrow \beta_{21}, \ldots, \neg\blacktriangle_1 \rightarrow \beta_{2m_2}, \blacktriangle_1 \rightarrow \beta_{11}, \ldots, \blacktriangle_1 \rightarrow \beta_{1m_1}, \blacktriangle_1)$ is such a sequence. In fact, we will show that $Bel([\rho, \blacktriangle_1 \vee \blacktriangle_2] * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_1, \blacktriangle_1] * \varphi_1 * \ldots * \varphi_i)$ or $Bel([\rho, \blacktriangle_1 \vee \blacktriangle_2] * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_2, \blacktriangle_2] * \varphi_1 * \ldots * \varphi_i)$ for all $1 \leq i \leq n$. Then the proposition immediately follows as $[\rho_1, \blacktriangle_1]$ and $[\rho_2, \blacktriangle_2]$ are explanations for o. Fixing an i we will now show that either $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)) \equiv f(\rho_1 \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_1))$ or $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2)) \equiv f(\rho_2 \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_2))$.

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

Proposition 2.13 then tells us how to treat the implications in ρ with respect to the original formulae in ρ_1 and ρ_2. In the first case where $\blacktriangle_1$ is entailed all $\blacktriangle_1 \rightarrow \beta_{1j}$ are treated exactly like the β_{1j} from ρ_1 and the $\neg\blacktriangle_1 \rightarrow \beta_{2k}$ from ρ_2 can be ignored. As a consequence, we get the same beliefs as for the epistemic state $[\rho_1, \blacktriangle_1]$. In the second case $\neg\blacktriangle_1$ is entailed and hence all $\blacktriangle_1 \rightarrow \beta_{1j}$ from ρ_1 can be ignored and all $\neg\blacktriangle_1 \rightarrow \beta_{2k}$ are treated exactly like the β_{2k} from ρ_2. Consequently, we get the same beliefs as for the epistemic state $[\rho_2, \blacktriangle_2]$.

First assume $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \vdash \blacktriangle_1$ which entails $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \nvdash \neg\blacktriangle_1$. By Proposition 2.10 $f(\blacktriangle_1, \varphi_1, \ldots, \varphi_i, \blacktriangle_1 \vee \blacktriangle_2) \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1, \blacktriangle_1 \vee \blacktriangle_2)$ which is equivalent to $f(\varphi_1, \ldots, \varphi_i, f(\blacktriangle_1, \blacktriangle_1 \vee \blacktriangle_2))$ and as $\blacktriangle_1 \vdash \blacktriangle_1 \vee \blacktriangle_2$ we can apply Proposition 2.7 to show that this is equivalent to $f(\varphi_1, \ldots, \varphi_i, \blacktriangle_1)$ as claimed.

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

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

Proposition 2.37. *The function $\blacktriangle_\vee(\cdot)$ satisfies the following properties of a function $\blacktriangle(\cdot)$ mapping observations to formulae, for any observations o, o':*

(Acceptability)	*If an o-acceptable core exists then $\blacktriangle(o)$ is o-acceptable.*
(Consistency)	*If $\blacktriangle(o) \not\equiv \bot$ then there is an o-acceptable core.*
(Right Monotony)	$\blacktriangle(o \cdot o') \vdash \blacktriangle(o)$
(Left Monotony)	$\blacktriangle(o' \cdot o) \vdash \blacktriangle(o)$

Proof. Consistency follows directly from the Definition 2.36 of $\blacktriangle_\vee(o)$ which can be consistent only if an o-acceptable core exists. Acceptability follows from Definition 2.36 and Proposition 2.35 which states that the disjunction of two o-acceptable cores is again o-acceptable.

Left Monotony and Right Monotony follow from Proposition 2.33 which states that an o-acceptable core is also acceptable for any sub-observation of o, and Definition 2.36 which entails that any o-acceptable core entails $\blacktriangle_\vee(o)$. In case there is no acceptable core for the extended observation, $\blacktriangle_\vee(\cdot)$ will return $\bot$ which entails any formula. □

Proposition 2.38. *Let $\blacktriangle(\cdot)$ be any function which returns a formula given any observation o. Then the following are equivalent:*
(i) $\blacktriangle(\cdot)$ satisfies Acceptability, Consistency and Right Monotony.
(ii) $\blacktriangle(\cdot)$ satisfies Acceptability, Consistency and Left Monotony.
(iii) $\blacktriangle(o) \equiv \blacktriangle_\vee(o)$ for all observations o.

Proof. Proposition 2.37 immediately yields (iii)→(i) and (iii)→(ii), so it suffices to show (i)→(iii) and (ii)→(iii).

Let $\blacktriangle(\cdot)$ satisfy Acceptability, Consistency and Right Monotony. Assume there is an observation o such that $\blacktriangle(o) = \psi$ and $\blacktriangle_\vee(o) = \lambda \not\equiv \psi$. Both λ and ψ must be consistent. If both are inconsistent then $\lambda \not\equiv \psi$ is violated. Without loss of generality, assume only ψ is inconsistent. $\lambda \not\equiv \bot$ and the property Consistency of $\blacktriangle_\vee(\cdot)$ imply the existence of an o-acceptable core, but ψ is not o-acceptable contradicting Acceptability of $\blacktriangle(\cdot)$, so ψ is consistent as well. As any o-acceptable core entails $\blacktriangle_\vee(o)$, we know $\psi \vdash \lambda$. But as $\lambda \not\equiv \psi$, we get $\lambda \nvdash \psi$.

Now consider $o' = o \cdot (\neg\psi, \neg\psi, \emptyset)$. λ is o'-acceptable. It explains the prefix o and does not prevent $\neg\psi$ from being introduced into the belief set upon receiving it, which is the only condition needed to satisfy the additional piece of observation $(\neg\psi, \neg\psi, \emptyset)$. Hence, there is an o'-acceptable core.

Due to Acceptability $\blacktriangle(o')$ must be consistent and due to Right Monotony $\blacktriangle(o') \vdash \psi$. Consequently, an agent with that core belief will not believe $\neg\psi$ upon receiving it as a revision input. Hence $\blacktriangle(o')$ is not $\langle(\neg\psi, \neg\psi, \emptyset)\rangle$-acceptable and by Proposition 2.33 $\blacktriangle(o')$ is not o'-acceptable. As a consequence $\blacktriangle(\cdot)$ violates Acceptability, leading to a contradiction. Hence, the assumed observation o for which $\blacktriangle(\cdot)$ and $\blacktriangle_\vee(\cdot)$ return different formulae does not exist.

Now let $\blacktriangle(\cdot)$ satisfy Acceptability, Consistency and Left Monotony. Assume there is an observation o such that $\blacktriangle(o) = \psi$ and $\blacktriangle_\vee(o) = \lambda \not\equiv \psi$. As above we have $\psi \vdash \lambda$ and $\lambda \nvdash \psi$, but we further have $\lambda \nvdash \neg\psi$ (otherwise $\psi \vdash \neg\psi$, but ψ is consistent).

Consider $o' = \langle(\neg\psi, \neg\psi, \emptyset), (\psi, \psi, \emptyset)\rangle \cdot o$. Due to Left Monotony $\blacktriangle(o') \vdash \psi$. Consequently, an agent with that core belief will not believe $\neg\psi$ upon receiving it as a revision input (if $\blacktriangle(o')$ is inconsistent it is not o'-acceptable, even if in this case $\neg\psi$ is believed). Hence $\blacktriangle(o')$ is not $\langle(\neg\psi, \neg\psi, \emptyset)\rangle$-acceptable and by Proposition 2.33 $\blacktriangle(o')$ is not o'-acceptable. We will show that λ is o'-acceptable. This then tells us that there is an o'-acceptable core. However, $\blacktriangle(o')$ is not o'-acceptable and hence $\blacktriangle(\cdot)$ violates Acceptability, leading to a contradiction. So the assumed observation o for which $\blacktriangle(\cdot)$ and $\blacktriangle_\vee(\cdot)$ return different formulae does not exist. We now need to show that λ is indeed o'-acceptable.

As this proof is a constructive one, we need to look into the observation we assumed to exist. Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be that observation and consequently $o' = \langle(\neg\psi, \neg\psi, \emptyset), (\psi, \psi, \emptyset), (\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. λ and ψ are both o-acceptable so there exist sequences $\rho_1 = (\beta_{11}, \ldots, \beta_{1m_1})$ and $\rho_2 = (\beta_{21}, \ldots, \beta_{2m_2})$ such that $[\rho_1, \lambda]$ and $[\rho_2, \psi]$ explain o. We will show that there is a sequence ρ such that $[\rho, \lambda]$ explains o'. $\rho = (\psi \to \beta_{21}, \ldots, \psi \to \beta_{2m_2}, \neg\psi \to \beta_{11}, \ldots, \neg\psi \to \beta_{1m_1})$ is such a sequence.

Note that $[\sigma, \lambda]$ explains the prefix $\langle(\neg\psi, \neg\psi, \emptyset), (\psi, \psi, \emptyset)\rangle$ of o' using *any* sequence σ. This is because that observation only requires $\neg\psi$ and ψ to be believed upon receiving them, but this is guaranteed as λ is consistent with both. We will show that for all the remaining inputs $\varphi_1, \ldots, \varphi_n$, $Bel([\rho, \lambda] * \neg\psi * \psi * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_1, \lambda] * \varphi_1 * \ldots * \varphi_i)$ or $Bel([\rho, \lambda] * \neg\psi * \psi * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_2, \psi] * \varphi_1 * \ldots * \varphi_i)$ which yields that $[\rho, \lambda]$ indeed explains o'. The argument is basically identical to that in the proof for Proposition 2.35.

$f(\neg\psi, \psi, \varphi_1, \ldots, \varphi_i, \lambda) \vdash \psi$ or $f(\neg\psi, \psi, \varphi_1, \ldots, \varphi_i, \lambda) \vdash \neg\psi$ because ψ is an element of the sequence and hence it is either collected yielding the first case or it is rejected yielding the second one (Proposition 2.6). For the first case, Proposition 2.10 tells us that

$f(\neg\psi, \psi, \varphi_1, \ldots, \varphi_i, \lambda) \equiv f(\neg\psi, \varphi_1, \ldots, \varphi_i, \psi, \lambda)$, but as $\psi \vdash \lambda$ which implies that $f(\psi, \lambda) \equiv \psi$ and $f(\neg\psi, \varphi_1, \ldots, \varphi_i, \psi, \lambda) \equiv f(\neg\psi, \varphi_1, \ldots, \varphi_i, f(\psi, \lambda))$ (Proposition 2.3), we get $f(\neg\psi, \psi, \varphi_1, \ldots, \varphi_i, \lambda) \equiv f(\neg\psi, \varphi_1, \ldots, \varphi_i, \psi)$. This definitely entails ψ, so the $\neg\psi$ in the beginning is irrelevant. Hence in this case $f(\neg\psi, \psi, \varphi_1, \ldots, \varphi_i, \lambda) \equiv f(\varphi_1, \ldots, \varphi_i, \psi)$.

In other words, before ρ is processed, a formula has been constructed that is equivalent to that which has been collected before processing ρ_2. Proposition 2.13 yields that the $\neg\psi \rightarrow \beta_{1j}$ can now be ignored when processing ρ and that the $\psi \rightarrow \beta_{2k}$ are treated exactly like the β_{2k} in ρ_2. Hence, in this case $Bel([\rho, \lambda] * \neg\psi * \psi * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_2, \psi] * \varphi_1 * \ldots * \varphi_i)$ as claimed.

In the second case ($f(\neg\psi, \psi, \varphi_1, \ldots, \varphi_i, \lambda) \vdash \neg\psi$), we know $f(\varphi_1, \ldots, \varphi_i, \lambda) \vdash \neg\psi$ (otherwise the next formula ψ is accepted). But this means $f(\neg\psi, \psi, \varphi_1, \ldots, \varphi_i, \lambda) \equiv f(\varphi_1, \ldots, \varphi_i, \lambda)$, i.e., before ρ is processed a formula has been constructed that is equivalent to that which has been collected before processing ρ_1. Proposition 2.13 yields that the $\psi \rightarrow \beta_{2j}$ can now be ignored when processing ρ and that the $\neg\psi \rightarrow \beta_{1k}$ are treated exactly like the β_{1k} in ρ_1 and hence, $Bel([\rho, \lambda] * \neg\psi * \psi * \varphi_1 * \ldots * \varphi_i) = Bel([\rho_1, \lambda] * \varphi_1 * \ldots * \varphi_i)$. □

Proposition 2.41. $\Rightarrow_{[\rho,\blacktriangle]}$ *is a rational consequence relation.*

Proof. We need to show that the following conditions are satisfied by $\Rightarrow_{[\rho,\blacktriangle]}$. These conditions are given in [54]. The notation we use is in line with that in [14]. Recall that $Bel([\rho, \blacktriangle] * \varphi) = Cn(f(\rho \cdot (\varphi, \blacktriangle)))$. For inconsistent $\blacktriangle$ we have $\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ for all θ and ϕ so the conditions are all satisfied trivially. So we can assume $\blacktriangle$ to be consistent. If $\blacktriangle \vdash \neg\theta$ most of the conditions are trivially satisfied as in those cases $\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ for any ϕ, so we give the proofs only if they are non-trivial and assume $\blacktriangle \nvdash \neg\theta$ if not stated otherwise.

$\theta \Rightarrow_{[\rho,\blacktriangle]} \theta$: $f(\theta, \blacktriangle) = \blacktriangle \wedge \theta$ already entails θ.

$\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ and $\theta \equiv \psi$ implies $\psi \Rightarrow_{[\rho,\blacktriangle]} \phi$: immediate from Proposition 2.4

$\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ and $\phi \models \psi$ implies $\theta \Rightarrow_{[\rho,\blacktriangle]} \psi$: immediate from closure of $Cn(\cdot)$.

$\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ and $\theta \Rightarrow_{[\rho,\blacktriangle]} \psi$ implies $\theta \Rightarrow_{[\rho,\blacktriangle]} \phi \wedge \psi$: immediate from closure of $Cn(\cdot)$.

$\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ and $\psi \Rightarrow_{[\rho,\blacktriangle]} \phi$ implies $\theta \vee \psi \Rightarrow_{[\rho,\blacktriangle]} \phi$: This is trivial if $\blacktriangle \vdash \neg\theta$ as then $f(\theta \vee \psi, \blacktriangle) \equiv f(\psi, \blacktriangle)$ — analogously for $\blacktriangle \vdash \neg\psi$. Otherwise it basically follows from $\lambda \wedge \theta \vdash \phi$ and $\lambda \wedge \psi \vdash \phi$ which implies $\lambda \wedge (\theta \vee \psi) \vdash \phi$. If $f(\rho \cdot (\theta \vee \psi, \blacktriangle))$ collects exactly the same formulae from ρ as both $f(\rho \cdot (\theta, \blacktriangle))$ and $f(\rho \cdot (\psi, \blacktriangle))$ then we can use this argument. However, if it collects the same formulae (their conjunction being λ) as both up to a particular point and then accepts a formula α (without loss of generality) $f(\rho \cdot (\theta, \blacktriangle))$ does not accept, we have $\lambda \wedge \theta \wedge \blacktriangle \vdash \neg\alpha$ so $\lambda \wedge \alpha \wedge \blacktriangle \vdash \neg\theta$ and hence $\lambda \wedge (\theta \vee \psi) \wedge \blacktriangle \wedge \alpha \equiv \lambda \wedge \psi \wedge \blacktriangle \wedge \alpha$. Then it follows immediately that $f(\rho \cdot (\theta \vee \psi, \blacktriangle)) \equiv f(\rho \cdot (\psi, \blacktriangle))$.

$\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ and $\theta \Rightarrow_{[\rho,\blacktriangle]} \psi$ implies $\theta \wedge \phi \Rightarrow_{[\rho,\blacktriangle]} \psi$: In case $\blacktriangle \vdash \neg\theta$ then also $\blacktriangle \vdash \neg(\theta \wedge \phi)$. Otherwise, $\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ implies $f(\rho \cdot (\theta, \blacktriangle)) \nvdash \neg\phi$ and consequently by Proposition 2.10 $f(\rho \cdot (\phi, \theta, \blacktriangle)) = f(\rho \cdot f(\phi, \theta, \blacktriangle)) \equiv f(\rho \cdot (\theta, \blacktriangle)) \wedge \phi$. We are done in case we can show $f(\theta \wedge \phi, \blacktriangle) \equiv f(\phi, \theta, \blacktriangle)$. $f(\phi, \theta, \blacktriangle) = \theta \wedge \phi \wedge \blacktriangle$ ($\blacktriangle \nvdash \neg\theta$ and $f(\theta, \blacktriangle) \nvdash \neg\phi$). Assume $f(\theta \wedge \phi, \blacktriangle) \neq \theta \wedge \phi \wedge \blacktriangle$, i.e., $\blacktriangle \vdash \neg(\theta \wedge \phi)$ and hence $\theta \wedge \blacktriangle = f(\theta, \blacktriangle) \vdash \neg\phi$, contradiction.

$\theta \Rightarrow_{[\rho,\blacktriangle]} \phi$ and $\theta \not\Rightarrow_{[\rho,\blacktriangle]} \neg\psi$ implies $\theta \wedge \psi \Rightarrow_{[\rho,\blacktriangle]} \phi$: we have $f(\rho \cdot (\theta, \blacktriangle)) \nvdash \neg\psi$ immediately and hence analogously to the above argument $f(\rho \cdot (\theta \wedge \psi, \blacktriangle)) \equiv f(\rho \cdot (\theta, \blacktriangle)) \wedge \psi$ which entails ϕ. □

Proposition 2.48. *$\rho_R(\mathcal{C}, \mathcal{N}) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0)$ is a logical chain, that is, $\bigwedge U_i \vdash \bigwedge U_{i+1}$ for $0 \leq i \leq m-1$.*

Proof. This follows immediately from $U_i \supseteq U_{i+1}$. That this indeed holds can be seen from the way $\mathcal{C}_{i+1}$, $\mathcal{N}_{i+1}$, and U_{i+1} are calculated from $\mathcal{C}_i$, $\mathcal{N}_i$. and U_i. Condition 3 in Definition 2.45 immediately yields $\mathcal{C}_i \supseteq \mathcal{C}_{i+1}$ and $\mathcal{N}_i \supseteq \mathcal{N}_{i+1}$. As a consequence $\tilde{\mathcal{C}}_i \supseteq \tilde{\mathcal{C}}_{i+1}$. We will now argue that in the least fixpoint construction any $\neg\lambda$ that is added to U_{i+1} will also be added to U_i, yielding the desired inclusion. Assume $\lambda \Rightarrow \mu \in \mathcal{N}_{i+1}$ is n-exceptional for $\tilde{\mathcal{C}}_{i+1}$, then it is also n-exceptional for $\tilde{\mathcal{C}}_i$ but by $\mathcal{N}_i \supseteq \mathcal{N}_{i+1}$ the conditional also belongs to $\mathcal{N}_i$. Hence, $\neg\lambda$ is added to both U_{i+1} and U_i. For an inductive argument assume that up to now every $\neg\lambda'$ that was added to U_{i+1} has also been added to U_i. So any additional conditional that is n-exceptional for U_{i+1} must again be n-exceptional for U_i. Hence, the negated antecedent is added to both, keeping the superset relation. □

Proposition 2.49. *Given $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and a core belief $\blacktriangle \not\equiv \bot$ let $\rho_R(o, \blacktriangle) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0)$ be the rational prefix of o with respect to $\blacktriangle$.*

If $\bigwedge U_m \equiv \top$ then $[\rho_R(o, \blacktriangle), \blacktriangle]$ explains o.

Proof. By Proposition 2.24 and Definition 2.17 it suffices to show that for all i we have $f(\rho_R(o, \blacktriangle) \cdot f(\iota_i \cdot \blacktriangle)) \vdash \theta_i$ and $f(\rho_R(o, \blacktriangle) \cdot f(\iota_i \cdot \blacktriangle)) \nvdash \delta$ for all $\delta \in D_i$. As $\rho_R(o, \blacktriangle)$ is a logical chain (Proposition 2.48) we know $f(\rho_R(o, \blacktriangle) \cdot f(\iota_i \cdot \blacktriangle)) \equiv \bigwedge U_j \wedge f(\iota_i \cdot \blacktriangle)$ where j is minimal such that this formula is consistent. Such a j must exist. Firstly, $\blacktriangle$ is consistent and hence $f(\iota_i \cdot \blacktriangle)$ is consistent. Secondly, at least $\bigwedge U_m \equiv \top$ is consistent with $f(\iota_i \cdot \blacktriangle)$. Thirdly, adding as conjuncts the $\bigwedge U_k$ with $k > j$ does not change the logical content of the formula as they are already entailed by $\bigwedge U_j$. We now need to show that $\bigwedge U_j \wedge f(\iota_i \cdot \blacktriangle)$ entails θ_i and does not entail any $\delta \in D_i$.

$f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i \in \mathcal{C}_{\blacktriangle}(o)$ and hence this conditional belongs to $\mathcal{C}_0$. As $f(\iota_i \cdot \blacktriangle)$ is inconsistent with all $\bigwedge U_k$ for $k < j$ we know $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ is p-exceptional for all U_k, $k < j$. Hence

$f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i \in \tilde{\mathcal{C}}_j$ and consequently $\bigwedge U_j \vdash f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i$ which yields that the conjunction entails θ_i.

Now assume $\bigwedge U_j \wedge f(\iota_i \cdot \blacktriangle) \vdash \delta$ for some $\delta \in D_i$. $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta \in \mathcal{N}_{\blacktriangle}(o)$ and hence it belongs to $\mathcal{C}_0$. As $f(\iota_i \cdot \blacktriangle)$ is inconsistent with all $\bigwedge U_k$ for $k < j$ we know that $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ is n-exceptional for all U_k, $k < j$. Hence $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta \in \mathcal{N}_j$. Our assumption yields that $\bigwedge U_j \wedge f(\iota_i \cdot \blacktriangle) \vdash \delta$, i.e., the conditional is n-exceptional for U_j, and by Definition 2.45 $\neg f(\iota_i \cdot \blacktriangle) \in U_j$ contradicting $\bigwedge U_j \wedge f(\iota_i \cdot \blacktriangle)$ is consistent. Hence, the assumption was wrong and such a δ does not exist. □

Proposition 2.50. *Given* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *and a core belief* $\blacktriangle \not\equiv \bot$ *let* $\rho_R(o, \blacktriangle) = (\bigwedge U_m, \bigwedge U_{m-1}, \ldots, \bigwedge U_0)$ *be the rational prefix of* o *with respect to* $\blacktriangle$.

If $\bigwedge U_m \not\equiv \top$ *then* $\blacktriangle$ *is not* o*-acceptable.*

Proof. We need to show two things. $[\rho_R(o, \blacktriangle), \blacktriangle]$ is not an explanation and there is no ρ such that $[\rho, \blacktriangle]$ explains o.

Assume $[\rho_R(o, \blacktriangle), \blacktriangle]$ explains o, that is, for all i we have $f(\rho_R(o, \blacktriangle) \cdot f(\iota_i \cdot \blacktriangle)) \vdash \theta_i$ and $f(\rho_R(o, \blacktriangle) \cdot f(\iota_i \cdot \blacktriangle)) \nvdash \delta$ for all $\delta \in D_i$. As $\bigwedge U_m \not\equiv \top$ we know that $U_m \neq \emptyset$. So $\mathcal{C}_m \neq \emptyset$ or $\mathcal{N}_m \neq \emptyset$. First assume that $\mathcal{C}_m$ contained at least one non-trivial conditional $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ where $f(\iota_i \cdot \blacktriangle) \nvdash \theta_i$, i.e., the corresponding material counterpart $f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i$ is not a tautology. As this conditional is in $\mathcal{C}_m$ it must have been p-exceptional for all U_j (including U_m as otherwise the rational prefix calculation would not have stopped). Hence $f(\rho_R(o, \blacktriangle) \cdot f(\iota_i \cdot \blacktriangle)) = f(\iota_i \cdot \blacktriangle) \nvdash \theta_i$ violating the condition for $[\rho_R(o, \blacktriangle), \blacktriangle]$ to be an explanation for o.

So $\mathcal{C}_m$ is either empty or contains only trivial conditionals, i.e., conditionals whose material counterparts are tautologies. Hence $\tilde{\mathcal{C}}_m = \emptyset$ or it contains only tautologies. Logically, both amounts to $\tilde{\mathcal{C}}_m = \emptyset$. Now, the least fixpoint construction of U_m tells us that there must be a conditional $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ in $\mathcal{N}_m$ that is n-exceptional for $\tilde{\mathcal{C}}_m$. If this was not the case, then $\bigwedge U_m \equiv \top$ which it is not. But if this conditional is n-exceptional for $\emptyset$ it follows that $f(\iota_i \cdot \blacktriangle) \vdash \delta$. Consequently, $f(\rho_R(o, \blacktriangle) \cdot f(\iota_i \cdot \blacktriangle)) \vdash \delta$ violating the condition for $[\rho_R(o, \blacktriangle), \blacktriangle]$ to be an explanation for o. Hence, $[\rho_R(o, \blacktriangle), \blacktriangle]$ does not explain o.

We now show that there is no ρ such that $[\rho, \blacktriangle]$ explains o. Assume that $\mathcal{C}_m$ contains only trivial conditionals, i.e., conditionals whose material counterparts are tautologies. As above, we can then infer that $\mathcal{N}_m$ must contain a conditional $f(\iota_i \cdot \blacktriangle) \rightarrow \delta$ such that $f(\iota_i \cdot \blacktriangle) \vdash \delta$ and hence $f(\rho \cdot f(\iota_i \cdot \blacktriangle)) \vdash \delta$ for *any* sequence ρ. Hence in this case, no explanation can exist.

Assume there is a $\rho = (\beta_l, \ldots, \beta_0)$ such that $[\rho, \blacktriangle]$ explains o. Proposition 2.27 yields that then there is a logical chain with equivalent behaviour, as well. So we can assume that ρ itself is a logical chain. And from the above considerations we can infer that $\mathcal{C}_m$ contains non-trivial conditionals. Let $I = \{i \mid f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i \in \mathcal{C}_m \text{ and } f(\iota_i \cdot \blacktriangle) \not\vdash \theta_i\}$ be the index set of the non-trivial conditionals in $\mathcal{C}_m$. Since $[\rho, \blacktriangle]$ is an explanation for o, it must in particular treat the conditionals $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ such that $i \in I$ correctly. As ρ is a logical chain, we know that for all $i \in I$ there is a smallest s_i such that $f(\rho \cdot f(\iota_i \cdot \blacktriangle)) = \beta_{s_i} \wedge f(\iota_i \cdot \blacktriangle)$ is consistent and $\beta_{s_i} \wedge f(\iota_i \cdot \blacktriangle) \vdash \theta_i$ which implies $\beta_{s_i} \vdash f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i$. Among those s_i there is obviously a smallest s and at least one corresponding conditional.

Wrapping this up, there is a conditional $f(\iota_j \cdot \blacktriangle) \Rightarrow \theta_j \in \mathcal{C}_m$ such that $\beta_s \wedge f(\iota_j \cdot \blacktriangle) \vdash \theta_j$ and $\beta_s \wedge f(\iota_j \cdot \blacktriangle)$ is consistent. As ρ is a logical chain we know $\beta_s \vdash \beta_{s_i}$ for all s_i. This means $\beta_s \vdash f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i$ for all $i \in I$ and hence $\beta_s \vdash \bigwedge \tilde{\mathcal{C}}_m$. We will now show that β_s also entails all the negated antecedents of the conditionals in $\mathcal{N}_m$, i.e., $\beta_s \vdash \neg f(\iota_k \cdot \blacktriangle)$ for all $f(\iota_k \cdot \blacktriangle) \Rightarrow \delta \in \mathcal{N}_m$. But this implies $\beta_s \vdash \bigwedge U_m$ and as $f(\iota_j \cdot \blacktriangle) \Rightarrow \theta_j$ is p-exceptional for U_m, i.e., $U_m \vdash \neg f(\iota_j \cdot \blacktriangle)$ we get a contradiction to $\beta_s \wedge f(\iota_j \cdot \blacktriangle)$ being consistent. Hence, $[\rho, \blacktriangle]$ does not satisfy the conditional $f(\iota_j \cdot \blacktriangle) \Rightarrow \theta_j$ and cannot be an explanation for o.

We still have to show that $\beta_s \vdash \neg f(\iota_k \cdot \blacktriangle)$ for all $f(\iota_k \cdot \blacktriangle) \Rightarrow \delta \in \mathcal{N}_m$. We already know $\beta_s \vdash \bigwedge \tilde{\mathcal{C}}_m$. We also know that there is a conditional $f(\iota_k \cdot \blacktriangle) \Rightarrow \delta \in \mathcal{N}_m$ such that $\bigwedge \tilde{\mathcal{C}}_m \wedge f(\iota_k \cdot \blacktriangle) \vdash \delta$ (otherwise the rational prefix construction would not have stopped). But this means β_s (and therefore β_t, $t < s$) must not be consistent with $f(\iota_k \cdot \blacktriangle)$ as otherwise $[\rho, \blacktriangle]$ could not be an explanation for o. To see this, assume some β_t, $t \leq s$ and t minimal, was consistent with $f(\iota_k \cdot \blacktriangle)$. Then $f(\rho \cdot f(\iota_k \cdot \blacktriangle)) = \beta_t \wedge f(\iota_k \cdot \blacktriangle) \vdash \beta_s \wedge f(\iota_k \cdot \blacktriangle) \vdash \delta$ contradicting the observation o.

So, $\beta_s \vdash \bigwedge \tilde{\mathcal{C}}_m \wedge \neg f(\iota_k \cdot \blacktriangle)$. We can now go on like in the fixpoint construction of U_m, showing that β_s must entail the negated antecedents of any negative conditional in $\mathcal{N}_m$ and therefore $\beta_s \vdash \bigwedge U_m$ as required. □

Proposition 2.51. *Given $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and a core belief $\blacktriangle \equiv \bot$. Then $\rho_R(o, \blacktriangle) = (\top)$.*

Proof. If $\blacktriangle \equiv \bot$ then the antecedent $f(\iota_i \cdot \blacktriangle)$ of any conditional in $\mathcal{C}_\blacktriangle(o)$ or $\mathcal{N}_\blacktriangle(o)$ is inconsistent. This means that the material counterpart of every positive conditional $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ is a tautology and the negated antecedent $\neg f(\iota_i \cdot \blacktriangle)$ of any negative conditional $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ is a tautology, as well. So the conjunction of any subset of these will be a tautology, in particular any $\bigwedge U_j$ will be a tautology.

We now need to show that the calculation stops after the first iteration, i.e., $\mathcal{C}_1 = \mathcal{C}_0$ and $\mathcal{N}_1 = \mathcal{N}_0$. We already showed that U_0 contains only tautologies. $\mathcal{C}_0$ contains all positive

conditionals, $\mathcal{N}_0$ contains all negative ones, so it suffices to show that all conditionals are exceptional for U_0 as then all of them go into the next level, yielding the desired equalities.

So let $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ be any positive conditional. $U_0 \vdash \top$, so $U_0 \vdash \neg f(\iota_i \cdot \blacktriangle)$ and hence $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ is p-exceptional for U_0. Let $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ be any negative conditional. $\bot \vdash \lambda$ for any λ, so $U_0 \cup \{f(\iota_i \cdot \blacktriangle)\} \vdash \delta$ and hence $f(\iota_i \cdot \blacktriangle) \Rightarrow \delta$ is n-exceptional for U_0.

This proves $\mathcal{C}_1 = \mathcal{C}_0$ and $\mathcal{N}_1 = \mathcal{N}_0$ and hence the minimal m such that $\mathcal{C}_{m+1} = \mathcal{C}_m$ and $\mathcal{N}_{m+1} = \mathcal{N}_m$ is 0. Consequently, $\rho_R(o, \blacktriangle) = (\bigwedge U_0) = (\top)$. □

Proposition 2.52. *The decision problem of whether a given core belief ▲ is o-acceptable is Δ_2^P-complete.*

Proof. We will prove this by showing that slightly extending the rational prefix construction we get an algorithm with a polynomial number of NP-oracle calls deciding whether ▲ is o-acceptable and once more reducing the entailment problem for linear base revision to this problem. As input for the decision problem, we are given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and a core belief ▲.

One satisfiability test tells us whether ▲ is a contradiction, in which case it could not be o-acceptable. A polynomial number of SAT-tests suffices to construct the antecedents $f(\iota_i \cdot \blacktriangle)$ of all the positive and negative conditionals yielded by the observation and the core. The size of these formulae is polynomially bounded by the size of the input, as f constructs a conjunction of some of the formulae given. So the sets of conditionals $\mathcal{C}$ and $\mathcal{N}$ passed on to the rational closure construction contain $n + \sum_{j=1}^{n} |D_j|$ conditionals whose size is properly bounded. The constructions of $\mathcal{C}_{i+1}$ and $\mathcal{N}_{i+1}$ are not problematic. For each conditional in $\mathcal{C}_i$ and $\mathcal{N}_i$ one SAT-test decides the exceptionality for U_i and as the size of that set is properly bounded, we are on the safe side. The possible sources of complexity in that construction are the least fixpoint construction of the U_i and the number of iterations.

In each iteration one of the sets $\mathcal{C}_i$ or $\mathcal{N}_i$ has to contain at least one conditional less than $\mathcal{C}_{i-1}$ or $\mathcal{N}_{i-1}$, respectively, otherwise the process stops. As there is only a polynomial number of conditionals there can only be a polynomial number of iterations. The construction of U_i starts with the unproblematic initialisation of $\tilde{\mathcal{C}}_i$. In each iteration of the least fixpoint construction the negated antecedent of one negative conditional may be added. The corresponding conditional need not be checked again. So there are at most $|\mathcal{N}_i|$ iterations before a fixpoint is found and in each iteration at most $|\mathcal{N}|$ conditionals have to be checked for n-exceptionality which corresponds to that number of SAT-tests. The size of U_i keeps a proper bound as it grows only linearly in each iteration.

Hence, in each iteration of the rational closure construction a polynomial number of SAT-tests suffices. To check o-acceptability of $\blacktriangle$ we have to test whether $\bigwedge U_m \equiv \top$ (Propositions 2.49 and 2.50), but this corresponds to testing whether both $\mathcal{C}_m$ and $\mathcal{N}_m$ are empty. If one of them was not, the weakest formula of the rational prefix could not be a tautology (unless $\blacktriangle \equiv \bot$ which we tested before even starting). This concludes the proof of the decision problem belonging to Δ_2^P.

Again, the Δ_2^P-hardness part of the proof that $(\alpha_1, \ldots, \alpha_m) *_L \alpha \vdash \beta$ is Δ_2^P-complete [70, 71] uses a reduction that guarantees that α is consistent. We extend this reduction to the o-acceptability of a core belief. We claim that $(\alpha_1, \ldots, \alpha_m) *_L \alpha \vdash \beta$ if and only if $\top$ is o-acceptable for $o = \langle(\neg\beta, \top, \emptyset), (\alpha_1, \top, \emptyset), \ldots, (\alpha_m, \top, \emptyset), (\alpha, \beta, \emptyset)\rangle$. Obviously, this transformation is polynomial.

Assume $(\alpha_1, \ldots, \alpha_m) *_L \alpha \vdash \beta$ which is equivalent to $f(\alpha_1, \ldots, \alpha_m, \alpha) \vdash \beta$ and thus also to $f(\alpha_1, \ldots, \alpha_m, \alpha, \top) \vdash \beta$ (Proposition 2.7 and α is consistent). Consider the epistemic state $[(), \top]$. All D_i are empty and the only relevant θ_i is $\theta_n = \beta$; the others are trivially dealt with as any formula entails $\top$. $f(\neg\beta, \alpha_1, \ldots, \alpha_m, \alpha, \top)$ is equivalent to $f(\neg\beta, f(\alpha_1, \ldots, \alpha_m, \alpha, \top))$ (Proposition 2.3), but as $f(\alpha_1, \ldots, \alpha_m, \alpha, \top) \vdash \beta$ this is equivalent to $f(\alpha_1, \ldots, \alpha_m, \alpha, \top)$ which entails β as necessary. So $[(), \top]$ explains o and $\top$ is o-acceptable.

Now assume $(\alpha_1, \ldots, \alpha_m) *_L \alpha \nvdash \beta$ which is equivalent to $f(\alpha_1, \ldots, \alpha_m, \alpha, \top) \nvdash \beta$ as above. In particular, this means $\beta \not\equiv \top$ and $\neg\beta \not\equiv \bot$. Consider *any* epistemic state $[\rho, \top]$. As above only $\theta_n = \beta$ is relevant. $f(\rho \cdot (\neg\beta, \alpha_1, \ldots, \alpha_m, \alpha, \top)) \equiv f(\rho \cdot f(\neg\beta, f(\alpha_1, \ldots, \alpha_m, \alpha, \top)))$. $f(\alpha_1, \ldots, \alpha_m, \alpha, \top)$ does not entail β and is hence consistent with $\neg\beta$. As a consequence, $f(\neg\beta, f(\alpha_1, \ldots, \alpha_m, \alpha, \top)) = \neg\beta \wedge f(\alpha_1, \ldots, \alpha_m, \alpha, \top)$ and thus $f(\rho \cdot (\neg\beta, \alpha_1, \ldots, \alpha_m, \alpha, \top))$ entails $\neg\beta$ and cannot consistently entail β as required. So, $[\rho, \top]$ does not explain o no matter which sequence ρ is used and $\top$ is not o-acceptable — proving our claim. □

Proposition 2.55. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation and $\blacktriangle$ be an o-acceptable core.*

$$\textit{If } [\sigma, \blacktriangle] \textit{ explains } o \textit{ then } \rho_R(o, \blacktriangle) \preceq_1 \sigma.$$

Proof. As $\blacktriangle$ is an o-acceptable core we know that $[\rho_R(o, \blacktriangle), \blacktriangle]$ is indeed an explanation for o (Propositions 2.49 and 2.50). By Proposition 2.27 we can restrict our attention to logical chains. To ease notation, let us denote $\rho_R(o, \blacktriangle)$ in this proof by just $\rho_R = (\alpha_m, \ldots, \alpha_0)$, $\sigma = (\beta_l, \ldots, \beta_0)$ such that $[\sigma, \blacktriangle]$ explains o. Both sequences are logical chains. Again, we abbreviate $(\varphi_1, \ldots, \varphi_i)$ by ι_i.

Also, let us introduce the following notation: Given any formula λ,

$$rank_{\rho_R}(\lambda) = \begin{cases} \min\{k \mid \lambda \wedge \alpha_k \text{ is consistent}\} \\ \infty \qquad\qquad \text{if no such } k \text{ exists} \end{cases}$$

Note that since $\alpha_m \equiv \top$ the second case applies only if $\lambda \equiv \bot$). Analogously,

$$rank_{\sigma}(\lambda) = \begin{cases} \min\{k \mid \lambda \wedge \beta_k \text{ is consistent}\} \\ \infty \qquad\qquad \text{if no such } k \text{ exists} \end{cases}$$

Recall that as ρ_R is a logical chain we have for any λ such that $\lambda \not\equiv \bot$, $f(\rho_R \cdot \lambda)) = \lambda \wedge \alpha_s$ where $s = rank_{\rho_R}(\lambda)$.

To show $\rho_R \preceq_1 \sigma$ we must prove that for any $i \in \{0, \ldots, n\}$, if $Bel_j^{\rho_R} \equiv Bel_j^{\sigma}$ for all $j < i$ then $Bel_i^{\sigma} \vdash Bel_i^{\rho_R}$. So fix $i \in \{0, \ldots, n\}$ and assume $Bel_j^{\rho_R} \equiv Bel_j^{\sigma}$ for all $j < i$. We have $Bel_i^{\rho_R} = f(\rho_R \cdot \iota_i \cdot \blacktriangle)$. By Proposition 2.3 this is the same as $f(\rho_R \cdot f(\iota_i \cdot \blacktriangle))$. As $\blacktriangle$ is consistent and consequently $f(\iota_i \cdot \blacktriangle)$ is consistent, we have $f(\rho_R \cdot f(\iota_i \cdot \blacktriangle)) = f(\iota_i \cdot \blacktriangle) \wedge \alpha_s$, where $s = rank_{\rho_R}(f(\iota_i \cdot \blacktriangle))$. What does α_s look like? By the construction of ρ_R, we know

$$\begin{aligned} &\alpha_s \equiv \textstyle\bigwedge_{k \in I} \big(f(\iota_k \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \bigwedge_{k \in J} \big(\neg f(\iota_k \cdot \blacktriangle)\big), \text{ where} \\ &I = \{k \mid 1 \leq k \leq n \text{ and } rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) \geq s\} \text{ and} \\ &J = \{k \in I \mid \bigwedge_{i \in I} \big(f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i\big) \wedge \bigwedge_{k' \prec_e k} \big(\neg f(\iota_{k'} \cdot \blacktriangle)\big) \big) \wedge f(\iota_k \cdot \blacktriangle) \vdash \delta \\ &\qquad\qquad \text{for some negative conditional } f(\iota_k \cdot \blacktriangle) \Rightarrow \delta\}. \end{aligned}$$

$\prec_e$ is a total order on the indexes, indicating in which order the corresponding negative conditionals become exceptional in the least fixpoint calculation of the rational prefix construction. In other words I is the index set of the positive conditionals that are p-exceptional for U_{s-1} (some of them may still be for U_s but not necessarily all) and J is the index set of the negative conditionals that are n-exceptional for U_s. Thus we have obtained

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

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

(a) $Bel_i^{\sigma} \vdash f(\iota_i \cdot \blacktriangle)$

(b) $Bel_i^{\sigma} \vdash \big(f(\iota_k \cdot \blacktriangle) \rightarrow \theta_k\big)$ for all $k \in I$, equivalently
$Bel_i^{\sigma} \wedge f(\iota_k \cdot \blacktriangle) \vdash \theta_k$, for all $k \in I$

(c) $Bel_i^{\sigma} \vdash \neg f(\iota_k \cdot \blacktriangle)$, for all $k \in J$.

(a): $Bel_i^{\sigma} = f(\sigma \cdot \iota_i \cdot \blacktriangle) = f(\sigma \cdot f(\iota_i \cdot \blacktriangle)) \vdash f(\iota_i \cdot \blacktriangle)$

(b): Let $k \in I$. $Bel_i^{\sigma} \wedge f(\iota_k \cdot \blacktriangle) \equiv \beta_t \wedge f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle)$, $t = rank_{\sigma}(f(\iota_i \cdot \blacktriangle))$. If this is inconsistent, we trivially get the desired conclusion. So suppose $Bel_i^{\sigma} \wedge f(\iota_k \cdot \blacktriangle)$ is

consistent, which implies that $f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle)$ is consistent. We have to consider two cases $i \leq k$ and $i > k$.

$i \leq k$ Proposition 2.15 tells us that $f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$ and as a consequence $\beta_t \wedge f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle) \equiv \beta_t \wedge f(\iota_k \cdot \blacktriangle)$. We claim that $rank_\sigma(f(\iota_k \cdot \blacktriangle)) = t$ which implies that $\beta_t \wedge f(\iota_k \cdot \blacktriangle) \equiv Bel_k^\sigma \vdash \theta_k$ — yielding the desired result.

To prove our claim assume $rank_\sigma(f(\iota_k \cdot \blacktriangle)) = u < t$. Then $\beta_u \wedge f(\iota_k \cdot \blacktriangle) \nvdash \bot$, as $f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$ we know $\beta_u \wedge f(\iota_i \cdot \blacktriangle) \nvdash \bot$ contradicting that $f(\iota_i \cdot \blacktriangle)$ is inconsistent with all β_u where $u < t$ (definition of *rank*). $u > t$ is not possible as $\beta_t \wedge f(\iota_k \cdot \blacktriangle)$ is consistent and *rank* yields the smallest possible index. This proves our claim.

$i > k$ Proposition 2.15 tells us that $f(\iota_i \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$ and by Proposition 2.12 we know $Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) = f(\sigma \cdot f(\iota_i \cdot \blacktriangle)) \wedge f(\iota_k \cdot \blacktriangle) \equiv f(\sigma \cdot (f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle)))$. If we could show $f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle)$ is consistent then Proposition 2.10 would yield

$$\begin{aligned} Bel_i^\sigma \wedge f(\iota_k \cdot \blacktriangle) &\equiv f(\sigma \cdot (f(\iota_i \cdot \blacktriangle) \wedge f(\iota_k \cdot \blacktriangle))) \\ &\equiv f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle) \text{ by Proposition 2.12} \\ &\vdash f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) = Bel_k^\sigma, \end{aligned}$$

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

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

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

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

Now since $k \in I$ we know $t \geq s$, hence $\alpha_s \vdash \alpha_t$. We know already that $\alpha_s \wedge f(\iota_i \cdot \blacktriangle)$ is consistent. Thus it follows that $\alpha_t \wedge f(\iota_i \cdot \blacktriangle)$ is consistent and so $f(\sigma \cdot f(\iota_k \cdot \blacktriangle)) \wedge f(\iota_i \cdot \blacktriangle)$ is consistent as required.

(c): We know by construction of the rational prefix that we can order the elements of J using a total order $\prec_e$. For any $k \in J$ there is a negative conditional $f(\iota_k \cdot \blacktriangle) \Rightarrow \delta$ such that

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

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

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

Recall, $Bel_i^{\rho_R} = f(\iota_i \cdot \blacktriangle) \wedge \alpha_s$. As $Bel_i^{\rho_R}$ is consistent we know that $Bel_i^{\rho_R} \nvdash \neg f(\iota_i \cdot \blacktriangle)$. This implies $i \notin J$ (cf. the structure of α_s). Hence, we only need to consider the cases $i > k$ and $i < k$.

$i > k$ $rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) = u > s$ as $k \in J$ (since $\alpha_s \vdash \neg f(\iota_k \cdot \blacktriangle)$ for $k \in J$). Proposition 2.15 again tells us $f(\iota_i \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$ implying $rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) = u \leq s$. This is because $rank_{\rho_R}(f(\iota_i \cdot \blacktriangle)) = s$ and hence α_s must be consistent with $f(\iota_k \cdot \blacktriangle)$. So this case is impossible, as well.

$i < k$ Proposition 2.15 tells us that $f(\iota_k \cdot \blacktriangle) \vdash f(\iota_i \cdot \blacktriangle)$. $Bel_i^\sigma = \beta_t \wedge f(\iota_i \cdot \blacktriangle)$ with $t = rank_\sigma(f(\iota_i \cdot \blacktriangle))$.

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

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

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

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

So $Bel_k^\sigma \vdash \delta$. Hence $[\sigma, \blacktriangle]$ cannot be an explanation — contradiction — and consequently $Bel_i^\sigma \vdash \neg f(\iota_k \cdot \blacktriangle)$.

□

Proposition 2.56. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation and $\blacktriangle$ be an o-acceptable core.*

$$\textit{If } [\sigma, \blacktriangle] \textit{ explains } o \textit{ and } \sigma \preceq_1 \rho_R(o, \blacktriangle) \textit{ then } \rho_R(o, \blacktriangle) \preceq_2 \sigma.$$

Proof. The proof is almost identical to the last part of Proposition 2.55. Again it suffices to restrict the argument to logical chains. We use $\rho_R = (\alpha_m, \ldots, \alpha_0)$ to denote $\rho_R(o, \blacktriangle)$. Let $[\sigma, \blacktriangle]$ be an explanation of o and suppose $\sigma \preceq_1 \rho_R$. We already know by Proposition 2.55 that

also $\rho_R \preceq_1 \sigma$. Taking these two inequalities together means we must have $Bel_i^{\sigma} \equiv Bel_i^{\rho_R}$ for *all* $i = 0, \ldots, n$. Now to show $\rho_R \preceq_2 \sigma$ choose any formula λ. We have to show $Bel([\sigma, \blacktriangle]{*}\varphi_1{*}\cdots{*}\varphi_n{*}\lambda) \vdash Bel([\rho_R, \blacktriangle]{*}\varphi_1{*}\cdots{*}\varphi_n{*}\lambda)$, i.e., $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle)$.

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

$$\begin{aligned}
&\alpha_s \equiv \textstyle\bigwedge_{k \in I} \big(f(\iota_k \cdot \blacktriangle) \rightarrow \theta_k\big) \wedge \bigwedge_{k \in J} \big(\neg f(\iota_k \cdot \blacktriangle)\big), \text{ where} \\
&I = \{k \mid 1 \leq k \leq n \text{ and } rank_{\rho_R}(f(\iota_k \cdot \blacktriangle)) \geq s\} \text{ and} \\
&J = \{k \in I \mid \textstyle\bigwedge\limits_{i \in I} \big(f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i\big) \wedge \bigwedge\limits_{k' \prec_e k} \big(\neg f(\iota_{k'} \cdot \blacktriangle)\big) \wedge f(\iota_k \cdot \blacktriangle) \vdash \delta \\
&\qquad\qquad \text{for some negative conditional } f(\iota_k \cdot \blacktriangle) \Rightarrow \delta\}.
\end{aligned}$$

So, again we have to prove

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

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

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

(a) and (b) are exactly as in the proof for Proposition 2.55. Note that for (b) only the case corresponding to $i > k$ is possible. In order to show (c) take an arbitrary $k \in J$.

If $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash \neg f(\iota_k \cdot \blacktriangle)$ we are done. So assume $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \nvdash \neg f(\iota_k \cdot \blacktriangle)$. Hence $f(\iota \cdot \lambda \cdot \blacktriangle)$ is consistent with $f(\iota_k \cdot \blacktriangle)$. Proposition 2.15 tells us that $f(\iota \cdot \lambda \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$. Consequently $f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash f(\iota_k \cdot \blacktriangle)$, but we already know $f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle) \vdash \neg f(\iota_k \cdot \blacktriangle)$ as $k \in J$. So we get a contradiction as $f(\rho_R \cdot \iota \cdot \lambda \cdot \blacktriangle)$ must be consistent ($\blacktriangle$ is). So it is impossible that $f(\sigma \cdot \iota \cdot \lambda \cdot \blacktriangle) \nvdash \neg f(\iota_k \cdot \blacktriangle)$. This concludes the proof. □

Proposition 2.58. *Given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and a core belief $\blacktriangle$. Ultimately exceptional conditionals in the rational prefix construction of o with respect to $\blacktriangle$ exist if and only if there are subsets $I_{\mathcal{C}} \neq \emptyset$ and $I_{\mathcal{N}}$ of $\{1, \ldots, n\}$ and a total order $\prec_e$ on $I_{\mathcal{N}}$ such that*

1. $\textstyle\bigwedge\limits_{i \in I_{\mathcal{C}}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \wedge \bigwedge\limits_{j \in I_{\mathcal{N}}} \neg f(\iota_j \cdot \blacktriangle) \vdash \bigwedge\limits_{i \in I_{\mathcal{C}}} \neg f(\iota_i \cdot \blacktriangle)$
2. $\forall j \in I_{\mathcal{N}} \exists \delta \in D_j : \textstyle\bigwedge\limits_{i \in I_{\mathcal{C}}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \wedge \bigwedge\limits_{k \in I_{\mathcal{N}} \wedge k \prec_e j} \neg f(\iota_k \cdot \blacktriangle) \wedge f(\iota_j \cdot \blacktriangle) \vdash \delta$

Proof. (i) If ultimately exceptional conditionals exist then these conditions are satisfied. Let $I_{\mathcal{C}}$ be the index set of the positive conditionals that are ultimately exceptional and $I_{\mathcal{N}}$ that of the negative ones. Note that $\textstyle\bigwedge\limits_{i \in I_{\mathcal{C}}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \wedge \bigwedge\limits_{j \in I_{\mathcal{N}}} \neg f(\iota_j \cdot \blacktriangle)$ is nothing but $\bigwedge U_m$,

as U_m contains the material counterparts of the p-exceptional conditionals and the negated antecedents of the n-exceptional ones.

The first condition simply expresses that any positive conditional with index $i \in I_{\mathcal{C}}$ will again be exceptional in the next step, which is true as they are all ultimately exceptional. The second condition is another way of looking at the least fixpoint construction of U_m. Let $\prec_e$ be the total order on the indexes expressing in which order the negative conditionals become n-exceptional. $\bigwedge_{i \in I_{\mathcal{C}}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i)$ is nothing but $\bigwedge \tilde{\mathcal{C}}_m$, i.e., it represents the set of formulae U_m is initialised with. Adding the antecedent of one the $\prec_e$-minimal negative conditionals must have made the corresponding consequent δ inferable, as that conditional was n-exceptional for $\tilde{\mathcal{C}}_m$. Consequently the negated antecedent was added to U_m. $k \prec_e j$ expresses that adding to $\tilde{\mathcal{C}}_m$ the negated antecedents of the n-exceptional conditionals less (with respect to $\prec e$) sufficed to make a conditional with index j n-exceptional as well. If the second condition did not hold, one of the conditionals in $\mathcal{N}_m$ could not be n-exceptional for U_m and hence not ultimately exceptional.

(ii) Let there be $I_{\mathcal{C}} \neq \emptyset$ and $I_{\mathcal{N}}$ of $\{1, \ldots, n\}$ and a total order $\prec_e$ on $I_{\mathcal{N}}$ such that the conditions are satisfied. We will now show by induction that if the conditionals corresponding to the index sets $I_{\mathcal{C}}$ and $I_{\mathcal{N}}$ belong to $\mathcal{C}_i$ and $\mathcal{N}_i$ then they are propagated to the $\mathcal{C}_{i+1}$ and $\mathcal{N}_{i+1}$. If this is true then we are done, as $\mathcal{C}_0$ and $\mathcal{N}_0$ contain all conditionals, in particular those corresponding to the given index sets.

So assume that all conditionals corresponding to the index sets $I_{\mathcal{C}}$ and $I_{\mathcal{N}}$ belong to $\mathcal{C}_i$ and $\mathcal{N}_i$. For the $\prec_e$-minimal element j of $I_{\mathcal{N}}$ the second condition translates to $\exists \delta \in D_j : \bigwedge_{i \in I_{\mathcal{C}}} (f(\iota_i \cdot \blacktriangle) \rightarrow \theta_i) \wedge f(\iota_j \cdot \blacktriangle) \vdash \delta$, or equivalently $\tilde{\mathcal{C}} \cup f(\iota_j \cdot \blacktriangle) \vdash \delta$. So in particular $\tilde{\mathcal{C}}_i \cup f(\iota_j \cdot \blacktriangle) \vdash \delta$ and hence $\neg f(\iota_j \cdot \blacktriangle)$ has to be added to U_i as otherwise the second condition of the rational closure construction is violated (c.f. Definition 2.45). The argument continues analogously for all the $j \in I_{\mathcal{N}}$ in increasing order w.r.t. $\prec_e$. That is, the negated antecedents of all conditionals with index in $I_{\mathcal{N}}$ are added to U_i and thus the corresponding negative conditionals will be n-exceptional for U_i and are propagated to $\mathcal{N}_{i+1}$.

As the first condition holds we have $U_i \vdash \bigwedge_{i \in I_{\mathcal{C}}} \neg f(\iota_i \cdot \blacktriangle)$ (all the positive conditionals corresponding to $I_{\mathcal{C}}$ belong to $\mathcal{C}_i$ and the material counterparts are hence entailed by $\tilde{\mathcal{C}}_i$) and as just seen all the negated antecedents of conditionals corresponding to $I_{\mathcal{N}}$ have been added as well. But this means that all positive conditionals corresponding to $\mathcal{C}$ are p-exceptional for U_i and are hence propagated to $\mathcal{C}_{i+1}$. This proves the claim. □

Proposition 2.59. *Let* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ *be an observation and* $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$ *the rational prefix for* o *using some core* $\blacktriangle$.

Then $\alpha_m \equiv \top$ and $\mathcal{C}_m \neq \emptyset$ implies $\blacktriangle \equiv \bot$.

Proof. We showed that $\alpha_m = \bigwedge U_m \equiv \bigwedge_{i \in I} \neg f(\iota_i \cdot \blacktriangle)$ where I is the index set of the ultimately exceptional positive conditionals. So if $\mathcal{C}_m \neq \emptyset$, then $\neg f(\iota_i \cdot \blacktriangle)$ must be a tautology for all $i \in I$ and $f(\iota_i \cdot \blacktriangle) \equiv \bot$. But this can be the case only if $\blacktriangle \equiv \bot$. □

Proposition 2.60. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$, $\blacktriangle$ a core belief and $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$.*

If $\blacktriangle' \vdash \blacktriangle$ and $\blacktriangle' \nvdash \alpha_m$ then $\blacktriangle'$ cannot be o-acceptable.

Proof. We need not consider the uninteresting case where $\blacktriangle$ is o-acceptable or inconsistent, implying $\alpha_m = \top$. So let $\blacktriangle$ be consistent and not o-acceptable. Then $\alpha_m = \bigwedge U_m \not\equiv \top$ (contraposition of Proposition 2.49) which implies that there were ultimately exceptional conditionals ($\mathcal{C}_m \neq \emptyset$ or $\mathcal{N}_m \neq \emptyset$). Let $I_\mathcal{C}$ and $I_\mathcal{N}$ be the index sets of the ultimately exceptional positive and negative conditionals and let $\prec_e$ be the order on $I_\mathcal{N}$ in which the corresponding conditionals became exceptional in the least fixpoint construction. So we know

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

We will need these entailments later in the proof and will refer back to them with (*). We will now show that if $\blacktriangle' \nvdash \bigwedge_{i \in I_\mathcal{C}} \neg f(\iota_i \cdot \blacktriangle)$ (implying $\blacktriangle'$ to be consistent) there will again be ultimately exceptional conditionals and as $\blacktriangle' \not\equiv \bot$ the formula $\bigwedge U'_{m'}$ cannot be a tautology. Hence $\blacktriangle'$ is not o-acceptable (Proposition 2.50).

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

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

This means a conditional with an index $j \in J$ will again be ultimately exceptional when using $\blacktriangle'$ as the core belief. We know $\blacktriangle' \vdash \blacktriangle$, so $\blacktriangle' \wedge \blacktriangle \equiv \blacktriangle'$ and for $j \in J$ we have $f(\iota_j \cdot \blacktriangle') \nvdash \neg\blacktriangle'$, so using Proposition 2.12 we get $f(\iota_j \cdot \blacktriangle') \equiv f(\iota_j \cdot \blacktriangle' \wedge \blacktriangle) \equiv f(\iota_j \cdot \blacktriangle) \wedge \blacktriangle'$. The corresponding equivalence does not hold for $f(\iota_j \cdot \blacktriangle')$, $j \notin J$. In this case we know that $f(\iota_j \cdot \blacktriangle) \wedge \blacktriangle' \vdash \bot$ so that $\bigwedge_{j \notin J} (f(\iota_j \cdot \blacktriangle) \wedge \blacktriangle' \rightarrow \theta_j)$ is a tautology. We start by proving $\bigwedge_{i \in J} (f(\iota_i \cdot \blacktriangle') \rightarrow \theta_i) \ \wedge \bigwedge_{j \in J \cap I_\mathcal{N}} \neg f(\iota_j \cdot \blacktriangle') \vdash \bigwedge_{i \in J} \neg f(\iota_i \cdot \blacktriangle')$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proposition 2.61. *For all observations o Algorithm 1 terminates.*

Proof. The termination condition for the repeat loop is that α_m from the rational prefix construction $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$ must be a tautology. Proposition 2.51 yields that this will definitely be the case if $\blacktriangle$ is inconsistent. The contraposition of Proposition 2.50 yields that if $\blacktriangle$ is o-acceptable, then $\alpha_m \equiv \top$, so in case during the iteration an o-acceptable core has been constructed, the algorithm will terminate, as well.

So let $\blacktriangle \not\equiv \bot$ and $\alpha_m \not\equiv \top$, i.e., there will be another iteration. If we show that $\blacktriangle \nvdash \alpha_m$ and consequently $\blacktriangle \wedge \alpha_m$ is strictly logically stronger, then this guarantees termination. This is because the calculation of the conditional beliefs and the rational prefix construction do not invent new propositional variables and as the language of o is finite, constructing strictly stronger cores will either yield an acceptable core or $\blacktriangle = \bot$ after finitely many iterations.

Assume $\blacktriangle \vdash \alpha_m$, but $\alpha_m = \bigwedge U_m$. However, we argued that $\bigwedge U_m \equiv \bigwedge_{i \in I} \neg f(\iota_i \cdot \blacktriangle)$ where I is the index set of the ultimately exceptional positive conditionals which are collected in $\mathcal{C}_m$. Let $f(\iota_i \cdot \blacktriangle) \Rightarrow \theta_i$ be one of these conditionals. So $\blacktriangle \vdash \neg f(\iota_i \cdot \blacktriangle)$ and as $f(\iota_i \cdot \blacktriangle) \vdash \blacktriangle$ we have $f(\iota_i \cdot \blacktriangle) \vdash \neg f(\iota_i \cdot \blacktriangle)$. Hence $f(\iota_i \cdot \blacktriangle)$ is inconsistent, but this is possible only if $\blacktriangle$ is inconsistent — contradiction, so $\blacktriangle \nvdash \alpha_m$ □

Proposition 2.62. *Given as input an observation o, Algorithm 1 outputs the rational explanation $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ for o, if an explanation for o exists. If no explanation exists it outputs "no explanation".*

Proof. If o does not have an explanation, the algorithm will terminate with $\blacktriangle \equiv \bot$ and hence output "no explanation". Assume $\blacktriangle$ was consistent, then α_m from $\rho_R(o, \blacktriangle) = (\alpha_m, \ldots, \alpha_0)$ would have to be a tautology, but then we could apply Proposition 2.49 saying that $\blacktriangle$ is o-acceptable, contradicting that o has no explanation.

So assume o does have an explanation. We need to show two things. Firstly, if the algorithm terminates with a consistent core $\blacktriangle$ then its output $[\rho, \blacktriangle]$ indeed matches the rational explanation. Secondly, the algorithm indeed does terminate with a consistent core $\blacktriangle$.

So assume the algorithm terminates with a consistent core $\blacktriangle$. It suffices to show that $\blacktriangle \equiv \blacktriangle_\vee(o)$ as then $\rho_R(o, \blacktriangle)$ is obviously equivalent to $\rho_R(o, \blacktriangle_\vee(o))$. $\blacktriangle \vdash \blacktriangle_\vee(o)$ is trivial as $\blacktriangle$ is o-acceptable (Proposition 2.49) and any o-acceptable core entails $\blacktriangle_\vee(o)$. So we need to show $\blacktriangle_\vee(o) \vdash \blacktriangle$. As the algorithm terminated, it must have done so after $k+1$ iterations of calculating the rational prefix for the current core belief and calculating a new core belief (in the last step the core does not change, as the old one is conjoined with a tautology). So consider the sequence $[\rho_0, \blacktriangle_0], \ldots, [\rho_k, \blacktriangle_k]$. We need to show $\blacktriangle_\vee(o) \vdash \blacktriangle_k$ as that is the core in the output of the algorithm.

Note $\blacktriangle_\vee(o) \vdash \top$ so $\blacktriangle_\vee(o) \vdash \blacktriangle_0$ as that is the core the algorithm starts off with. For the inductive step let $\blacktriangle_\vee(o) \vdash \blacktriangle_i$ and $\rho_i = \rho_R(o, \blacktriangle_i) = (\alpha_m, \ldots, \alpha_0)$. Now, Proposition 2.60 tells

us that if $\blacktriangle_\vee(o)$ does not entail α_m then $\blacktriangle_\vee(o)$ is not o-acceptable but $\blacktriangle_\vee(o)$ actually is o-acceptable. Hence $\blacktriangle_\vee(o)$ must entail α_m. Now, we have $\blacktriangle_\vee(o) \vdash \blacktriangle_i$ and $\blacktriangle_\vee(o) \vdash \alpha_m$ and hence $\blacktriangle_\vee(o) \vdash \blacktriangle_{i+1}$ as $\blacktriangle_{i+1} = \blacktriangle_i \wedge \alpha_m$.

This yields that $\blacktriangle_\vee(o)$ indeed entails $\blacktriangle_k$ as needed. We still need to show that the algorithm indeed terminates with a consistent core. Assume it does not. As it always terminates (Proposition 2.61), it will do so with an inconsistent core after $k+1$ iterations. So $\blacktriangle_k \equiv \bot$. We can now use the above argumentation that $\blacktriangle_\vee(o) \vdash \blacktriangle_i$ for all i, so in particular $\blacktriangle_\vee(o) \vdash \blacktriangle_k$, contradicting that $\blacktriangle_\vee(o)$ is an o-acceptable core. Hence, the algorithm terminates with a consistent core $\blacktriangle$. □

A.2 Proofs from Chapter 3

Proposition 3.4. *Let L be a propositional language and $\phi \in L$. Let I be a set of natural numbers, for each $i \in I$ let $\lambda_i(\chi)$ be a parametrised formula based on L, $\alpha_i = \lambda_i(\chi)[\chi/\phi]$, and $\alpha'_i = \lambda_i(\chi)[\chi/x]$ where $x \notin L$, i.e., the propositional variable x is not contained in any $\lambda_i(\chi)$ or ϕ. Then for all finite $S \subseteq I$*

$$\bigwedge_{i\in S} \alpha_i \vdash \bot \text{ if and only if } (x \leftrightarrow \phi) \wedge \bigwedge_{i\in S} \alpha'_i \vdash \bot.$$

Proof. If $\bigwedge_{i\in S} \alpha_i \nvdash \bot$ then $\exists m : m \models \bigwedge_{i\in S} \alpha_i$. Note that x does not appear in $\bigwedge_{i\in S} \alpha_i$. Let $m' \sim_x m$ with $m' \models x$ if and only if $m \models \phi$. Hence $m' \models (x \leftrightarrow \phi)$ and $m' \models \bigwedge_{i\in S} \alpha'_i$.

If $(x \leftrightarrow \phi) \wedge \bigwedge_{i\in S} \alpha'_i \nvdash \bot$ then $\exists m : m \models (x \leftrightarrow \phi) \wedge \bigwedge_{i\in S} \alpha'_i$. $m \models x$ if and only if $m \models \phi$. Hence $m \models \alpha'_i$ if and only if $m \models \alpha_i$. As a consequence $m \models \bigwedge_{i\in S} \alpha_i$. □

Proposition 3.5. *If $[\rho, \blacktriangle]$ explains $o[(\chi_i/\phi_i)_i]$ and x is a propositional variable not appearing in o, $\blacktriangle$, ρ or any ϕ_i then $[\rho, \blacktriangle \wedge (x \leftrightarrow \phi_i)]$ explains $o[\chi_i/x, (\chi_j/\phi_j)_{j\neq i}]$.*

Proof. Let $\iota_k(o)$ denote the sequence of the first k revision inputs from o. We will show that for any parametrised formula λ possibly containing the unknown subformulae and not containing the additional variable x

$$\begin{array}{lll} f(\rho \cdot \iota_k(o[(\chi_i/\phi_i)_i]) \cdot \blacktriangle) & \vdash \lambda[(\chi_i/\phi_i)_i] & \text{if and only if} \\ f(\rho \cdot \iota_k(o[\chi_i/x, (\chi_j/\phi_j)_{j\neq i}]) \cdot \blacktriangle \wedge (x \leftrightarrow \phi_i)) & \vdash \lambda[\chi_i/x, (\chi_j/\phi_j)_{j\neq i}]. & \end{array}$$

This immediately yields the desired result as this means the two correspond on any parametrised formula that can appear in the beliefs and non-beliefs recorded in the observation.

First we will show that the two collect the same elements from ρ and from the corresponding inputs in $\iota_k(o[(\chi_i/\phi_i)_i])$ and $\iota_k(o[\chi_i/x,(\chi_j/\phi_j)_{j\neq i}])$. This is a simple induction over the length of the sequence using Proposition 3.4. Note that any two corresponding formulae from the two sequences behave like α_i and α'_i in that proposition — the two are equivalent except for a different instantiation of χ_i.

The induction starts with proving $\blacktriangle$ is accepted iff $\blacktriangle \wedge (x \leftrightarrow \phi_i)$ is accepted. This is the case, as $\blacktriangle$ does not contain x and we can apply Proposition 3.4 by taking $\alpha = \blacktriangle = \alpha'$ ($\blacktriangle$ is independent of χ_i). That way, if the core is consistent, the $(x \leftrightarrow \phi_i)$ necessary for the proposition is introduced. Now assume the two have selected the corresponding elements up to some point, then by Proposition 3.4 the one will accept α_i if and only if the other accepts α'_i. Hence, the same elements are collected.

To see that both entail corresponding λs, we simply assume a further $\alpha_j = \neg\lambda[(\chi_i/\phi_i)_i]$ and $\alpha'_j = \neg\lambda[\chi_i/x,(\chi_j/\phi_j)_{j\neq i}]$ in Proposition 3.4 ($\psi \vdash \lambda$ iff $\psi \wedge \neg\lambda \vdash \bot$). □

Proposition 3.6. *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* $\left[\rho,\blacktriangle \wedge \bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i)\right]$ *explains* $o[(\chi_i/x_i)_i]$.

Proof. Apply proposition 3.5 n times each time replacing one χ_i by x_i. □

Proposition 3.8. *Let* $\lambda(\chi_1,\ldots,\chi_n)$ *be a parametrised formula based on* L, ψ *a formula and* $x_1,\ldots,x_n \notin L$ *be* n *propositional variables not appearing in* λ.

$$\textit{If} \quad \lambda[(\chi_i/x_i)_i] \equiv \alpha = \varphi \wedge \bigwedge_{1\leq j\leq l}(\theta_j \vee \bigvee_{p\in P_j} x_p \vee \bigvee_{q\in N_j} \neg x_q)$$

for a natural number l, *appropriate* $\varphi,\theta_j \in L$ *and subsets* P_j,N_j *of* $\{1,\ldots,n\}$, $1 \leq j \leq l$,

$$\textit{then } \lambda[(\chi_i/x_i\wedge\psi)_i] \equiv \alpha[(x_i/x_i\wedge\psi)_i] = \varphi \wedge \bigwedge_{1\leq j\leq l}(\theta_j \vee \bigvee_{p\in P_j}(x_p\wedge\psi) \vee \bigvee_{q\in N_j}\neg(x_q\wedge\psi)).$$

Proof. Note that α represents the CNF of the formula, where the clauses containing a variable x_i are explicitly represented. The proposition basically follows from the fact that any equivalence transformation on $\lambda[(\chi_i/x_i)_i]$, so in particular those that lead to the CNF, can also be carried out on $\lambda[(\chi_i/x_i\wedge\psi)_i]$.

Let $\lambda[(\chi_i/x_i)_i] \equiv \alpha$ and let α' denote $\alpha[(x_i/x_i\wedge\psi)_i]$. So assume $\lambda[(\chi_i/x_i\wedge\psi)_i] \not\equiv \alpha'$, i.e., $\exists m$ such that (i) $m \models \lambda[(\chi_i/x_i\wedge\psi)_i]$ and $m \not\models \alpha'$ or (ii) $m \not\models \lambda[(\chi_i/x_i\wedge\psi)_i]$ and $m \models \alpha'$.

(i) Assume $m \models \psi$, then $m \models x_i\wedge\psi$ iff $m \models x_i$, for all i. Hence $m \models \lambda[(\chi_i/x_i)_i]$, but $m \not\models \alpha$ — contradiction to both being equivalent. So assume $m \not\models \psi$. Let $m' \sim_{\{x_1,\ldots,x_n\}} m$ such that $m' \models \neg x_i$ for all i. Then $m' \models \lambda[(\chi_i/x_i)_i]$ but $m \not\models \alpha$ — contradiction to both being equivalent. (ii) is proved analogously. □

Proposition 3.10. *Let L be a finitely generated propositional language.*
Let $x_1, \ldots, x_n \notin L$ be additional propositional variables.
Let $\blacktriangle \wedge \psi$ be a formula such that $\blacktriangle \in L$ and $Cn(\blacktriangle) = Cn(\blacktriangle \wedge \psi) \cap L$.
Let $\sigma = (\alpha_m, \ldots, \alpha_1)$ be a sequence of formulae with
$\alpha_i \equiv \varphi_i \wedge \bigwedge_{1 \leq j \leq l_i} (\theta_{ij} \vee \bigvee_{p \in P_{ij}} x_p \vee \bigvee_{q \in N_{ij}} \neg x_q)$ *such that* $\varphi_i, \theta_{ij} \in L$
Let $\sigma' = (\alpha'_m, \ldots, \alpha'_1)$ with $\alpha'_i = \alpha_i[(x_k / x_k \wedge \psi)_k]$, that is,
$\alpha'_i \equiv \varphi_i \wedge \bigwedge_{1 \leq j \leq l_i} (\theta_{ij} \vee \bigvee_{p \in P_{ij}} (x_p \wedge \psi) \vee \bigvee_{q \in N_{ij}} \neg(x_q \wedge \psi))$
Then $f(\sigma \cdot \blacktriangle \wedge \psi) \equiv f(\psi \cdot \sigma' \cdot \blacktriangle)$.

Proof. Due to $Cn(\blacktriangle) = Cn(\blacktriangle \wedge \psi) \cap L$ we can restrict ourselves to the case where $\blacktriangle \wedge \psi$ is consistent as otherwise the two are trivially equivalent.

We will show that for any $C \subseteq \{1, \ldots, m\}$ we have $\blacktriangle \wedge \psi \wedge \bigwedge_{c \in C} \alpha_c \vdash \bot$ if and only if $\blacktriangle \wedge \bigwedge_{c \in C} \alpha'_c \vdash \bot$. The proposition then follows from a simple induction over the length of σ showing that both collect the corresponding elements from σ / σ'. Note that $\psi \wedge \alpha_j \equiv \psi \wedge \alpha'_j$ and $\psi \wedge \alpha_j \vdash \alpha'_j$. Adding ψ in $f(\psi \cdot \sigma' \cdot \blacktriangle)$ must be possible (otherwise $f(\sigma \cdot \blacktriangle \wedge \psi)$ would have to be inconsistent which is not possible given $\blacktriangle \wedge \psi$ is consistent) and make the two formulae equivalent.

$\blacktriangle \wedge \bigwedge_{c \in C} \alpha'_c \vdash \bot$ implies $\blacktriangle \wedge \psi \wedge \bigwedge_{c \in C} \alpha_c \vdash \bot$ because $\psi \wedge \alpha_j \vdash \psi \wedge \alpha'_j$.

$\blacktriangle \wedge \psi \wedge \bigwedge_{c \in C} \alpha_c \vdash \bot$ implies $\blacktriangle \wedge \bigwedge_{c \in C} \alpha_c \vdash \neg\psi$. We claim (and will later prove this claim) that in this case $\blacktriangle \wedge \bigwedge_{c \in C} \alpha'_c \vdash \neg\psi$. But then

$$\blacktriangle \wedge \bigwedge_{c \in C} \alpha'_c \equiv$$

$$\blacktriangle \wedge \bigwedge_{c \in C} \left[\varphi_c \wedge \bigwedge_{1 \leq j \leq l_c} (\theta_{cj} \vee \bigvee_{p \in P_{cj}} (x_p \wedge \psi) \vee \bigvee_{q \in N_{cj}} \neg(x_q \wedge \psi)) \right] \equiv$$

$$\blacktriangle \wedge \bigwedge_{c \in C} \left[\varphi_c \wedge \bigwedge_{1 \leq j \leq l_c} (\theta_{cj} \vee \bigvee_{q \in N_{cj}} \top) \right].$$

Note that the last formula does not contain any x_i and hence is an element of L. So $\blacktriangle \wedge \bigwedge_{c \in C} \left[\varphi_c \wedge \bigwedge_{1 \leq j \leq l_c} (\theta_{cj} \vee \bigvee_{q \in N_{cj}} \top) \right] \vdash \neg\psi$.

$\blacktriangle \wedge \psi \vdash \neg \bigwedge_{c \in C} \left[\varphi_c \wedge \bigwedge_{1 \leq j \leq l_c} (\theta_{cj} \vee \bigvee_{q \in N_{cj}} \top) \right]$. But any $\theta \in L$ entailed by $\blacktriangle \wedge \psi$ is already entailed by $\blacktriangle$ alone.

So $\blacktriangle \vdash \neg \bigwedge_{c \in C} \left[\varphi_c \wedge \bigwedge_{1 \leq j \leq l_c} (\theta_{cj} \vee \bigvee_{q \in N_{cj}} \top) \right]$ and hence $\blacktriangle \wedge \bigwedge_{c \in C} \left[\varphi_c \wedge \bigwedge_{1 \leq j \leq l_c} (\theta_{cj} \vee \bigvee_{q \in N_{cj}} \top) \right] \vdash \bot$.
Using the above equivalences, this means $\blacktriangle \wedge \bigwedge_{c \in C} \alpha'_c \vdash \bot$ as needed, concluding this part of the proof.

We still need to prove the claim that $\blacktriangle \wedge \bigwedge_{c \in C} \alpha_c \vdash \neg\psi$ implies $\blacktriangle \wedge \bigwedge_{c \in C} \alpha'_c \vdash \neg\psi$. Assume this is not the case, that is there is an assignment m with $m \models \blacktriangle \wedge \bigwedge_{c \in C} \alpha'_c$ and $m \models \psi$. Consider an arbitrary α'_c $(c \in C)$. $m \models \varphi_c \wedge \bigwedge_{1 \leq j \leq l_c} (\theta_{cj} \vee \bigvee_{p \in P_{cj}} (x_p \wedge \psi) \vee \bigvee_{q \in N_{cj}} \neg(x_q \wedge \psi))$. So $m \models \varphi_c$ further consider an arbitrary $j : 1 \leq j \leq l_c$ If $m \models \bigvee_{p \in P_{cj}} (x_p \wedge \psi)$ then $m \models \bigvee_{p \in P_{cj}} x_p$ and if $m \models \bigvee_{q \in N_{cj}} \neg(x_q \wedge \psi))$ then $m \models \bigvee_{q \in N_{cj}} \neg x_q$. But this means $m \models \alpha_c$

So $m \models \alpha_c$ for any $c \in C$ and consequently $m \models \blacktriangle \wedge \psi \wedge \bigwedge_{c \in C} \alpha_c$ — contradiction. □

Proposition 3.11. *If $[\rho, \blacktriangle]$ explains $o[(\chi_i/x_i)_i]$ then there exist $\blacktriangle'$ and ψ such that $\blacktriangle'$ does not contain any x_i and $[\rho \cdot \psi, \blacktriangle']$ explains $o[(\chi_i/x_i \wedge \psi)_i]$.*

Proof. Let L be the language the parametrised observation o is based on and $x_i \notin L$ for any i. $\blacktriangle'$ and ψ such that $\blacktriangle' \wedge \psi \equiv \blacktriangle$ and $Cn(\blacktriangle') = Cn(\blacktriangle' \wedge \psi) \cap L$ can be constructed from $\blacktriangle$. This is possible by transforming $\blacktriangle$ into CNF, calculating the closure of all resolvents, which is finite as L is finite. Then choose $\blacktriangle'$ such that it represents all clauses not containing any x_i and ψ to represent the rest. We can then apply Proposition 3.10 directly. Note that any formula appearing in $o[(\chi_i/x_i)_i]$ can be written as $\alpha \equiv \varphi \wedge \bigwedge_{1 \leq j \leq l} (\theta_j \vee \bigvee_{p \in P_j} x_p \vee \bigvee_{q \in N_j} \neg x_q)$ such that $\varphi, \theta_j \in L$. This is because α is nothing but the conjunctive normal form of the formula where the additional variables x_i are explicitly represented. Proposition 3.8 further yields that $o[(\chi_i/x_i \wedge \psi)_i]$ gives rise to the correct α'_i that are used in 3.10.

By setting $\sigma = \iota_j(o[(\chi_i/x_i)_i])$ and $\sigma' = \iota_j(o[(\chi_i/x_i \wedge \psi)_i])$ for all j, we get that for any point in the observation both construct an equivalent formula before processing ρ, hence the two initial epistemic states yield the same belief set at every step of the corresponding observations. □

Proposition 3.12. *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 $Cn(\blacktriangle'') = Cn(\blacktriangle') \cap L$.*

Then $\blacktriangle \vdash \blacktriangle''$.

Proof. Proposition 3.6 yields that $\left[\rho, \blacktriangle \wedge \bigwedge_{1 \leq i \leq n} x_i \leftrightarrow \phi_i\right]$ is an explanation for $o[(\chi_i/x_i)_i]$. By Proposition 2.35 we get $\blacktriangle \wedge \bigwedge_{1 \leq i \leq n} x_i \leftrightarrow \phi_i \vdash \blacktriangle'$ and hence $\blacktriangle \wedge \bigwedge_{1 \leq i \leq n} x_i \leftrightarrow \phi_i \vdash \blacktriangle''$.

Now assume $\blacktriangle \nvdash \blacktriangle''$, i.e., there is an assignment m such that $m \models \blacktriangle$ but $m \not\models \blacktriangle''$. As $\blacktriangle''$ does not contain any x_i we can construct an assignment $m' \sim_{\{x_1, \ldots, x_n\}} m$ such that $m' \models \blacktriangle \wedge \bigwedge_{1 \leq i \leq n} x_i \leftrightarrow \phi_i$ but $m' \not\models \blacktriangle''$ by setting $m' \models x_i$ iff $m \models \phi_i$. But this leads to a contradiction to $\blacktriangle \wedge \bigwedge_{1 \leq i \leq n} x_i \leftrightarrow \phi_i \vdash \blacktriangle''$, so indeed $\blacktriangle \vdash \blacktriangle''$. □

Proposition 3.13. *Let $[\rho, \blacktriangle]$ be the rational explanation for $o[(\chi_i/x_i)_i]$ and $\blacktriangle'$ such that $Cn(\blacktriangle') = Cn(\blacktriangle) \cap L(o)$.*

Then $\blacktriangle'$ is the unique weakest o-acceptable core.

Proof. Proposition 3.12 yields that any o-acceptable core will entail $\blacktriangle'$ and (the proof of) Proposition 3.11 tells us that $\blacktriangle'$ is indeed o-acceptable. □

Proposition 3.15. *If $\blacktriangle_1$ is o_1-acceptable for $o_1 = o[(\chi_i/\phi_i^1)_i]$ and $\blacktriangle_2$ is o_2-acceptable for $o_2 = o[(\chi_i/\phi_i^2)_i]$ then there are formulae $\phi_1', \ldots, \phi_n'$ such that $\blacktriangle_1 \vee \blacktriangle_2$ is o'-acceptable for $o' = o[(\chi_i/\phi_i')_i]$.*

Proof. Let $L = L(o)$ and $x_1, \ldots, x_n \notin L$ be propositional variables not appearing in $\blacktriangle_j$ or ϕ_i^j, $j \in \{1, 2\}$, $1 \leq i \leq n$. Proposition 3.6 yields that both $\blacktriangle_1 \wedge \bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i^1)$ and $\blacktriangle_2 \wedge \bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i^2)$ are o''-acceptable with $o'' = o[(\chi_i/x_i)_i]$. Let ψ_j denote $\bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i^j)$. Proposition 2.35 yields that $[\rho, (\blacktriangle_1 \wedge \psi_1) \vee (\blacktriangle_2 \wedge \psi_2)]$ explains o'' for some ρ. The core can equivalently be written as $(\blacktriangle_1 \vee \blacktriangle_2) \wedge (\blacktriangle_1 \vee \psi_2) \wedge (\blacktriangle_2 \vee \psi_1) \wedge (\psi_1 \vee \psi_2)$. Let ψ denote $(\blacktriangle_1 \vee \psi_2) \wedge (\blacktriangle_2 \vee \psi_1) \wedge (\psi_1 \vee \psi_2)$. Hence we have $(\blacktriangle_1 \vee \blacktriangle_2) \wedge \psi$ is o''-acceptable. If we can show that $Cn((\blacktriangle_1 \vee \blacktriangle_2) \wedge \psi) \cap L = Cn(\blacktriangle_1 \vee \blacktriangle_2)$ then we can apply Proposition 3.10 to show that $\blacktriangle_1 \vee \blacktriangle_2$ is o'-acceptable for $o' = o[(\chi_i/x_i \wedge \psi)_i]$. By setting $\sigma = \iota_j(o[(\chi_i/x_i)_i])$ and $\sigma' = \iota_j(o[(\chi_i/x_i \wedge \psi)_i])$ for all j, we get that for any point in the observation both construct an equivalent formula before processing ρ, hence the two initial epistemic states yield the same belief set at every step of the corresponding observations.

Let $\theta \in L$. $\blacktriangle_1 \vee \blacktriangle_2 \vdash \theta$ implies $(\blacktriangle_1 \vee \blacktriangle_2) \wedge \psi \vdash \theta$. So assume $(\blacktriangle_1 \vee \blacktriangle_2) \wedge \psi \vdash \theta$ and $\blacktriangle_1 \vee \blacktriangle_2 \nvdash \theta$. This means $\exists m : m \models \blacktriangle_1 \vee \blacktriangle_2$ but $m \not\models \theta$. Without loss of generality assume $m \models \blacktriangle_1$. Let $m' \sim_{\{x_1,\ldots,x_n\}} m$ such that $m' \models x_i$ if and only if $m \models \phi_i^1$. As a consequence $m' \models \blacktriangle_1$ and $m' \models \bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i^1)$ and hence $m' \models (\blacktriangle_1 \vee \blacktriangle_2) \wedge (\blacktriangle_1 \vee \psi_2) \wedge (\blacktriangle_2 \vee \psi_1) \wedge (\psi_1 \vee \psi_2)$. In other words, $m' \models (\blacktriangle_1 \vee \blacktriangle_2) \wedge \psi$ and $m' \not\models \theta$, contradicting $(\blacktriangle_1 \vee \blacktriangle_2) \wedge \psi \vdash \theta$. □

Proposition 3.17. *Let o be a parametrised observation based on L. An explanation of o restricted to L exists if and only if $\blacktriangle \wedge \bigwedge_{1\leq i\leq n} x_i \leftrightarrow \phi_i$ is $o[(\chi_i/x_i)_i]$-acceptable for some $\blacktriangle, \phi_i \in L$ and $x_i \notin L$.*

Proof. Assume there are instantiations $\phi_i \in L$ for the unknown subformulae χ_i in o such that $o[(\chi_i/\phi_i)_i]$ has an explanation. Using the rational explanation construction on $o[(\chi_i/\phi_i)_i]$ yields that in this case there must exist an explanation restricted to L, as that construction does not invent new variables and returns an explanation if one exists (Proposition 2.62). Let $\blacktriangle$ be the corresponding core belief. Now, Proposition 3.6 immediately yields that $\blacktriangle \wedge \bigwedge_{1\leq i\leq n} x_i \leftrightarrow \phi_i$ is $o[(\chi_i/x_i)_i]$-acceptable.

Now assume that $\blacktriangle \wedge \bigwedge_{1\leq i\leq n} x_i \leftrightarrow \phi_i$ is $o[(\chi_i/x_i)_i]$-acceptable for some $\blacktriangle, \phi_i \in L$ and $x_i \notin L$, ρ being the rational prefix of $o[(\chi_i/x_i)_i]$ using that core belief. We can than use Proposition 3.4 to show that $[\rho[(x_i/\phi_i)_i], \blacktriangle]$ explains $o[(\chi_i/\phi_i)_i]$. This is because we can think of ρ as sequence equivalent to some $\rho'[(\chi_i/x_i)_i]$ and hence $\rho[(x_i/\phi_i)_i] \equiv \rho'[(\chi_i/\phi_i)_i]$. □

Proposition 3.23. *Let* $o = \langle (\varphi_1, \theta_1, D_1), (\psi_{11}, \top, \emptyset), \ldots, (\psi_{1m_1}, \top, \emptyset),$
$(\varphi_2, \theta_2, D_2), (\psi_{21}, \top, \emptyset), \ldots, (\psi_{2m_2}, \top, \emptyset),$
$\ldots,$
$(\varphi_{n-1}, \theta_{n-1}, D_{n-1}), (\psi_{(n-1)1}, \top, \emptyset), \ldots, (\psi_{(n-1)m_{n-1}}, \top, \emptyset),$
$(\varphi_n, \theta_n, D_n)\rangle$
and $\iota'_i(o) = (\varphi_1, \psi_{11}, \ldots, \psi_{1m_1}, \varphi_2, \ldots, \psi_{(i-1)m_{i-1}}, \varphi_i)$ *denote the prefix of* $\iota(o)$ *with* φ_i *being the last element.*

If $[\rho, \blacktriangle]$ *explains* o*, then it also explains*

$$\begin{aligned} o' = \langle & (\varphi_1, \theta_1, D_1), (f(\iota'_2(o) \cdot \blacktriangle), \top, \emptyset), \\ & (\varphi_2, \theta_2, D_2), (f(\iota'_3(o) \cdot \blacktriangle), \top, \emptyset), \\ & \ldots, \\ & (\varphi_{n-1}, \theta_{n-1}, D_{n-1}), (f(\iota'_n(o) \cdot \blacktriangle), \top, \emptyset), \\ & (\varphi_n, \theta_n, D_n)\rangle. \end{aligned}$$

Proof. It suffices to show that $f\big(\varphi_1, \psi_{11}, \ldots, \psi_{1m_1}, \varphi_2, \ldots, \psi_{(i-1)m_{i-1}}, \varphi_i, \blacktriangle\big)$ is equivalent to $f(\varphi_1, f(\iota'_2(o) \cdot \blacktriangle), \varphi_2, \ldots, f(\iota'_i(o) \cdot \blacktriangle), \varphi_i, \blacktriangle)$. This implies that in both cases an equivalent formula has been collected before processing ρ and hence the beliefs after receiving the i^{th} known input φ_i are the same in both cases.

By Proposition 2.3 $f(\varphi_1, f(\iota'_2(o) \cdot \blacktriangle), \varphi_2, \ldots, f(\iota'_i(o) \cdot \blacktriangle), \varphi_i, \blacktriangle)$ is equivalent to $f(\varphi_1, f(\iota'_2(o) \cdot \blacktriangle), \varphi_2, \ldots, f(\iota'_i(o) \cdot \blacktriangle), f(\varphi_i, \blacktriangle))$. Obviously, $f(\iota'_i(o) \cdot \blacktriangle) \vdash f(\varphi_i, \blacktriangle)$ so using Proposition 2.7 we get $f(\varphi_1, f(\iota'_2(o) \cdot \blacktriangle), \varphi_2, \ldots, f(\iota'_i(o) \cdot \blacktriangle), f(\varphi_i, \blacktriangle))$ is equivalent to $f(\varphi_1, f(\iota'_2(o) \cdot \blacktriangle), \varphi_2, \ldots, f(\iota'_i(o) \cdot \blacktriangle))$. Now, by the definition of $\iota'_i(o)$ we know that $f\big(\varphi_1, \psi_{11}, \ldots, \psi_{1m_1}, \varphi_2, \ldots, \psi_{(i-1)m_{i-1}}, \varphi_i, \blacktriangle\big)$ just the same as $f(\iota'_i(o) \cdot \blacktriangle)$, so if we can show that $f(\varphi_1, f(\iota'_2(o) \cdot \blacktriangle), \varphi_2, \ldots, f(\iota'_i(o) \cdot \blacktriangle))$ is equivalent to $f(\iota'_i(o) \cdot \blacktriangle)$ we are done.

But this follows from the fact that each remaining formula is either already entailed by $f(\iota'_i(o) \cdot \blacktriangle)$ or inconsistent with it. For the φ_j with $j < i$ this is obvious. It appears in $\iota'_i(o)$ and hence is entailed or inconsistent (Proposition 2.6). For the $f\Big(\iota'_j(o) \cdot \blacktriangle\Big)$, $j < i$, recall the definition of $\iota'_j(o)$. Clearly $\iota'_j(o)$ is a prefix of $\iota'_i(o)$, so either $f(\iota'_i(o) \cdot \blacktriangle)$ accepts all the formulae accepted by $f\Big(\iota'_j(o) \cdot \blacktriangle\Big)$ in which case it is entailed by $f(\iota'_i(o) \cdot \blacktriangle)$, or $f\Big(\iota'_j(o) \cdot \blacktriangle\Big)$ accepts a formula $f(\iota'_i(o) \cdot \blacktriangle)$ rejects, in which case the two are inconsistent. □

Proposition 3.25. *If* $x \notin L(o_1 \cdot o_2)$ *and* $Cn(\blacktriangle) = Cn(\blacktriangle_\vee(o_1 \cdot \langle (x, \top, \emptyset)\rangle \cdot o_2)) \cap L(o_1 \cdot o_2)$ *then* $\blacktriangle_\vee(o_1 \cdot o_2) \vdash \blacktriangle$.

Proof. $\blacktriangle_\vee(o_1 \cdot o_2) \equiv \blacktriangle_\vee(o_1 \cdot \langle(\top, \top, \emptyset)\rangle \cdot o_2)$. The additional revision input $\top$ has no effect on the epistemic state whatsoever. By Proposition 3.5 $\blacktriangle_\vee(o_1 \cdot \langle(\top, \top, \emptyset)\rangle \cdot o_2) \wedge x$ is $o_1 \cdot \langle(x, \top, \emptyset)\rangle \cdot o_2$-acceptable and hence entails $\blacktriangle_\vee(o_1 \cdot \langle(x, \top, \emptyset)\rangle \cdot o_2)$. So by Proposition 3.12 we know $\blacktriangle_\vee(o_1 \cdot o_2) \wedge x \vdash \blacktriangle$, where $Cn(\blacktriangle) = Cn(\blacktriangle_\vee(o_1 \cdot \langle(x, \top, \emptyset)\rangle \cdot o_2))$. Applying Proposition 3.4 we can show that every formula from $L(o_1 \cdot o_2)$ that follows from $\blacktriangle_\vee(o_1 \cdot o_2) \wedge x$ is already entailed by $\blacktriangle_\vee(o_1 \cdot o_2)$. □

Proposition 3.26. *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 $f(\sigma \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \equiv f(\sigma' \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle))$ for all $1 \leq i \leq n$.*

Proof. The proposition is trivial for $n \leq m$. In this case we can simply set $\sigma' = \sigma \cdot (\top, \ldots, \top)$, appending tautologies until σ' has the right length. These tautologies have no impact on the logical content of the formula constructed by f (Proposition 2.8).

Now assume $m > n$. Due to Proposition 2.3 it suffices to show $f(\sigma \cdot f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$ is equivalent to $f(\sigma' \cdot f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$. The proof of Proposition 2.27 tells us that for every σ there is a logical chain behaving exactly as that sequence. We will now argue that a logical chain of length at most n exists that behaves like σ at least for the required formulae $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$. Note that the above requirement translates into the observation $o = \langle(f(\varphi_1), f(\sigma \cdot (\varphi_1, \blacktriangle)), \emptyset), \ldots, (f(\varphi_1, \ldots, \varphi_n, \blacktriangle), f(\sigma \cdot f(\varphi_1, \ldots, \varphi_n, \blacktriangle)), \emptyset)\rangle$, i.e., all D_i are empty to start with. However, we interpret the $f(\sigma \cdot f(\varphi_1, \ldots, \varphi_n, \blacktriangle))$ to be complete characterisations of the beliefs. As there is a sequence σ giving rise to this observation, the rational explanation construction must find one as well. Ultimately we will choose σ' to be the sequence calculated. We know the core belief $\blacktriangle$, so we can start the algorithm with the correct instantiation of the core. The rational explanation of o (or rather the calculation of $\rho_R(o, \blacktriangle)$ the rational prefix of o with respect to $\blacktriangle$) need not yield the exact beliefs required. Assume there is an i such that $\lambda = f(\sigma' \cdot f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$ is logically stronger than $f(\sigma \cdot f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$ (which is possible only because $f(\sigma')$ is weaker than $f(\sigma)$). In that case we simply add λ to D_i and start again until the beliefs indicated in the observation are exactly realised. This must be the case at some point since such a sequence exists.

Now, what length can σ' have? There are n positive conditionals. At each iteration of the rational closure algorithm (Definition 2.45) at least one positive conditional must leave $\mathcal{C}_i$. If not, all positive conditionals are p-exceptional and will remain so which implies that $\blacktriangle$ is not o-acceptable which it is. But if one positive conditional leaves $\mathcal{C}_i$ at every iteration, there can be at most n non-tautological $\bigwedge U_i$ and the tautologies at the beginning of the sequence can be omitted.

A note on the case, where ρ contains intermediate inputs as well, i.e., where we try to replace an earlier block of intermediate inputs. Here we need the equivalence of σ and σ' only for

those appended formulae $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ where φ_i is a recorded input. It does not matter how many intermediate inputs there are between φ_1 and φ_i. This is because we are not interested in the beliefs after an intermediate input has been received but only those after recorded inputs. And as shown above those intermediate inputs can be replaced by a block of correct length that allows construction of an equivalent formula. That is, we can assume to have an equivalent formula before σ is processed even if later intermediate inputs are already replaced.

Now, the implication constructed from a positive conditional belief where the input is an intermediate one is a tautology as the corresponding θ is one. However, if only a tautology is eliminated from $\mathcal{C}_i$, then all the remaining conditionals must remain exceptional. So the only way to ensure termination is that in every iteration of the rational closure construction at least one non-trivial positive conditional, i.e., one arising from a recorded revision input, is satisfied. Hence the we need only as many intermediate inputs as recorded inputs appearing later and can ignore intermediate inputs coming afterwards. □

A.3 Proofs from Chapter 4

Proposition 4.4. *Let (φ_i, k_i), $1 \leq i \leq n$, be a set of n revision inputs with corresponding indexes. Further let $\blacktriangle$ be a core belief and ρ_p, ρ_s and ρ_s^i, $1 \leq i \leq n$, sequences such that $[\rho_p \cdot \rho_s, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i) = \left[\rho_p \cdot \rho_s^i, \blacktriangle\right]$ for all $1 \leq i \leq n$.*

Then $Bel([\rho_p \cdot \rho_s, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i)) = Bel([\rho_p, \blacktriangle] *_I (f\left(\rho_s^i \cdot \blacktriangle\right), 1))$ for all $1 \leq i \leq n$.

Proof.

$$
\begin{aligned}
Bel([\rho_p \cdot \rho_s, \blacktriangle] *_I (\varphi_1, k_1) *_I \ldots *_I (\varphi_i, k_i)) &= Bel(\left[\rho_p \cdot \rho_s^i, \blacktriangle\right]) \quad \text{above condition} \\
&= Cn(f\left(\rho_p \cdot \rho_s^i \cdot \blacktriangle\right)) \quad \text{Definition 4.2} \\
&= Cn(f\left(\rho_p \cdot f\left(\rho_s^i \cdot \blacktriangle\right)\right)) \quad \text{Proposition 2.3} \\
&= Cn(f\left(\rho_p \cdot f\left(\rho_s^i \cdot \blacktriangle\right) \cdot \blacktriangle\right)) \quad \text{Proposition 2.7} \\
&= Bel([\rho_p, \blacktriangle] *_I (f\left(\rho_s^i \cdot \blacktriangle\right), 1)) \quad \text{Definition 4.2}
\end{aligned}
$$

□

Proposition 4.10. *Let $*_k \in \{*, *_\blacktriangle\}$ for all $1 \leq k \leq n$. $[\rho, \rho_\blacktriangle]$ explains*

$$
\begin{aligned}
o = \langle &((\varphi_1, *_1), \theta_1, D_1), \ldots, ((\varphi_{i-1}, *_{i-1}), \theta_{i-1}, D_{i-1}), \\
&((\bot, *_\blacktriangle), \theta_i, \emptyset), \\
&((\varphi_{i+1}, *), \theta_{i+1}, \emptyset), \ldots, ((\varphi_j, *), \theta_j, \emptyset),
\end{aligned}
$$

$$((\varphi_{j+1}, *_{\blacktriangle}), \theta_{j+1}, D_{j+1}),$$
$$((\varphi_{j+2}, *_{j+2}), \theta_{j+2}, D_{j+2}), \ldots, ((\varphi_n, *_n), \theta_n, D_n)\rangle$$

if and only if it explains

$$\begin{aligned} o' = \langle &((\varphi_1, *_1), \theta_1, D_1), \ldots, ((\varphi_{i-1}, *_{i-1}), \theta_{i-1}, D_{i-1}), \\ &((\varphi_{i+1}, *), \top, \emptyset), \ldots, ((\varphi_j, *), \top, \emptyset), \\ &((\varphi_{j+1}, *_{\blacktriangle}), \theta_{j+1}, D_{j+1}), \\ &((\varphi_{j+2}, *_{j+2}), \theta_{j+2}, D_{j+2}), \ldots, ((\varphi_n, *_n), \theta_n, D_n)\rangle \end{aligned}$$

Proof. First note that if there is no core revision input following $\varphi_i \equiv \bot$, i.e. $j = n$, then the observation ends after the record for the input φ_j and all the entries in o beginning with that of the inconsistent core revision input are modified in order to obtain o'. Both observations record exactly the same revision inputs in the same order using the same revision function, the only exception being the additional inconsistent core revision input in o.

We will now look at the epistemic state of the agent after an arbitrary number of revision steps according to o and the corresponding revisions steps in o'. With respect to o, the agent may have received the additional inconsistent core revision input φ_i. We will see that the beliefs are either equivalent or irrelevant implying that the epistemic state explains one of the two observations if and only if it also explains the other one. We distinguish three cases.

- The agent has not received the inconsistent core revision input φ_i. In this case the resulting epistemic state is identical for both observations — $[\rho \cdot \sigma_1, \rho_{\blacktriangle} \cdot \sigma_2]$, where σ_1 is the sequence of regular revision inputs and σ_2 the sequence of core revision inputs received up to that point. As the epistemic states are identical, the beliefs will be the exactly the same.

- With respect to o the agent has received the inconsistent core revision input φ_i but not yet the next core revision input φ_{j+1}. That is, the agent's epistemic state is $[\rho \cdot \sigma_1, \rho_{\blacktriangle} \cdot \sigma_2 \cdot \bot]$ with the above interpretation of the σ_i. With respect to o' where φ_i is not recorded the epistemic state would be $[\rho \cdot \sigma_1, \rho_{\blacktriangle} \cdot \sigma_2]$.

  Note that in this case, the conditions for the epistemic states explaining an observation are trivially satisfied. The beliefs in the first epistemic state are inconsistent, that is any formula is entailed. The beliefs in the second one are irrelevant as any formula entails a tautology. The non-beliefs in this case are empty sets, so we need not worry about formulae that are not to be entailed by the agent's beliefs.

- With respect to o, the agent has received the inconsistent core revision input φ_i but also the next core revision input φ_{j+1}. In this case the epistemic states are

$[\rho \cdot \sigma_1, \rho_{\blacktriangle} \cdot \sigma_2 \cdot \bot \cdot \sigma_3]$ and $[\rho \cdot \sigma_1, \rho_{\blacktriangle} \cdot \sigma_2 \cdot \sigma_3]$. Here σ_2 is the sequence of core revision inputs received before $\varphi_i \equiv \bot$ and σ_3 the non-empty sequence of core revision inputs received after φ_i. Due to Proposition 2.8 the additional contradiction has no impact on the beliefs as it is inserted in a proper prefix of the argument sequence passed to f. That is, the belief sets will be equivalent. □

Appendix B

A Note on Computational Complexity

In this chapter we want to investigate a few complexity related issues concerning observations and their explanations. We already gave two complexity results. Propositions 2.31 and 2.52 state that deciding whether an epistemic state $[\rho, \blacktriangle]$ explains an observation o as well as deciding o-acceptability of a core belief $\blacktriangle$ are both Δ_2^P-complete. Some problems for which complexity results might also be interesting are whether a given core belief $\blacktriangle$ is the weakest o-acceptable core ($\blacktriangle \equiv \blacktriangle_\vee(o)$) and whether an explanation for o exists at all.

Using the rational explanation algorithm, these problems can be decided. We claimed that this algorithm may not be suitable for giving a complexity bound of these problems as it may need an exponential number of iterations. This would yield that these problems are in $EXPTIME$. We will first illustrate the example requiring so many iterations and then develop more specific (but not necessarily tight) complexity bounds for those two problems.

B.1 Why the rational explanation may need exponentially many iterations

Consider the observation $o = \langle (p_1, \top, \emptyset), \ldots, (p_n, \top, \emptyset), (p_{n+1}, \theta, \emptyset) \rangle$ where $\theta = \bigwedge_{1 \leq i \leq n+1} \neg p_i$. All p_j, $1 \leq j \leq n+1$, are distinct propositional variables. o does not give rise to any negative conditionals as $D_i = \emptyset$ for all i.

The material counterparts of positive conditionals with index smaller than $n+1$ will be tautologies, no matter what the core will look like. This is because the consequent of every such conditional is a tautology. That is, in the rational prefix construction $\bigwedge \tilde{\mathcal{C}}_i \equiv \top$ in

case $f(\iota \cdot \blacktriangle) \Rightarrow \theta \notin \mathcal{C}$, $\bigwedge \tilde{\mathcal{C}}_i \equiv f(\iota \cdot \blacktriangle) \rightarrow \theta$ otherwise. As there are no negative conditionals we have $U_i = \tilde{\mathcal{C}}_i$. We will start by showing that whenever there are ultimately exceptional conditionals when using some core belief $\blacktriangle$, then $\alpha_m = \bigwedge U_m \equiv \neg f(\iota \cdot \blacktriangle)$. That is, in the rational explanation algorithm the core belief $\blacktriangle$ will be strengthened by $\neg f(\iota \cdot \blacktriangle)$.

Let there be ultimately exceptional conditionals when constructing the rational prefix of o with respect to some consistent core $\blacktriangle$. Assume the conditional with index $n+1$ is not among the ultimately exceptional ones. This implies $\alpha_m = \bigwedge U_m \equiv \top$. But no conditional can be exceptional for U_m as $f(\iota_i \cdot \blacktriangle)$ is consistent whenever $\blacktriangle$ is — contradiction. So the conditional with index $n+1$ is among the ultimately exceptional ones. Using the above remark we get $\bigwedge U_m = \bigwedge \tilde{\mathcal{C}}_m \equiv f(\iota \cdot \blacktriangle) \rightarrow \theta$. Due to the exceptionality of the conditional we know $f(\iota \cdot \blacktriangle) \rightarrow \theta \vdash \neg f(\iota \cdot \blacktriangle)$ which implies $f(\iota \cdot \blacktriangle) \rightarrow \theta \equiv \neg f(\iota \cdot \blacktriangle)$ as claimed.

From the structure of the observation it is obvious that as long as $f(\iota \cdot \blacktriangle) \vdash p_i$ for a single i the core belief $\blacktriangle$ cannot be o-acceptable, i.e., as long as $\blacktriangle \not\equiv \theta$. The agent cannot consistently believe both p_i and $\neg p_i$. In this case there are ultimately exceptional conditionals implying that the core belief in the next iteration will be $\blacktriangle \wedge \neg f(\iota \cdot \blacktriangle)$. As $\blacktriangle$ is constructed from variables appearing in o and due to Proposition 2.6 we know that $f(\iota \cdot \blacktriangle)$ must be a conjunction of literals for each variable p_i, $1 \leq i \leq n+1$. Hence, $\neg f(\iota \cdot \blacktriangle)$ will always be a *disjunction* of the negated literals.

The question is how $\blacktriangle$ and $f(\iota \cdot \blacktriangle)$ will evolve when starting the rational explanation algorithm with $\blacktriangle_0 = \top$. The following table intends to illustrate that for the case $n = 3$. We will not formally prove for arbitrary n that the evolution will be analogous, but the example should make plausible that this is indeed the case. j is the number of the iteration in the rational explanation construction. $\blacktriangle_j$ is the core belief used in that iteration. $f(\iota \cdot \blacktriangle_j)$ the antecedent of the only relevant conditional and $\neg f(\iota \cdot \blacktriangle_j)$ its negation. For the sake of readability we will use $\overline{p_i}$ instead of $\neg p_i$.

j	$\blacktriangle_j$	$f(\iota \cdot \blacktriangle_j)$	$\neg f(\iota \cdot \blacktriangle_j)$
0	$\top$	$p_1 \wedge p_2 \wedge p_3 \wedge p_4$	$\overline{p_1} \vee \overline{p_2} \vee \overline{p_3} \vee \overline{p_4}$
1	$\overline{p_1} \vee \overline{p_2} \vee \overline{p_3} \vee \overline{p_4}$	$\overline{p_1} \wedge p_2 \wedge p_3 \wedge p_4$	$p_1 \vee \overline{p_2} \vee \overline{p_3} \vee \overline{p_4}$
2	$\overline{p_2} \vee \overline{p_3} \vee \overline{p_4}$	$p_1 \wedge \overline{p_2} \wedge p_3 \wedge p_4$	$\overline{p_1} \vee p_2 \vee \overline{p_3} \vee \overline{p_4}$
3	$\blacktriangle_2 \wedge (\overline{p_1} \vee p_2 \vee \overline{p_3} \vee \overline{p_4})$	$\overline{p_1} \wedge \overline{p_2} \wedge p_3 \wedge p_4$	$p_1 \vee p_2 \vee \overline{p_3} \vee \overline{p_4}$
4	$\overline{p_3} \vee \overline{p_4}$	$p_1 \wedge p_2 \wedge \overline{p_3} \wedge p_4$	$\overline{p_1} \vee \overline{p_2} \vee p_3 \vee \overline{p_4}$
5	$\blacktriangle_4 \wedge (\overline{p_1} \vee \overline{p_2} \vee \overline{p_4})$	$\overline{p_1} \wedge p_2 \wedge \overline{p_3} \wedge p_4$	$p_1 \vee \overline{p_2} \vee p_3 \vee \overline{p_4}$
6	$\blacktriangle_4 \wedge (\overline{p_2} \vee \overline{p_4})$	$p_1 \wedge \overline{p_2} \wedge \overline{p_3} \wedge p_4$	$\overline{p_1} \vee p_2 \vee p_3 \vee \overline{p_4}$
7	$\blacktriangle_6 \wedge (\overline{p_1} \vee \overline{p_4})$	$\overline{p_1} \wedge \overline{p_2} \wedge \overline{p_3} \wedge p_4$	$p_1 \vee p_2 \vee p_3 \vee \overline{p_4}$
8	$\overline{p_4}$	$p_1 \wedge p_2 \wedge p_3 \wedge \overline{p_4}$	$\overline{p_1} \vee \overline{p_2} \vee \overline{p_3} \vee p_4$
	$\vdots$		

In the last complete row it is finally settled that the core belief must entail $\neg p_4$. Note that this core does not restrict any of the variables with subscript lower than 4. After another 3 iterations it will be clear that also $\neg p_3$ must be entailed and $\blacktriangle_{12}$ will be $\neg p_3 \wedge \neg p_4$. $\blacktriangle_{14}$ will be $\neg p_2 \wedge \neg p_3 \wedge \neg p_4$ and $\blacktriangle_{15}$ finally $\neg p_1 \wedge \neg p_2 \wedge \neg p_3 \wedge \neg p_4$.

This also shows that of the many (logically different) formulae from a finitely generated language, only very few may be o-acceptable. In this example, only formulae equivalent to the conjunction of the negation of all propositional variables work. This illustrates that even if we are not after $\blacktriangle_{\vee}(o)$ but are willing to settle for *any* o-acceptable core, the problem is not easy. For an actual implementation much thought will have to be put in for a heuristic generating candidate core beliefs.

B.2 Deciding whether *o* has an explanation

From the complexity point of view, the "problem" of the rational explanation algorithm is that it goes deterministically through all potential core beliefs. If we were able to guess a core belief (in a proper way) and and then just test whether it is o-acceptable, we would have a non-deterministic algorithm with one Δ_2^P-oracle call[1] which yields that o has an explanation if and only if an o-acceptable core belief can be guessed. If this were possible, the problem would belong to Σ_2^P as the single oracle call can be simulated. The general problem with guessing formulae is that they are potentially exponential in size. That is, simply guessing an arbitrary formula cannot be used for deriving that complexity bound. In order to restrict the guess we make use of the particular structure of our problem. This will allow us to impose restrictions on the potential core beliefs such that the size of the formula can be polynomially bounded.

Assume the agent's real epistemic state is $[\rho, \blacktriangle]$ ($\blacktriangle$ being consistent) and it gave rise to an observation $o = \{(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\}$. Consequently o has an explanation and an o-acceptable core. Which role does $\blacktriangle$ play for that observation? The main (and arguably only) purpose of $\blacktriangle$ is to select the correct inputs φ_j when calculating $f(\varphi_1, ..., \varphi_i, \blacktriangle)$ for all $1 \leq j \leq i \leq n$.[2] That is, $\blacktriangle$ determines for which φ_j we have $f(\iota_i \cdot \blacktriangle) \vdash \varphi_j$ and for which we have $f(\iota_i \cdot \blacktriangle) \vdash \neg\varphi_j$. Formally, we can capture this impact of $\blacktriangle$ for the observation as follows.

[1] Recall that deciding o-acceptability of a given core belief is Δ_2^P-complete.

[2] The proof of Proposition B.3 reveals that the remaining beliefs at every point during the observation caused by the core can be simulated by adding the core to the sequence of the epistemic state. That is, the part of $\blacktriangle$ irrelevant for the observation can be absorbed by ρ.

Definition B.1. *Let* $o = \{(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\}$ *be an observation and* $\blacktriangle$ *a core belief. We then define for all* $1 \leq i \leq n$

$A_i(o, \blacktriangle) = \{j \in \{1, \ldots, i\} \mid \varphi_j$ *is accepted in the calculation of* $f(\iota_i \cdot \blacktriangle)\}$ *and*
$R_i(o, \blacktriangle) = \{j \in \{1, \ldots, i\} \mid \varphi_j$ *is rejected in the calculation of* $f(\iota_i \cdot \blacktriangle)\}$.

We call $AR(o, \blacktriangle) = ((A_1(o, \blacktriangle), R_1(o, \blacktriangle)), \ldots, (A_n(o, \blacktriangle), R_n(o, \blacktriangle)))$ *the* AR-characteristic *of* $\blacktriangle$ *with respect to* o.

When core belief and observation are fixed, we will omit "$(o, \blacktriangle)$" and simply write A_i and R_i. Given an observation o and a consistent core belief $\blacktriangle$, the corresponding AR-characteristic will satisfy the following obvious properties for all $1 \leq i \leq n$. The first two express that every φ_j in question must have been accepted or rejected and none can have been both accepted and rejected (c.f. Proposition 2.6). The last two say that accepted formulae must be believed and that of rejected formulae the negation must be believed.

- $A_i \cup R_i = \{1, \ldots, i\}$
- $A_i \cap R_i = \emptyset$
- $f(\iota_i \cdot \blacktriangle) \vdash \bigwedge_{j \in A_i} \varphi_j$
- $f(\iota_i \cdot \blacktriangle) \vdash \bigwedge_{j \in R_i} \neg\varphi_j$

Given the first two properties, A_i can be constructed from R_i and vice versa. Instead of guessing a potential core belief we will guess a potential AR-characteristic (or equivalently only the A_i), construct a corresponding core belief and test whether it is o-acceptable. Note that given an observation with n recorded revision inputs, the AR-characteristic contains n pairs, each pair containing sets with at most n elements.[3] We want to stress that not every sequence of pairs satisfying the first two conditions is the AR-characteristic of some core belief for some observation. Consider $((\emptyset, \{1\}), (\{1, 2\}, \emptyset))$. It expresses that the first revision input recorded was immediately rejected when received. This implies $\blacktriangle \vdash \neg\varphi_1$. But then after receiving the second input, it is suddenly believed, so in particular $\blacktriangle \nvdash \neg\varphi_1$. This is clearly impossible. In the current context, this is not a relevant problem. An explanation for o exists if and only if there is an AR-characteristic that works for some core belief. We will now give a condition relating a core belief to its AR-characteristic. This will allow us to construct a core belief *from* a potential AR-characteristic.

[3] In order for the first two conditions to be satisfied, the first pair contains exactly one number, the second two, etc. So the AR-characteristic will contain exactly $n(n+1)/2$ numbers.

Proposition B.2. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation and $\blacktriangle$ a core belief. Further let $AR = ((A_1, R_1), \ldots, (A_n, R_n))$ be the AR-characteristic of $\blacktriangle$ with respect to o. Then for all $1 \leq i \leq n$ and for all $l \in R_i$*

$$\blacktriangle \vdash \left(\bigwedge_{j \in A_i \wedge j > l} \varphi_j \wedge \bigwedge_{k \in R_i \wedge k > l} \neg\varphi_k \right) \rightarrow \neg\varphi_l.$$

Proof. (Sketch) We will show that if this condition is violated, then AR cannot be the AR-characteristic of $\blacktriangle$ with respect to o. Assume this condition is violated for some i and some l, i.e. $\blacktriangle \nvdash \left(\bigwedge_{j \in A_i \wedge j > l} \varphi_j \wedge \bigwedge_{k \in R_i \wedge k > l} \neg\varphi_k \right) \rightarrow \neg\varphi_l$. We consider the calculation of $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ which is equivalent to $f(\varphi_1, \ldots, \varphi_l, f(\varphi_{l+1}, \ldots, \varphi_i, \blacktriangle))$ (Proposition 2.3). Now $f(\varphi_{l+1}, \ldots, \varphi_i, \blacktriangle) \equiv \blacktriangle \wedge \bigwedge_{j \in A_i \wedge j > l} \varphi_j \wedge \bigwedge_{k \in R_i \wedge k > l} \neg\varphi_k$, as the AR-characteristic tells us which revision inputs are accepted and which are rejected. Due to the above assumption, this formula does not entail $\neg\varphi_l$ ($\alpha \vdash (\beta_1 \rightarrow \beta_2)$ iff $\alpha_1 \wedge \beta_1 \vdash \beta_2$). But this means φ_l is accepted contradicting $l \in R_i$! □

This proposition informs us about a number of formulae that must be entailed by the core belief. The next one expresses that these formulae suffice to construct a core with the same AR-characteristic. This will allow us to construct an o-acceptable core from the AR-characteristic of some (unknown) o-acceptable core.

Proposition B.3. *Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ be an observation, $\blacktriangle$ an o-acceptable core and $AR = ((A_1, R_1), \ldots, (A_n, R_n))$ the corresponding AR-characteristic. Let* $\psi_{il} = \left(\bigwedge_{j \in A_i \wedge j > l} \varphi_j \wedge \bigwedge_{k \in R_i \wedge k > l} \neg\varphi_k \right) \rightarrow \neg\varphi_l.$
Then $\psi = \bigwedge\{\psi_{il} \mid 1 \leq i \leq n, l \in R_i\}$ is o-acceptable. Further if $\blacktriangle \equiv \blacktriangle_\vee(o)$ then $\blacktriangle \equiv \psi$.

Proof. (Sketch) By Proposition B.2 we immediately have $\blacktriangle \vdash \psi$. If we can show that ψ is o-acceptable the second part is trivial as all o-acceptable cores entail $\blacktriangle_\vee(o)$.

The proof of ψ being o-acceptable is based on the following idea. As $\blacktriangle$ is o-acceptable there is a sequence ρ such that $[\rho, \blacktriangle]$ explains o. We will show for all $1 \leq i \leq n$ that $f(\iota_i \cdot \blacktriangle) \equiv f(\blacktriangle \cdot \iota_i \cdot \psi)$ which yields that $[\rho \cdot \blacktriangle, \psi]$ also explains o as both have constructed the same formula before processing ρ.

Let i be an arbitrary element of $\{1, \ldots, n\}$. As $\blacktriangle$ is o-acceptable it is consistent and hence ψ is consistent ($\blacktriangle \vdash \psi$). So both $f(\iota_i \cdot \blacktriangle)$ and $f(\blacktriangle \cdot \iota_i \cdot \psi)$ select the last element of the argument sequence. Assume they both have selected the same elements from ι_i down to $l + 1$. So $f(\varphi_{l+1}, \ldots, \varphi_i, \blacktriangle)$ and $f(\varphi_{l+1}, \ldots, \varphi_i, \psi)$ both entail $\bigwedge_{k \in A_i \wedge k > l} \varphi_k$ and $\bigwedge_{k \in R_i \wedge k > l} \neg\varphi_k$. If $f(\iota_i \cdot \blacktriangle)$ rejects φ_l then $l \in R_i$. Hence, $\psi \vdash \psi_{il}$, $f(\varphi_{l+1}, \ldots, \varphi_i, \psi)$ is inconsistent with φ_l and so it will

also be rejected by $f(\blacktriangle \cdot \iota_i \cdot \psi)$. If $f(\blacktriangle \cdot \iota_i \cdot \psi)$ rejects φ_l then $\psi \wedge \bigwedge_{k \in A_i \wedge k > l} \varphi_k \wedge \bigwedge_{k \in R_i \wedge k > l} \neg\varphi_k$ entails $\neg\varphi_l$ and as $\blacktriangle \vdash \psi$ also $f(\varphi_{l+1}, \ldots, \varphi_i, \blacktriangle) = \blacktriangle \wedge \bigwedge_{k \in A_i \wedge k > l} \varphi_k \wedge \bigwedge_{k \in R_i \wedge k > l} \neg\varphi_k$ entails $\neg\varphi_l$. This means $f(\iota_i \cdot \blacktriangle)$ also rejects φ_l.

As both select the same elements from ι_i and $\blacktriangle \vdash \psi$ we know that $f(\iota_i \cdot \psi)$ is consistent with $\blacktriangle$. Propositions 2.10 and 2.12 now yield that $f(\blacktriangle \cdot \iota_i \cdot \psi) \equiv f(\iota_i \cdot \psi) \wedge \blacktriangle \equiv f(\iota_i \cdot \psi \wedge \blacktriangle)$. As $\blacktriangle \vdash \psi$ this is equivalent to $f(\iota_i \cdot \blacktriangle)$ as claimed. □

Note that ψ is polynomial in the length of o. This is because there are n sets R_i and each contains at most n numbers. Consequently there are at most n^2 formulae ψ_{il}. Each ψ_{il} is an implication from the conjunction of at most $n-1$ formulae bounded by the length of o to one such formula. We are now ready to give an improved complexity result for the decision problem whether an observation o has an explanation.

Proposition B.4. *Deciding whether an observation o has an explanation is in Σ_2^P.*

Proof. It suffices to give a non-deterministic algorithm. We guess a potential AR-characteristic. This can be done by guessing n bit sequences of length $1, \ldots, n$. The j^{th} bit in the i^{th} bit sequence expresses whether $j \in A_i$. Then ψ can be constructed according to Proposition B.3 which can be done deterministically in polynomial time. Finally we check o-acceptability of ψ in polynomial time using NP-oracle calls (Proposition 2.52).

If o has an explanation then there is an o-acceptable core with a corresponding AR-characteristic. This will be among the guesses and hence the algorithm returns *true* as an o-acceptable core was constructed and tested (Proposition B.3). The algorithm returns *true* only if an explanation exists as an o-acceptable core must have been constructed for the o-acceptability test to succeed. □

Currently it is open whether this problem is also Σ_2^P-hard.

B.3 Deciding whether $\blacktriangle \equiv \blacktriangle_\vee(o)$

Now that we have a proper methodology of guessing a potential core belief, it is relatively easy to transfer the complexity result to the question of whether a given core belief is indeed $\blacktriangle_\vee(o)$. We will start with the simplified case where we know whether the given observation can be explained at all or not.

Given an observation o which has *no* explanation (and hence $\blacktriangle_\vee(o) = \bot$ by definition) and a core $\blacktriangle$, deciding $\blacktriangle_\vee(o) \equiv \blacktriangle$ is *coNP*-complete as that decision problem is simply an

unsatisfiability test. Given an observation o which has an explanation and a core $\blacktriangle$, deciding $\blacktriangle_\vee(o) \equiv \blacktriangle$ is in Π_2^P. We will show this by proving that the complement, deciding whether $\blacktriangle$ is not the weakest o-acceptable core, is in Σ_2^P.

$\blacktriangle$ is not the weakest o-acceptable core belief if and only if (i) $\blacktriangle$ is not o-acceptable or (ii) there exists an o-acceptable core $\blacktriangle'$ and $\blacktriangle' \not\models \blacktriangle$ (any o-acceptable core entails $\blacktriangle_\vee(o)$). (i) is in Δ_2^P and (ii) can be tested as follows. We guess potential core beliefs by guessing their AR-characteristic, construct the corresponding ψ, test if ψ is o-acceptable. If so we check satisfiability of $\psi \wedge \neg\blacktriangle$. If this is satisfiable then $\psi \not\models \blacktriangle$ and hence $\blacktriangle$ cannot be the weakest core belief. In particular we can guess the real $\blacktriangle_\vee(o)$ — by Proposition B.3 the corresponding ψ is equivalent to $\blacktriangle_\vee(o)$. Except for the guessing of $\blacktriangle_\vee(o)$ all steps can be done deterministically in polynomial time with NP-oracle calls.

But what if we do not know the status of o? Given an observation o and a core belief $\blacktriangle$, deciding $\blacktriangle_\vee(o) \equiv \blacktriangle$ is in Π_3^P. It may be surprising that the lack of knowledge about the o-acceptability of $\blacktriangle$ should make such a difference. However, we could not find an algorithm in a lower complexity class (which does not imply that none exists). Again, we will show that the complement, deciding $\blacktriangle_\vee(o) \not\equiv \blacktriangle$ is in Σ_3^P.

$\blacktriangle_\vee(o) \not\equiv \blacktriangle$ if and only if (i) o has no explanation and $\blacktriangle$ is satisfiable, (ii) o has an explanation but $\blacktriangle$ is not o-acceptable, or (iii) o has an explanation, $\blacktriangle$ is o-acceptable and there is a core $\blacktriangle'$ such that $\blacktriangle' \not\models \blacktriangle$ (in particular $\blacktriangle' = \blacktriangle_\vee(o)$). Note that a formula φ is consistent if and only if $\langle(\top, \varphi, \emptyset)\rangle$ has an explanation. So we can use this oracle also for checking satisfiability of a formula.

A possible non-deterministic algorithm (with Σ_2^P-oracles) could start by checking whether o has an explanation (one oracle call). If not then satisfiability of $\blacktriangle$ (one oracle call) corresponds to the answer of the decision problem (this deals with case (i)). If o has an explanation, we can simulate the test for o-acceptability of $\blacktriangle$ as that is in Δ_2^P. If $\blacktriangle$ is not o-acceptable then $\blacktriangle \not\equiv \blacktriangle_\vee(o)$ (case (ii)). Up to this point non-determinism was not used. If $\blacktriangle$ is o-acceptable we guess all possible AR-characteristics and construct the corresponding cores ψ (according to Proposition B.3). The answer to the decision problem is *yes* if one guess yields an o-acceptable ψ for which $\psi \not\models \blacktriangle$.

Again, hardness of the problems is still open.

Appendix C

Algorithms

In this thesis, we have verbally described a number of procedures. Several functions have also been given via definitions. In this chapter, we want to collect the pseudo-code of the most important algorithms. We did not include this in the main text for a number of reasons. Firstly, at the core of this work are the formal results. The algorithms are derived from these results and presenting them in the main text could be distracting. Secondly, presenting them all in one chapter should help possible implementers of the described methods. Here they find a compact representation of what needs to be done for building a system that reasons about an observed agent. Starting with this overview, they can refer to the main text for clarifying descriptions of the methods. Conversely, the "academic" reader can refer to this chapter for a compact representation of the methods described in the text. The notation used in the algorithms requires little explanation. $x \Leftarrow y$ denotes that x is assigned the value of y.

C.1 Basic functions

Once more, we will start with the basic operations of the assumed belief revision framework. For f we will give an iterative and a recursive version. Recall that the argument sequence is assumed to be non-empty.

Function $f(\alpha_1, \ldots, \alpha_m)$

$\lambda \Leftarrow \alpha_m$
$i \Leftarrow m - 1$
while $i \geq 1$ **do**
  if $\lambda \wedge \alpha_i \not\vdash \bot$ **then** $\lambda \Leftarrow \lambda \wedge \alpha_i$ **end**
  $i \Leftarrow i - 1$
end
return λ

Function $f(\alpha_1, \ldots, \alpha_m)$

switch *case of* m **do**
  case $m = 1$ **return** α_1
  case $m = 2$
    if $\alpha_1 \wedge \alpha_2 \vdash \bot$ **then return** α_2
    else return $\alpha_1 \wedge \alpha_2$
  end
  otherwise return
  $f(\alpha_1, f(\alpha_2, \ldots, \alpha_m))$
end

The following function is trivial, as the definition of the revision function is very simple. We include the function for completeness and will not use it in what follows. Instead we use the result of the function directly. For convenience, we extend this definition to revision by a sequence $\sigma = (\varphi_1, \ldots, \varphi_n)$ of formulae with the obvious meaning. The prefix notation may look strange but should not confuse the reader.

Function $*([\rho, \blacktriangle], \varphi)$
return $[\rho \cdot \varphi, \blacktriangle]$

Function $*([\rho, \blacktriangle], \sigma)$
return $[\rho \cdot \sigma, \blacktriangle]$

Rather than the agent's belief *set*, here $Bel([\rho, \blacktriangle])$ is intended to be the formula representing that set. We use the same function name for the belief trace, but as the two functions are of different type they can easily be distinguished. Whereas the beliefs are calculated from an epistemic state, the belief trace needs an additional argument — a sequence of revision inputs.

Function $Bel([\rho, \blacktriangle])$
return $f(\rho \cdot \blacktriangle)$

Function $Bel([\rho, \blacktriangle], (\varphi_1, \ldots, \varphi_n))$
$\alpha_0 \Leftarrow Bel([\rho, \blacktriangle])$
for $i = 1$ **to** n **do** $\alpha_i \Leftarrow Bel([\rho \cdot (\varphi_1, \ldots, \varphi_i), \blacktriangle])$ **end**
return $(\alpha_0, \ldots, \alpha_n)$

The functions presented so far allow us to check whether a given epistemic state explains an observation.

Algorithm Does $[\rho, \blacktriangle]$ explain o?
Input: $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ and $[\rho, \blacktriangle]$
if $\blacktriangle \vdash \bot$ **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
$(\alpha_0, \ldots, \alpha_n) \Leftarrow Bel([\rho, \blacktriangle], (\varphi_1, \ldots, \varphi_n))$
for $i = 1$ **to** n **do**
  if $\alpha_i \nvdash \theta_i$ **then** answer $\Leftarrow$ no **end**
  foreach $\delta \in D_i$ **do**
    if $\alpha_i \vdash \delta$ **then** answer $\Leftarrow$ no **end**
  end
end
return answer

C.2 Rational closure

The following functions calculate the positive and negative conditionals given an observation o and a core belief $\blacktriangle$. In the main text, we wrote $\mathcal{C}_{\blacktriangle}(o)$ instead of $\mathcal{C}(o, \blacktriangle)$ and $\mathcal{N}_{\blacktriangle}(o)$ instead of $\mathcal{N}(o, \blacktriangle)$. For optimisations the conditionals could be labelled with their index. If several observations are to be explained using the same state, the label would also have to indicate which observation the conditional was obtained from.

Function $\mathcal{C}(\langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle, \blacktriangle)$
$\mathcal{C} \Leftarrow \emptyset$
for $i = 1$ **to** n **do** $\mathcal{C} \Leftarrow \mathcal{C} \cup \{f((\varphi_1, \ldots, \varphi_i) \cdot \blacktriangle) \Rightarrow \theta_i\}$ **end**
return $\mathcal{C}$

Function $\mathcal{N}(\langle(\varphi_1,\theta_1,D_1),\ldots,(\varphi_n,\theta_n,D_n)\rangle,\blacktriangle)$

$\mathcal{N} \Leftarrow \emptyset$
for $i = 1$ **to** n **do**
  foreach $\delta \in D_i$ **do** $\mathcal{N} \Leftarrow \mathcal{N} \cup \{f((\varphi_1,\ldots,\varphi_i)\cdot\blacktriangle) \Rightarrow \delta\}$ **end**
end
return $\mathcal{N}$

Before turning to the rational closure, we first give some auxiliary functions. `mat(`$\mathcal{C}$`)` returns the material counterparts of the conditionals in $\mathcal{C}$. In the main text they were denoted by $\tilde{\mathcal{C}}$. `pex(`$U,\mathcal{C}$`)` returns the subset of conditionals of $\mathcal{C}$ that is p-exceptional for U. `nex(`$U,\mathcal{N}$`)` returns the subset of conditionals of $\mathcal{N}$ that is n-exceptional for U.

Function `mat(`$\mathcal{C}$`)`

$\tilde{\mathcal{C}} \Leftarrow \emptyset$
foreach $\lambda \Rightarrow \mu \in \mathcal{C}$ **do** $\tilde{\mathcal{C}} \Leftarrow \tilde{\mathcal{C}} \cup \{\lambda \rightarrow \mu\}$ **end**
return $\tilde{\mathcal{C}}$

Function `pex(`$U,\mathcal{C}$`)`

$S \Leftarrow \emptyset$
foreach $\lambda \Rightarrow \mu \in \mathcal{C}$ **do**
  if $U \vdash \neg\lambda$ **then** $S \Leftarrow S \cup \{\lambda \Rightarrow \mu\}$ **end**
end
return S

Function `nex(`$U,\mathcal{N}$`)`

$S \Leftarrow \emptyset$
foreach $\lambda \Rightarrow \mu \in \mathcal{N}$ **do**
  if $U \cup \{\lambda\} \vdash \mu$ **then** $S \Leftarrow S \cup \{\lambda \Rightarrow \mu\}$ **end**
end
return S

As mentioned in the main text, the rational closure construction works for arbitrary sets of positive and negative conditionals. The following reformulation of the algorithm given in [14] is not yet optimised for dealing with conditionals obtained from an observation.[1]

Algorithm sequence $\rho_R(\mathcal{C},\mathcal{N})$ corresponding to the rational closure of $\mathcal{C}$ and $\mathcal{N}$

Input: a set of positive conditionals $\mathcal{C}$ and a set of negative conditionals $\mathcal{N}$

$\mathcal{C}_0 \Leftarrow \mathcal{C}$
$\mathcal{N}_0 \Leftarrow \mathcal{N}$
$i \Leftarrow 0$
repeat
  $U_i \Leftarrow$ `mat(`$\mathcal{C}_i$`)`
  $S \Leftarrow$ `nex(`$U_i,\mathcal{N}_i$`)`
  $T \Leftarrow \mathcal{N}_i$
  repeat
    foreach $\lambda \Rightarrow \mu \in S$ **do** $U_i \Leftarrow U_i \cup \{\neg\lambda\}$ **end**
    $T \Leftarrow T \setminus S$
    $S \Leftarrow$ `nex(`U_i,T`)`
    `/* only conditionals in` T `could have additionally become`
      `n-exceptional by enlarging` U_i `*/`
  until $S = \emptyset$
  $\mathcal{C}_{i+1} \Leftarrow$ `pex(`$U_i,\mathcal{C}_i$`)`
  $\mathcal{N}_{i+1} \Leftarrow$ `nex(`$U_i,\mathcal{N}_i$`)`
  $i \Leftarrow i + 1$
until $\mathcal{C}_i = \mathcal{C}_{i-1}$ *and* $\mathcal{N}_i = \mathcal{N}_{i-1}$
return $(\bigwedge U_{i-1},\ldots,\bigwedge U_0)$

[1] For example, when calculating $\mathcal{N}_{i+1}$ we could simply take all the negative conditionals that belonged to S in the innermost repeat-loop. The negations of their antecedents were added to U_i. Further, T would only have to to contain those negative conditionals whose index was not yet contained in S as conditionals with the same index are either all exceptional or all not exceptional.

C.3 Rational explanation

The following algorithm calculates the rational prefix $\rho_R(o, \blacktriangle)$ of an observation o with respect to a given core belief $\blacktriangle$. Abusing notation, we can extend this algorithm to a set of observations O which started in the same initial state. Note that $\rho_R(O, \blacktriangle)$ need not be equivalent to the rational prefix of any observation contained in O.

Algorithm rational prefix $\rho_R(o, \blacktriangle)$

Input: o and $\blacktriangle$

$\mathcal{C} \Leftarrow \mathcal{C}(o, \blacktriangle)$
$\mathcal{N} \Leftarrow \mathcal{N}(o, \blacktriangle)$
return $\rho_R(\mathcal{C}, \mathcal{N})$

Algorithm rational prefix $\rho_R(O, \blacktriangle)$

Input: set O of observations and $\blacktriangle$

$\mathcal{C} \Leftarrow \bigcup\{\mathcal{C}(o, \blacktriangle) \mid o \in O\}$
$\mathcal{N} \Leftarrow \bigcup\{\mathcal{N}(o, \blacktriangle) \mid o \in O\}$
return $\rho_R(\mathcal{C}, \mathcal{N})$

This allows for testing whether a given core belief is o-acceptable for a given observation or acceptable for a set of observations O starting in the same state.

Algorithm Is $\blacktriangle$ o-acceptable?

Input: observation o and core $\blacktriangle$

answer $\Leftarrow$ no
$\rho \Leftarrow \rho_R(o, \blacktriangle)$
`/* now` $\rho = (\alpha_m, \ldots, \alpha_0)$ `*/`
if $\alpha_m \equiv \top$ **then** answer $\Leftarrow$ yes **end**
return answer

Algorithm Is $\blacktriangle$ acceptable for O?

Input: set O of observations and core $\blacktriangle$

answer $\Leftarrow$ no
$\rho \Leftarrow \rho_R(O, \blacktriangle)$
`/* now` $\rho = (\alpha_m, \ldots, \alpha_0)$ `*/`
if $\alpha_m \equiv \top$ **then** answer $\Leftarrow$ yes **end**
return answer

In the following we will denote the rational explanation of an observation o by `ratexp(`o`)`. Abusing notation, we can also define the rational explanation of a set of observations O starting in the same initial state which we will denote by `ratexp(`O`)`. Again this need not be the rational explanation for any observation in o.

Algorithm rational explanation of o

Input: observation o
Output: `ratexp(`o`)`

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

Algorithm rational explanation of O

Input: set O of observations
Output: `ratexp(`O`)`

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

If additional information about the core belief is given, slight variants of the rational explanation construction may be needed. If the core is known to entail a formula λ then $\blacktriangle$ is initialised with λ rather than with $\top$. The result is an explanation using the weakest acceptable core belief entailing λ. We will denote this explanation with $\texttt{ratexp}_{\vdash}(O, \lambda)$. If the core is known to be λ then we simply complete the explanation using the rational prefix construction. This explanation will be denoted by $\texttt{ratexp}_{\equiv}(O, \lambda)$.

Algorithm explanation for O such that core entails λ

Input: set O of observations and λ
Output: $\texttt{ratexp}_{\vdash}(O, \lambda)$

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

Algorithm explanation for O such that core is λ

Input: set O of observations and $\blacktriangle$
Output: $\texttt{ratexp}_{\equiv}(O, \lambda)$

$\rho \Leftarrow \rho_R(O, \blacktriangle)$ /* $\rho = (\alpha_m, \ldots, \alpha_0)$ */
return $[\rho, \blacktriangle]$ if $\alpha_m \equiv \top$, "no such explanation" otherwise

C.4 Conclusions about $\mathcal{A}$ and hypothetical reasoning

We now turn to conclusions we draw about an agent based on an observation, first basing them on the rational explanation and later testing whether they are safe. Throughout this section we assume the observation o to have the form $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. That is, φ_i denotes the i^{th} revision input and θ_i the known beliefs after its having been received etc.

conclusions based on the rational explanation

An important question is whether a particular revision input is accepted or rejected by the agent. We give the general case for an arbitrary formula φ. If this formula is one of the inputs recorded in o then we are told whether the agent accepted or rejected that input.

Algorithm Is φ accepted or rejected?

Input: o and φ

if *o has no explanation* **then**
  answer $\Leftarrow$ o has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}(o)$
  if $\blacktriangle \vdash \neg\varphi$ **then** answer $\Leftarrow$ φ is rejected **else** answer $\Leftarrow$ φ is accepted **end**
end
return answer

The following algorithms are slightly more complicated than they need to be. We present them in this way in order to make explicit that conclusions about the beliefs of an agent at a certain point during the observation o are based on the belief trace calculated from o.

Algorithm beliefs in initial state

Input: o

if *o has no explanation* **then**
  answer $\Leftarrow$ o has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`o`)`
  $(\alpha_0, \ldots, \alpha_n) \Leftarrow Bel([\rho, \blacktriangle], (\varphi_1, \ldots, \varphi_n))$
  answer $\Leftarrow \alpha_0$
end
return answer

Algorithm beliefs after receiving φ_i

Input: o and i

if *o has no explanation* **then**
  answer $\Leftarrow$ o has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`o`)`
  $(\alpha_0, \ldots, \alpha_n) \Leftarrow Bel([\rho, \blacktriangle], (\varphi_1, \ldots, \varphi_n))$
  answer $\Leftarrow \alpha_i$
end
return answer

Apart from the beliefs during o one might also be interested in beliefs after *further* revision steps or after an *alternative* revision sequence starting in the same initial state. Note that the first is a special case of the second algorithm.

Algorithm beliefs after a further revision sequence σ

Input: o and σ

if *o has no explanation* **then**
  answer $\Leftarrow$ o has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`o`)`
  answer $\Leftarrow Bel([\rho \cdot (\varphi_1, \ldots, \varphi_n) \cdot \sigma, \blacktriangle])$
end
return answer

Algorithm beliefs after an alternative revision sequence σ

Input: o and σ

if *o has no explanation* **then**
  answer $\Leftarrow$ o has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`o`)`
  answer $\Leftarrow Bel([\rho \cdot \sigma, \blacktriangle])$
end
return answer

These questions can be answered using any explanation of o. It does not have to be the rational explanation. This is only one possible way of generating these hypotheses. We gave algorithms for the case where there is only one observation. It should be clear that the only line that would need to be modified for drawing conclusions from multiple observations is $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`o`)`. The algorithm then receives a set O of observations and uses `ratexp(`O`)` instead of `ratexp(`o`)`.

conclusions that are safe with respect to all possible explanations

These conclusions might not be safe with respect to all possible explanations (and hence with respect to the agent's actual initial state). We will present the general case where a set O of observations is given. O could be a singleton set $\{o\}$ which corresponds to the case mainly investigated in this thesis.

Note that we first have to test whether we can draw the conclusions at all. This is why the rational explanation for O is calculated first. Possibly, this is already a counterexample. If not then O is modified according to the hypothesis. We start with the question whether a certain revision input φ is accepted or rejected by $\mathcal{A}$. That formula might be one of the recorded inputs.

Algorithm Is φ (necessarily) accepted?

Input: O and φ

if O *has no explanation* **then**
  answer $\Leftarrow$ O has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`O`)`
  if $\blacktriangle \not\models \neg\varphi$ **then**
    $O' \Leftarrow O \cup \{\langle(\varphi, \top, \{\varphi\})\rangle\}$
    if O' *has explanation* **then**
      answer $\Leftarrow$ no
    else
      answer $\Leftarrow$ yes
    end
  else
    answer $\Leftarrow$ no
  end
end
return answer

Algorithm Is φ (necessarily) rejected?

Input: O and φ

if O *has no explanation* **then**
  answer $\Leftarrow$ O has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`O`)`
  if $\blacktriangle \vdash \neg\varphi$ **then**
    answer $\Leftarrow$ yes
  else
    answer $\Leftarrow$ no
  end
end
return answer

Algorithm Was θ (necessarily) believed at point $0 \leq i \leq n$ during o?

Input: O, $o \in O$, i and θ

if O *has no explanation* **then**
  answer $\Leftarrow$ O has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`O`)`
  $(\alpha_0, \ldots, \alpha_n) \Leftarrow Bel([\rho, \blacktriangle], (\varphi_1, \ldots, \varphi_n))$
  if $\alpha_i \vdash \theta$ **then**
    $\varphi_0 \Leftarrow \top$, $\theta_0 \Leftarrow \top$, $D_0 \Leftarrow \emptyset$
    $o' \Leftarrow \langle(\varphi_0, \theta_0, D_0)\rangle \cdot o$
    $D_i \Leftarrow D_i \cup \{\theta\}$
    `/*` $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$`,` i^{th} `entry in` o' `modified */`
    if $O \cup \{o'\}$ *has explanation* **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

Algorithm Was δ (necessarily) not believed at point $0 \leq i \leq n$ during o?

Input: O, $o \in O$, i and δ

if O *has no explanation* **then**
  answer $\Leftarrow$ O has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`O`)`
  $(\alpha_0, \ldots, \alpha_n) \Leftarrow Bel([\rho, \blacktriangle], (\varphi_1, \ldots, \varphi_n))$
  if $\alpha_i \not\vdash \delta$ **then**
    $\varphi_0 \Leftarrow \top$, $\theta_0 \Leftarrow \top$, $D_0 \Leftarrow \emptyset$
    $o' \Leftarrow \langle(\varphi_0, \theta_0, D_0)\rangle \cdot o$
    $\theta_i \Leftarrow \theta_i \wedge \delta$
    `/*` $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$`,` i^{th} `entry in` o' `modified */`
    if $O \cup \{o'\}$ *has explanation* **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

The next two tests deal with beliefs and non-beliefs held by the agent after an arbitrary revision sequence σ starting in the initial state. Note that these subsume the above tests by letting $\sigma = (\varphi_1, \ldots, \varphi_i)$.

Algorithm Would θ (necessarily) be believed after the revision sequence σ?

Input: O, $\sigma = (\psi_1, \ldots, \psi_m)$ and θ

if O *has no explanation* **then**
  answer $\Leftarrow$ O has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`O`)`
  $\alpha \Leftarrow Bel([\rho \cdot \sigma, \blacktriangle])$
  if $\alpha \vdash \theta$ **then**
    $o' \Leftarrow \langle(\psi_1, \top, \emptyset), \ldots, (\psi_m, \top, \{\theta\})\rangle$
    if $O \cup \{o'\}$ *has explanation* **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

Algorithm Would δ (necessarily) be not believed after the revision sequence σ?

Input: O, $\sigma = (\psi_1, \ldots, \psi_m)$ and δ

if O *has no explanation* **then**
  answer $\Leftarrow$ O has no explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`O`)`
  $\alpha \Leftarrow Bel([\rho \cdot \sigma, \blacktriangle])$
  if $\alpha \not\vdash \delta$ **then**
    $o' \Leftarrow \langle(\psi_1, \top, \emptyset), \ldots, (\psi_m, \delta, \emptyset)\rangle$
    if $O \cup \{o'\}$ *has explanation* **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

weaker notions of safety

There are cases where the above notion of safety is too strong. For example, if further information about the core belief is available, not all explanations need to be considered but only those corresponding to the additional information. In the following we will only give some examples. The other cases are analogous to the ones above. Note that σ is always an alternative revision sequence, i.e., it starts in the same initial state as all observations in O.
(i) $\mathcal{A}$'s core belief is known to entail λ.

Algorithm Would φ (necessarily) be accepted given $\blacktriangle \vdash \lambda$?

Input: O, φ, λ

if *O has no explanation such that* $\blacktriangle \vdash \lambda$ **then**
  answer $\Leftarrow$ O has no such explanation
else
  $[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}_{\vdash}(O, \lambda)$
  if $\blacktriangle \nvdash \neg\varphi$ **then**
    $O' \Leftarrow O \cup \{\langle(\varphi, \top, \{\varphi\})\rangle\}$
    if *O' has explanation such that* $\blacktriangle \vdash \lambda$ **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

Algorithm Would δ (necessarily) be not believed after the revision sequence σ given $\blacktriangle \vdash \lambda$?

Input: O, $\sigma = (\psi_1, \ldots, \psi_m)$, δ and λ

if *O has no explanation such that* $\blacktriangle \vdash \lambda$ **then**
  answer $\Leftarrow$ O has no such explanation
else
  $[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}_{\vdash}(O, \lambda)$
  $\alpha \Leftarrow Bel([\rho \cdot \sigma, \blacktriangle])$
  if $\alpha \nvdash \delta$ **then**
    $O' \Leftarrow O \cup \langle(\psi_1, \top, \emptyset), \ldots, (\psi_m, \delta, \emptyset)\rangle$
    if *O' has explanation such that* $\blacktriangle \vdash \lambda$ **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

(ii) $\mathcal{A}$'s core is known not to entail λ. Here we simply calculate rational explanation. If $\blacktriangle_{\vee}(O)$ — the core belief in `ratexp(`O`)` — does not entail λ then a counterexample is found.

Algorithm Would φ (necessarily) be rejected given $\blacktriangle \nvdash \lambda$?

Input: O, φ and λ

if *O has no explanation or* $\blacktriangle_{\vee}(O) \vdash \lambda$ **then**
  answer $\Leftarrow$ O has no such explanation
else
  $[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}(O)$
  if $\blacktriangle \vdash \neg\varphi$ **then** answer $\Leftarrow$ yes **else** answer $\Leftarrow$ no **end**
end
return answer

Algorithm Would θ (necessarily) be believed after the revision sequence σ given $\blacktriangle \nvdash \lambda$?

Input: O, $\sigma = (\psi_1, \ldots, \psi_m)$, θ and λ

if *O has no explanation or* $\blacktriangle_\vee(O) \vdash \lambda$ **then**
  answer $\Leftarrow$ O has no such explanation
else
  $[\rho, \blacktriangle] \Leftarrow$ `ratexp(`O`)`
  $\alpha \Leftarrow Bel([\rho \cdot \sigma, \blacktriangle])$
  if $\alpha \vdash \theta$ **then**
    $O' \Leftarrow O \cup \{\langle(\psi_1, \top, \emptyset), \ldots, (\psi_m, \top, \{\theta\})\rangle\}$
    if *O' has explanation such that* $\blacktriangle \nvdash \lambda$ **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

(iii) $\mathcal{A}$'s core is known to be λ. In this case we check O-acceptability of the given core.

Algorithm Would θ (necessarily) be believed after the revision sequence σ given core λ?

Input: O, $\sigma = (\psi_1, \ldots, \psi_m)$, θ and λ

if *λ is not O-acceptable* **then**
  answer $\Leftarrow$ O has no such explanation
else
  $\rho \Leftarrow \rho_R(O, \lambda)$
  $\alpha \Leftarrow Bel([\rho \cdot \sigma, \lambda])$
  if $\alpha \vdash \theta$ **then**
    $o' \Leftarrow \langle(\psi_1, \top, \emptyset), \ldots, (\psi_m, \top, \{\theta\})\rangle$
    if *λ is acceptable for $O \cup \{o'\}$* **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

Algorithm Would δ (necessarily) be not believed after the revision sequence σ given core λ?

Input: O, $\sigma = (\psi_1, \ldots, \psi_m)$, δ and λ

if *λ is not O-acceptable* **then**
  answer $\Leftarrow$ O has no such explanation
else
  $\rho \Leftarrow \rho_R(O, \lambda)$
  $\alpha \Leftarrow Bel([\rho \cdot \sigma, \lambda])$
  if $\alpha \nvdash \delta$ **then**
    $o' \Leftarrow \langle(\psi_1, \top, \emptyset), \ldots, (\psi_m, \delta, \emptyset)\rangle$
    if *λ is acceptable for $O \cup \{o'\}$* **then** answer $\Leftarrow$ no **else** answer $\Leftarrow$ yes **end**
  else
    answer $\Leftarrow$ no
  end
end
return answer

C.5 Parametrised observations

In the following, n will denote the number of unknown subformulae. $m + 1$ will be the number of recorded revision inputs in o. So $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_{m+1}, \theta_{m+1}, D_{m+1})\rangle$ will be a parametrised observation containing the unknown subformulae $\chi_1, \ldots, \chi_n$. First we

give a function that returns instantiations for the unknown subformulae. Then we use this function to define the "rational" explanation for a parametrised observation o given unknown subformulae $\chi_1, \ldots, \chi_n$. It will be denoted by $\texttt{ratexp}^p(o, (\chi_1, \ldots, \chi_n))$.

Function $\texttt{varsub}(o, (\chi_1, \ldots, \chi_n))$

Input: parametrised observation o and sequence $(\chi_1, \ldots, \chi_n)$ indicating placeholders
Output: instantiations $(x_1, \ldots, x_n)$ for placeholders

$V \Leftarrow$ variables used in o ignoring placeholders
$X \Leftarrow \{x_1, \ldots, x_n\}$ such that $V \cap \{x_1, \ldots, x_n\} = \emptyset$
return $(x_1, \ldots, x_n)$

Algorithm "rational" explanation for o

Input: parametrised observation o and sequence $(\chi_1, \ldots, \chi_n)$ indicating placeholders
Output: $\texttt{ratexp}^p(o, (\chi_1, \ldots, \chi_n))$

$(x_1, \ldots, x_n) \Leftarrow \texttt{varsub}(o, (\chi_1, \ldots, \chi_n))$
$o' \leftarrow o[(\chi_i/x_i)_i]$
$[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}(o')$
return $[\rho, \blacktriangle]$ if o' has explanation; "no explanation" otherwise

Note that this is not the explanation with the weakest core belief. But restricting conclusions to $L(o)$ (which can be determined as we know the placeholders $\chi_1, \ldots, \chi_n$), using this explanation yields the same results. One possibility for determining the weakest core is to calculate all resolvents for the CNF of $\blacktriangle$ and eliminating those clauses containing variables outside $L(o)$. The following algorithms assume that o has an explanation.

Algorithm $\blacktriangle_\vee(o)$ for the parametrised observation o

Input: parametrised observation o and sequence $(\chi_1, \ldots, \chi_n)$ indicating placeholders

$[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}^p(o, (\chi_1, \ldots, \chi_n))$
$\blacktriangle_\vee \Leftarrow \psi$ such that $Cn(\psi) = Cn(\blacktriangle) \cap L(o)$
return $\blacktriangle_\vee$

Algorithm belief trace for the parametrised observation o

Input: parametrised observation o, sequence $(\chi_1, \ldots, \chi_n)$ indicating the placeholders

$[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}^p(o, (\chi_1, \ldots, \chi_n))$
$(x_1, \ldots, x_n) \Leftarrow \texttt{varsub}(o, (\chi_1, \ldots, \chi_n))$ `/* same instantiation as for ratexp^p! */`
$(\alpha_0, \ldots, \alpha_m) \Leftarrow Bel([\rho, \blacktriangle], \iota[(\chi_i/x_i)_i])$ `/* ι is sequence of revision inputs in o */`
foreach α_i **do** $\beta_i \leftarrow \psi$ such that $Cn(\psi) = Cn(\alpha_i) \cap L(o)$ **end**
return $(\beta_0, \ldots, \beta_m)$

Note that the revision inputs in o may contain unknown subformulae. Consequently, the rational explanation for o has to use the same instantiations x_i for the unknown subformulae χ_i. The above algorithm assumes that.

Hypothetical reasoning can be done just as in the original case. The modifications corresponding to the conjecture should be made on the parametrised observation directly. This will avoid unintended interactions with the instantiations of the unknown subformulae. When reasoning based on a set of parametrised observations, one has to make sure that the same placeholders are used in different observations only if they stand for the same unknown subformulae. Otherwise the extension to sets of parametrised observations is straightforward.

C.6 Intermediate inputs

The basic case, which we will use to deal with the others, is that we are given a parametrised observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_{m+1}, \theta_{m+1}, D_{m+1})\rangle$, a sequence $(\chi_1, \ldots, \chi_n)$ indicating what the unknown subformulae are and a sequence $((1, i_1), \ldots, (m, i_m))$ indicating that after φ_j there were i_j intermediate inputs. Recall, that we should always assume the maximum possible number. We will do so by inserting i_j entries $(\chi_1^i, \top, \emptyset), \ldots, (\chi_{i_j}^i, \top, \emptyset)$ after the recorded revision input φ_j. Each χ_k^i is a *new* placeholder.

Algorithm explanation for o given the number of intermediate inputs for each position

Input: parametrised observation o, placeholders $(\chi_1, \ldots, \chi_n)$ and numbers of intermediate inputs at each position $((1, i_1), \ldots, (m, i_m))$
Output: $\texttt{ratexp}^i(o, (\chi_1, \ldots, \chi_n), ((1, i_1), \ldots, (m, i_m)))$

$o' \Leftarrow \langle(\varphi_1, \theta_i, D_1)\rangle$ `/* o' is parametrised observation incl. intermediate inputs */`
$\sigma \Leftarrow (\chi_1, \ldots, \chi_n)$ `/* σ will contain the sequence of all unknown subformulae */`
for $j = 1$ **to** m **do**
  if $i_j > 0$ **then**
    $o' \Leftarrow o' \cdot \langle(\chi_1^j, \top, \emptyset), \ldots, (\chi_{i_j}^j, \top, \emptyset)\rangle$
    $\sigma \Leftarrow \sigma \cdot (\chi_1^j, \ldots, \chi_{i_j}^i)$
  end
  $o' \Leftarrow o' \cdot \langle(\varphi_{j+1}, \theta_{j+1}, D_{j+1})\rangle$
end
$[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}^p(o', \sigma)$
return $[\rho, \blacktriangle]$ if o' has explanation; "no explanation" otherwise

Allowing any number of intermediate inputs at any time, we saw that assuming exactly one intermediate input between any two recorded ones suffices. When doing hypothetical reasoning about beliefs after a further revision step, an additional intermediate input after the last recorded one is necessary. However, the next algorithm ensures this because the additional input in hypothetical reasoning is treated like a recorded one.

Algorithm explanation for o with intermediate inputs at any time

Input: parametrised observation o and placeholders $(\chi_1, \ldots, \chi_n)$

$[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}^i(o, (\chi_1, \ldots, \chi_n), ((1, 1), \ldots, (m, 1)))$
return $[\rho, \blacktriangle]$ if o' has explanation; "no explanation" otherwise

If the positions of the intermediate inputs are known, the only problem is how many intermediate inputs should be assumed at each position. When doing hypothetical reasoning about a future input, an additional intermediate input has to be added to each position. Again, this is implicit as in this case the length of the modified observation is increased by one compared to the original observation. Recall that $m+1$ is the number of revision inputs recorded in o.

Algorithm explanation for o with intermediate inputs at given positions

Input: parametrised observation o, placeholders $(\chi_1, \ldots, \chi_n)$ and $P \subseteq \{1, \ldots, m\}$

for $j = 1$ **to** m **do**
  if $j \in P$ **then** $i_j \Leftarrow m + 1 - j$ **else** $i_j \Leftarrow 0$ **end**
end
$[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}^i(o, (\chi_1, \ldots, \chi_n), ((1, i_1), \ldots, (m, i_m)))$
return $[\rho, \blacktriangle]$ if o' has explanation; "no explanation" otherwise

If only the number l of intermediate inputs received by $\mathcal{A}$ is known — and their positions are not restricted — then the conclusions that can be drawn about the agent are least helpful. This is because all possible placements of the intermediate inputs have to be considered. Note however, that the number of intermediate inputs assumed at a certain position need not be greater than the number of recorded inputs after that position (Proposition 3.26). We will give only a construction of what $\mathcal{A}'s$ core belief must entail. The formula returned need not be o-acceptable.

Algorithm formula entailed by core belief given l intermediate inputs

Input: parametrised observation o, placeholders $(\chi_1, \ldots, \chi_n)$ and l

if $l \geq m$ **then**
  $[\rho, \blacktriangle] \Leftarrow \texttt{ratexp}^i(o, (\chi_1, \ldots, \chi_n), ((1, 1), \ldots, (m, 1)))$
  if $\exists$ *such an explanation* **then** $\blacktriangle_\vee \Leftarrow \psi$ s.t. $Cn(\psi) = Cn(\blacktriangle) \cap L(o)$ **end**
else
  $S \Leftarrow \{((1, i_1), \ldots, (m, i_m)) \mid \sum_{j=1}^{m} i_j \leq l \wedge \forall i_j : i_j \leq m + 1 - j\}$
  $T \Leftarrow \{\blacktriangle \mid [\rho, \blacktriangle] = \texttt{ratexp}^i(o, (\chi_1, \ldots, \chi_n), \sigma)$ for some $\sigma \in S\}$
  $T' \Leftarrow \{\psi \mid Cn(\psi) = Cn(\blacktriangle) \cap L(o) \wedge \blacktriangle \in T\}$
  if $T \neq \emptyset$ **then** $\blacktriangle_\vee \Leftarrow \bigvee T'$ **end**
end
return $\blacktriangle_\vee$ if there were explanations; "no such formula" otherwise

C.7 Variations

In this section, we give a selection of algorithms that might be used when considering variations of the reasoning task. We restrict ourselves to the case where there are no unknown subformulae. We do so not because the machinery is not powerful enough to deal with them but for readability. Note that we still make use of unknown subformulae and intermediate inputs for coming up with explanations.

graded observations

A graded observation has the form $o = \langle(\varphi_1, B_1, D_1), \ldots, (\varphi_{m+1}, B_{m+1}, D_{m+1})\rangle$ where B_i and D_i are finite sets of pairs (λ, k) such that λ is a formula and k a natural number indicating the reliability of λ. A regular observation is constructed with respect to a threshold t indicating the reliability the information must have.

Algorithm explanation for graded observation

Input: graded observation o, threshold t

$o' \Leftarrow \langle\rangle$
for $i = 1$ **to** $m + 1$ **do**
  $\theta'_i \Leftarrow \top$
  $D'_i \Leftarrow \emptyset$
  foreach $(\theta, k) \in B_i$ **do**
    if $k \leq t$ **then** $\theta'_i \Leftarrow \theta'_i \wedge \theta$ **end**
  end
  foreach $(\delta, k) \in D_i$ **do**
    if $k \leq t$ **then** $D'_i \Leftarrow D'_i \cup \{\delta\}$ **end**
  end
  $o' \Leftarrow o' \cdot \langle(\varphi_i, \theta'_i, D'_i)\rangle$
end
$[\rho, \blacktriangle] \Leftarrow$ `ratexp(`o'`)`
return $[\rho, \blacktriangle]$ if there is an explanation; "no explanation" otherwise

This algorithm extends naturally to sets of graded observations. When dealing with unknown subformulae, o' will be the parametrised observation containing the appropriate parts of o. Hypothetical reasoning can be done just as in the original case, but recall that conclusions need not be safe with respect to all information contained in o'.

revision inputs with priorities

In case priority information about the revision inputs is to be taken into account, we need to modify the revision function. The explanation for an observation is then calculated by simply applying the rational explanation algorithm using sets of conditionals that are calculated in a slightly different way. Observations have the following form: $o = \langle((\varphi_1, k_1), \theta_1, D_1), \ldots, ((\varphi_{m+1}, k_{m+1}), \theta_{m+1}, D_{m+1})\rangle$. The k_i indicate the position φ_i should have after revision has taken place.

Function $*_I([(\varphi_1, \ldots, \varphi_j), \blacktriangle], (\varphi, k))$

$i \Leftarrow 1$
$\sigma \Leftarrow ()$
$m \Leftarrow j$
while $i \leq k$ **do**
  if $m \geq 1$ **then**
    $\sigma \Leftarrow \varphi_m \cdot \sigma$
    $m \Leftarrow m - 1$
  else
    $\sigma \Leftarrow \top \cdot \sigma$
  end
  $i \Leftarrow i + 1$
end
$\sigma \Leftarrow \varphi \cdot \sigma$
if $m \geq 1$ **then**
  for $i = 1$ **to** m **do**
    $\sigma \Leftarrow \varphi_m \cdot \sigma$
    $m \Leftarrow m - 1$
  end
end
return $[\sigma, \blacktriangle]$

Function $\mathcal{C}_p(o, \blacktriangle, \rho_s)$

$\sigma \Leftarrow \rho_s$
$\mathcal{C}_p \Leftarrow \emptyset$
for $i = 1$ **to** $m + 1$ **do**
  $[\sigma, \blacktriangle] \Leftarrow *_I([\sigma, \blacktriangle], (\varphi_i, k_i))$
  $\mathcal{C}_p \Leftarrow \mathcal{C}_p \cup \{f(\sigma \cdot \blacktriangle) \Rightarrow \theta_i\}$
end
return $\mathcal{C}_p$

Function $\mathcal{N}_p(o, \blacktriangle, \rho_s)$

$\sigma \Leftarrow \rho_s$
$\mathcal{N}_p \Leftarrow \emptyset$
for $i = 1$ **to** $m + 1$ **do**
  $[\sigma, \blacktriangle] \Leftarrow *_I([\sigma, \blacktriangle], (\varphi_i, k_i))$
  foreach $\delta \in D_i$ **do** $\mathcal{N}_p \Leftarrow \mathcal{N}_p \cup \{f(\sigma \cdot \blacktriangle) \Rightarrow \delta\}$ **end**
end
return $\mathcal{N}_p$

Algorithm explanation for o

Input: observation o with priority information about revision inputs

$j \Leftarrow \max\{k_i - i \mid 1 \leq i \leq m + 1\}$
$\rho_s \Leftarrow (x_1, \ldots, x_j)$ such that $\{x_1, \ldots, x_j\} \cap L(o) = \emptyset$
$[\rho, \blacktriangle] \Leftarrow$ rational explanation for o using $\mathcal{C}_p(o, \blacktriangle, \rho_s)$ and $\mathcal{N}_p(o, \blacktriangle, \rho_s)$
return $[\rho \cdot \rho_s, \blacktriangle]$ if there is an explanation; "no explanation" otherwise

Here again, the explanation does not use the weakest possible core. The part of $\blacktriangle$ not talking about $L(o)$ still has to be moved to the unknown subformulae.

Algorithm belief trace based on explanation of o

Input: observation o with priority information about revision inputs

$[\rho, \blacktriangle] \Leftarrow$ explanation of o
$\alpha_0 \Leftarrow f(\rho \cdot \blacktriangle)$
$\beta_0 \Leftarrow \psi$ such that $Cn(\psi) = Cn(\alpha_0) \cap L(o)$
for $i = 1$ **to** $m + 1$ **do**
  $[\rho, \blacktriangle] \Leftarrow *_I([\rho, \blacktriangle], (\varphi_i, k_i))$
  $\alpha_i \Leftarrow f(\rho \cdot \blacktriangle)$
  $\beta_i \Leftarrow \psi$ such that $Cn(\psi) = Cn(\alpha_i) \cap L(o)$
end
return $(\beta_0, \ldots, \beta_{m+1})$ if there is an explanation; "no explanation" otherwise

core belief revision

Here, we consider observations of the following form where each $*_i$ is either $*$ or $*_\blacktriangle$: $o = \langle((\varphi_1, *_1), \theta_1, D_1), \ldots, ((\varphi_{m+1}, *_{m+1}), \theta_{m+1}, D_{m+1})\rangle$.

Function `rev`$([\rho, \rho_\blacktriangle], (\varphi, \circ))$

$\rho' \Leftarrow \rho$
$\rho'_\blacktriangle \Leftarrow \rho_\blacktriangle$
switch *case of* $\circ$ **do**
  case $\circ = *$: $\rho' \Leftarrow \rho' \cdot \varphi$
  case $\circ = *_\blacktriangle$: $\rho'_\blacktriangle \Leftarrow \rho'_\blacktriangle \cdot \varphi$
end
return $[\rho', \rho'_\blacktriangle]$

For reasoning about the agent we will give a way to calculate a possible belief trace given an observation. In the main text we illustrated a way to find an epistemic state whose sequence of revision inputs is as short as possible. Here, we initialise that sequence with a length such that an explanation is calculated in case one exists at all. A first important test is whether the observation contains non-beliefs for a point in time where the beliefs must be inconsistent due to an inconsistent core belief.

Algorithm Does o contain non-beliefs despite inconsistency?

Input: observation o indicating core and regular revision inputs

$i \Leftarrow 1$
answer $\Leftarrow$ o contains no impossible non-beliefs
while $i \leq m+1$ **do**
  if $*_i = *_\blacktriangle$ *and* $\varphi_i \equiv \bot$ **then**
    repeat
      if $D_i \neq \emptyset$ **then** answer $\Leftarrow$ o contains impossible non-beliefs **end**
      $i \Leftarrow i+1$
    until $*_i = *_\blacktriangle$ *or* $i > n$
  else
    $i \Leftarrow i+1$
  end
end
return answer

As the next step, we eliminate inconsistent core revision inputs that are recorded in o and adapt the recorded beliefs.

Algorithm observation without inconsistent core revision inputs

Input: observation o indicating core and regular revision inputs

$o' \Leftarrow \langle\rangle$
$i \Leftarrow 1$
while $i \leq m+1$ **do**
  if $*_i = *_\blacktriangle$ *and* $\varphi_i \equiv \bot$ **then**
    $i \Leftarrow i+1$
    while $*_i = *$ *and* $i \leq m+1$ **do**
      $o' \Leftarrow o' \cdot \langle(\varphi_i, \top, \emptyset)\rangle$
      $i \Leftarrow i+1$
    end
  else
    $o' \Leftarrow o' \cdot \langle(\varphi_i, \theta_i, D_i)\rangle$
    $i \Leftarrow i+1$
  end
end
return o'

For calculating an explaining epistemic state we apply the original rational explanation construction using alternative definitions for the positive and negative conditionals.

Function $\mathcal{C}_c(o, \rho_\blacktriangle)$

$\mathcal{C}_c \Leftarrow \emptyset$
$\sigma \Leftarrow ()$
$\rho'_\blacktriangle \Leftarrow \rho_\blacktriangle$
for $i = 1$ **to** $m+1$ **do**
  $[\sigma, \rho'_\blacktriangle] \Leftarrow \mathtt{rev}([\sigma, \rho_\blacktriangle], (\varphi_i, *_i))$
  $\mathcal{C}_c \Leftarrow \mathcal{C}_c \cup \{f(\sigma \cdot \rho'_\blacktriangle) \Rightarrow \theta_i\}$
end
return $\mathcal{C}_c$

Function $\mathcal{N}_c(o, \rho_\blacktriangle)$

$\mathcal{N}_c \Leftarrow \emptyset$
$\sigma \Leftarrow ()$
$\rho'_\blacktriangle \Leftarrow \rho_\blacktriangle$
for $i = 1$ **to** $m+1$ **do**
  $[\sigma, \rho'_\blacktriangle] \Leftarrow \mathtt{rev}([\sigma, \rho_\blacktriangle], (\varphi_i, *_i))$
  foreach $\delta \in D_i$ **do** $\mathcal{N}_c \Leftarrow \mathcal{N}_c \cup \{f(\sigma \cdot \rho'_\blacktriangle) \Rightarrow \delta\}$ **end**
end
return $\mathcal{N}_c$

The last algorithm we present calculates the belief trace based on the proposed explanation for an observation indicating which type of revision was triggered by an input. Note that the elements of the belief trace are calculated with respect to all revision inputs recorded in o, not just the consistent ones. As stated in the main text, it is possible that the beliefs in the initial state are inconsistent. In this case, the rational explanation construction for o' may

not be successful. So if no explanation is returned for o we can look for one for $\langle(\bot, \top, \emptyset)\rangle \cdot o$.[2] If there is no explanation for this observation then o cannot be explained at all.

Algorithm belief trace based on explanation of o

Input: observation o indicating core and regular revision inputs

if *o does not contain non-beliefs despite inconsistency* **then**
  $o' \Leftarrow o$ without inconsistent core revision inputs
  $k \Leftarrow 1 +$ number of core revision inputs in o'
  $\rho_{\blacktriangle} \Leftarrow (x_1, \ldots, x_k)$ such that $\{x_1, \ldots, x_k\} \cap L(o) = \emptyset$
  $[\rho, \blacktriangle] \Leftarrow$ rational explanation of o' using $\mathcal{C}_c(o', \rho_{\blacktriangle})$ and $\mathcal{N}_c(o', \rho_{\blacktriangle})$
  `/* ▲ is super core belief which is absorbed into the unknown subformulae,`
  `here we will not make this construction explicit */`
  if *explanation exists and* $L(o) \cap Cn(\blacktriangle) = Cn(\top)$ **then**
    $\alpha_0 \Leftarrow f(\rho \cdot \rho_{\blacktriangle} \cdot \blacktriangle)$
    $\beta_0 \Leftarrow \psi$ such that $Cn(\psi) = Cn(\alpha_0) \cap L(o)$
    for $i = 1$ **to** $m + 1$ **do**
      $[\rho, \rho_{\blacktriangle}] \Leftarrow rev([\rho, \rho_{\blacktriangle}], (\varphi_i, *_i))$
      if *last element of* $\rho_{\blacktriangle}$ *is inconsistent* **then**
        $\alpha_i \Leftarrow \bot$
      else
        $\alpha_i \Leftarrow f(\rho \cdot \rho_{\blacktriangle} \cdot \blacktriangle)$
      end
      $\beta_i \Leftarrow \psi$ such that $Cn(\psi) = Cn(\alpha_i) \cap L(o)$
    end
  end
end
return $(\beta_0, \ldots, \beta_{m+1})$ if there is an explanation; "no explanation" otherwise

[2]The same algorithm can now be used. However, as the first revision input corresponds to the initial state, β_0 will have to be eliminated from the belief trace returned and β_2 corresponds to the revision input φ_1 from o etc.