A Method for Reasoning about other Agents' Beliefs from Observations (Extended Abstract)

Alexander Nittka[1], Richard Booth[2]

[1] Department of Computer Science, University of Leipzig
Augustusplatz 10/11, 04109 Leipzig, Germany
nittka@informatik.uni-leipzig.de
[2] Faculty of Informatics, Mahasarakham University
Mahasarakham 44150, Thailand
richard.b@msu.ac.th

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

1 Introduction

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. In this paper, we want to give a brief overview over our investigations of what we can say about another agent's *actual* beliefs based on an observation of its belief revision behaviour. This issue is discussed in detail in the first author's PhD thesis [9] and in a number of (joint) papers [2,3,4,10,11].

The observed agent will be denoted by $\mathcal{A}$ and we work in a propositional language L. We make a number of simplifying assumptions. $\mathcal{A}$ employs a particular belief revision framework introduced in [1]. It can be seen as a non-prioritised version of Nayak's lexicographic revision [7] and is also closely related to Nebel's linear revision [8]. Further, we are interested only in propositional beliefs, that is, we will not deal with higher order beliefs. The information about $\mathcal{A}$ is given as an observation o containing the revision inputs received during a period of time and a partial description of beliefs held and not held after receiving an input.

We are interested in the following questions. Which inputs are accepted by $\mathcal{A}$ and which are rejected? (This need not be explicitly given in the observation.) What did $\mathcal{A}$ believe before the observation started? What did it believe during the time of observation apart from what o already tells us? What will $\mathcal{A}$ believe after receiving some further input(s)?

The general method is 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 and one of the tasks is to single out a good one.

Dagstuhl Seminar Proceedings 07351
Formal Models of Belief Change in Rational Agents
http://drops.dagstuhl.de/opus/volltexte/2007/1214

2 Belief revision framework, observation, and rational explanation

Definition 1. *The* epistemic state $[\rho, \blacktriangle]$ *of an agent consists of a sequence of formulae* ρ *and a formula* $\blacktriangle$. $\blacktriangle$ *is called the agent's* core belief. *The revision operator* $*$ *is defined for any epistemic state* $[\rho, \blacktriangle]$ *and formula* φ *by setting* $[\rho, \blacktriangle] * \varphi = [\rho \cdot \varphi, \blacktriangle]$. *The* belief set $Bel([\rho, \blacktriangle])$ *in any epistemic state* $[\rho, \blacktriangle]$ *is* $Bel([\rho, \blacktriangle]) = Cn(f(\rho \cdot \blacktriangle))$, *where*

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

The agent's epistemic state $[\rho, \blacktriangle]$ is made up of two components: *(i)* a sequence ρ of formulae and *(ii)* a single formula $\blacktriangle$, all formulae being elements of L. $\blacktriangle$ stands for the agent's set of core beliefs — the beliefs of the agent it considers "untouchable" and commits to at all times. One main effect of $\blacktriangle$ is that revision inputs contradicting it will not be accepted into the belief set. ρ is a record of the agent's revision history. Iterated revision is handled quite naturally. All revision steps are simply recorded and the problem of what $\mathcal{A}$ is to believe after each revision step, in particular whether the input just received is accepted, i.e., is believed, is deferred to the calculation of the beliefs in an epistemic state. The agent's full set of beliefs $Bel([\rho, \blacktriangle])$ in the state $[\rho, \blacktriangle]$ is determined by a particular calculation on ρ and $\blacktriangle$ which uses the function f mapping a sequence σ of propositional formulae to a formula. The agent starts with its core belief $\blacktriangle$ and then goes backwards through ρ, adding a formula as an additional conjunct if the resulting formula is consistent. If it is not, then the formula is simply ignored and the next element of ρ is considered. The belief set of $\mathcal{A}$ then is the set of logical consequences of the formula thus constructed.

Definition 2. *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* L. $[\rho, \blacktriangle]$ 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* $1 \leq i \leq n$:
   $Bel([\rho, \blacktriangle] * \varphi_1 * \dots * \varphi_i) \vdash \theta_i$ *and*
   $\forall \delta \in D_i : Bel([\rho, \blacktriangle] * \varphi_1 * \dots * \varphi_i) \nvdash \delta$

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

The intuitive interpretation of an observation is as follows. After having received the revision inputs φ_1 up to φ_i starting in some initial epistemic state, $\mathcal{A}$ believed at least θ_i but did not believe any element of D_i. Unless stated otherwise, we assume that during the time of the observation $\mathcal{A}$ received exactly the revision inputs recorded in o, in particular we assume that no input was received

between φ_i and φ_{i+1}, the observation being correct and complete in that sense. For the θ_i and D_i we assume the observation to be correct but possibly partial, i.e., the agent did indeed believe θ_i and did not believe any $\delta \in D_i$, but there may be formulae ψ for which nothing is known. 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 $D_i = \emptyset$.

An explanation of a given observation o is an epistemic state that verifies the information in o and has a consistent core belief. There are infinitely many explanations in case o can be explained. This is why our proposed method for reasoning about $\mathcal{A}$ is to choose one explanation. 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: $Bel([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = Bel([\rho, \blacktriangle] * f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$. Intuitively, the agent merges its core belief and all revision inputs received using f into a single formula and then conditions its epistemic state using it. This allows us to translate the observation into information about a single state — the initial epistemic state we are after. Note however, that $\blacktriangle$ needs to be known as otherwise $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ cannot be calculated. So, given a core belief $\blacktriangle$, o yields that $\mathcal{A}$ would believe θ_i (and would not believe any $\delta \in D_i$) in case it revised its initial epistemic state by $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$. This is nothing but conditional beliefs held and not held by $\mathcal{A}$ in its initial state $[\rho, \blacktriangle]$. That is, o is a partial description of $\mathcal{A}$'s conditional beliefs in $[\rho, \blacktriangle]$. The above equation also entails that if we had a *full* description of its conditional beliefs we could calculate the beliefs after any sequence of revision inputs, enabling us to answer the questions we posed in the introduction. It turns out that the assumed belief revision framework allows us to apply existing work ([6] and in particular [5]) on completing partial information about conditional beliefs and to construct a suitable ρ such that $[\rho, \blacktriangle]$ is indeed an explanation for o in case $\blacktriangle$ is o-acceptable. An iterative refinement of the core belief guarantees that an explanation, which we call the rational explanation, is found in case there is one.

The set of o-acceptable cores is closed under disjunction. We showed the rational explanation to satisfy a number of desirable properties, e.g., that it yields the logically weakest o-acceptable core and that the beliefs calculated satisfy a particular minimality property. However, not all conclusions drawn about $\mathcal{A}$ based on the rational explanation need to be correct. This is clear as *any other* explanation for o might correspond to the agent's true initial state. We suggested hypothetical reasoning for verifying conclusions. The basic idea is to modify the original observation o in a way such that an explanation for o', which also explains o, would be a counterexample to the conclusion we want to verify. For example, assume that the rational explanation for $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ tells us that after receiving φ_i the agent believes ψ (which is not entailed by θ_i). If we now replace the entry $(\varphi_i, \theta_i, D_i)$ in o by $(\varphi_i, \theta_i, D_i \cup \{\psi\})$, the resulting observation o' expresses that after receiving φ_i the agent does *not* believe ψ. So in case o' has an explanation — which can be tested by calculating the rational explanation of o' — the original conclusion is not safe.

3 Unknown logical content, missing revision inputs, etc.

The assumption that o contains exactly those inputs received by $\mathcal{A}$ excludes cases where the logical content of the inputs is only partially known or some inputs may have been missed by the observer. However, the original methods can be extended in order to deal with such cases as well. The idea is to allow *any* formula in o to contain unknown subformulae which are represented by place holders χ_j. The formula $p \wedge \chi_1$, for example, could represent a revision input that entails p. If the same χ_j appears in several places, this may carry information that can be exploited although the observer does not know the actual formula.

The method for dealing with this type of observation is as follows. Each χ_j is replaced by a new variable x_j, transforming an observation with unknown subformulae into a normal one. Now, the rational explanation construction can be applied. We showed that this transformation preserves explanations in the following sense. If there is an explanation for some instantiation of the χ_j then there is also one when using new variables. Conversely, if there is no explanation when using new variables x_j for the χ_j then no explanation exists using any instantiation of the unknown subformulae. Reasoning about $\mathcal{A}$ is then done by considering the rational explanation of the observation using new variables x_j but restricting the conclusions to the language constructed from those variables appearing in the observation, i.e., excluding the x_j.

This does not allow for missing inputs, yet. Note that the transformed observation is assumed to contain an entry for every input received by $\mathcal{A}$. These entries can be introduced into the observation by adding a new triple $(\chi_j, \top, \emptyset)$ for every revision input that may have been missed. This entry states that $\mathcal{A}$ received a revision input whose logical content is completely unknown and we have no information about what it believed or did not believe upon receiving that input. Of course, this methodology is problematic as we may not know how many inputs were missed at which position in the observation. However, we can provide results that limit the number of additional inputs that need to be assumed, depending on which information about number and positions of the missing inputs is available to the observer. The hypothetical reasoning methodology can still be applied.

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

References

1. Booth, R.: On the logic of iterated non-prioritised revision. In: Conditionals, Information and Inference, Springer's LNAI 3301 (2005) 86–107
2. Booth, R., Nittka, A.: Reconstructing an agent's epistemic state from observations. In: Proceedings of IJCAI'05. (2005) 394–399
3. Booth, R., Nittka, A.: Beyond the rational explanation. In: Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics. Number 05321 in Dagstuhl Seminar Proceedings (2005)
4. Booth, R., Nittka, A.: Reconstructing an agent's epistemic state from observations about its beliefs and non-beliefs. Journal of Logic and Computation (accepted for publication)
5. Booth, R., Paris, J.B.: A note on the rational closure of knowledge bases with both positive and negative knowledge. Journal of Logic, Language and Information **7** (1998) 165–190
6. Lehmann, D., Magidor, M.: What does a conditional knowledge base entail? Artificial Intelligence **55** (1992) 1–60
7. Nayak, A.: Iterated belief change based on epistemic entrenchment. Erkenntnis **41** (1994) 353–390
8. Nebel, B.: Base revision operations and schemes: Semantics, representation and complexity. In: Proceedings of ECAI'94. (1994) 342–345
9. Nittka, A.: A Method for Reasoning About Other Agents' Beliefs from Observations. PhD thesis, Leipzig University (submitted)
10. Nittka, A.: Reasoning about an agent based on its revision history with missing inputs. In: Proceedings of JELIA'06. (2006) 373–385
11. Nittka, A., Booth, R.: A method for reasoning about other agents' beliefs from observations. Texts in Logic and Games (accepted for publication)