INFORMATIK
BERICHTE

339 - 8/2007

Dynamics of Knowledge and Belief

Workshop at the 30th Annual German Conference on Artificial Intelligence, KI-2007
Osnabrück, Germany, September 10, 2007
Proceedings

Christoph Beierle, Gabriele Kern-Isberner (Eds.)

FernUniversität in Hagen

Fakultät für Mathematik und Informatik
Postfach 940
D-58084 Hagen

Christoph Beierle, Gabriele Kern-Isberner (Eds.)

Dynamics of Knowledge and Belief

Workshop at the 30th Annual German Conference
on Artificial Intelligence, KI-2007
Osnabrück, Germany, September 10, 2007
Proceedings

Workshop Organization

Workshop Organizers and Co-Chairs

Gabriele Kern-Isberner	Universität Dortmund, Germany
Christoph Beierle	FernUniversität in Hagen, Germany

Program Committee

Gerd Brewka	Universität Leipzig, Germany
James Delgrande	Simon Fraser University, Canada
Jürgen Dix	TU Clausthal-Zellerfeld, Germany
Didier Dubois	Université Paul Sabatier, Toulouse, France
Thomas Eiter	TU Wien, Austria
Esra Erdem	Sabanci University, Istanbul, Turkey
Christopher Habel	Universität Hamburg, Germany
Andreas Herzig	Université Paul Sabatier, Toulouse, France
Anthony Hunter	University College London, UK
Manfred Jaeger	Aalborg University, Denmark
Gerhard Lakemeyer	RWTH Aachen, Germany
Jerome Lang	Université Paul Sabatier, Toulouse, France
Bernhard Nebel	Albert-Ludwigs-Universität Freiburg, Germany
Torsten Schaub	Universität Potsdam, Germany
Guillermo Simari	Universidad Nacional del Sur, Bahia Blanca, Argentina
Gerhard Weiss	Software Competence Center Hagenberg, Austria

Preface

Knowledge Representation is one of the major topics in AI. Its concerns are (logical) formalisms and reasoning, with the intention to explore and model the basics of intelligent behaviour. In recent years, intelligent agents in the contexts of open environments and multi agent systems have become the leading paradigm of the field. Consequently, modern KR methods have to deal not only with static scenarios, but also with dynamic modifications in knowledge and belief, due to uncertain or incomplete information, or to changes in the environment. Moreover, agents are often expected to learn from past experiences, or to interact with other agents, making use of their knowledge and adjusting their beliefs during argumentation.

This volume contains the contributions that were presented at the Workshop *Dynamics of Knowledge and Belief* on September 10th, 2007, in Osnabrück, Germany, co-located with the 30th Annual German Conference on AI (KI-2007), and organized by the Special Interest Group on Knowledge Representation and Reasoning of the Gesellschaft für Informatik (*GI-Fachgrupppe Wissensrepräsentation und Schließen*). The particular focus of this workshop was on dynamic processes concerning any changes that an agent's state of knowledge and belief may undergo.

The first three papers use quantitative methods for knowledge representation. With their paper *From syntactical to semantical and expedient information - a survey*, Wilhelm Rödder and Elmar Reucher make a contribution to clarify the vague term "useful information" in economics and AI literature. In particular, they address issues like "value" and "price" of information, and present a study on creditworthiness. Jens Fisseler and Imre Feher make use of knowledge discovery techniques to combine data from different sources. The basic idea of their paper *A probabilistic approach to data fusion* is to generate a probabilistic rule base from each data set and to compute a joint distribution from the combined rule bases. The paper also presents a real world application with data from a telecommunication company. In the paper *On a conditional irrelevance relation for belief functions based on the operator of composition*, Radim Jirousek presents an approach how to define conditional irrelevance for belief functions via composition properties. The new composition operator is compared to Dempster's rule of combination, and relations to semigraphoids are pointed out.

Belief revision is the topic of the following papers. Haythem Ismail's paper *Reason maintenance and the Ramsey test* sheds new light on an old problem in belief revision, namely the incompatibility of handling conditionals according to the Ramsey test within the AGM framework. He proposes a theory to handle

conditionals adequately in a reason maintenance system which is based on relevance logic. With *Subjective models and multi-agent static belief revision*, Guillaume Aucher aims at generalising the famous AGM approach to multi-agent frameworks. He shows that his static belief revision operator satisfies the AGM-properties, and proposes some new postulates which are specific to the multi-agent scenario. The paper *What you should believe: Obligations and beliefs* by Guido Boella, Célia da Costa Pereira, Gabriella Pigozzi, Andrea Tettamanzi and Leendert van der Torre studies the interactions between obligations and beliefs when revising an agent's belief by new information. It is shown how obligations might help to choose between different possible options the agent has, thereby providing the logical grounds for modelling *conventional wisdom* agents.

Finally, the last two papers deal with conflicting and evolving ontologies. The paper *On the conservativity and stability of ontology-revision operators based on reinterpretation* by Özgür Özcep and Carola Eschenbach addresses the problem of resolving conflicts that are caused by agents using different ontologies in communication. The authors introduce ontology revision operators to establish consistency and encode semantic mappings between ontologies as formulas on the object level. The focus of *Dynamic T-Box-handling in agent-agent-communication* by Moritz Goeb, Peter Reiss, Bernhard Schiemann and Ulf Schreiber is on agent-agent-communication where the contents of messages are expressed in description logics. The authors study the process of merging ontologies that have been modified during communication.

We would like to thank all Program Committee members as well as the additional external reviewers Meghyn Bienvenu, Radim Jirousek, Thomas Lukasiewicz and Eric Neufeld for detailed and high-quality reviews for all submitted papers. Many thanks also to the organizers of KI-2007 for hosting the workshop at the KI-2007 conference.

August 2007 Gabriele Kern-Isberner and Christoph Beierle

Contents

Quantitative Approaches

Belief Revision

Ontologies and Description Logics

From syntactical to semantical and expedient information − a survey

Wilhelm Rödder and Elmar Reucher

University of Hagen, Germany
wilhelm.roedder@fernuni-hagen.de
elmar.reucher@fernuni-hagen.de

Abstract. In this contribution the frequently meaningless statements in the relevant economy literature, about what is knowledge and what is information, are overcome going back to the roots of information and communication theory. Information and entropy are defined precisely and then the theoretical concept is applied to an AI-model of knowledge processing. The result of this application is a powerful inference mechanism, permitting conclusions from given facts in a conditional environment. A creditworthiness problem for consumer credits demonstrates the performance of information based decision support. Here the external information factor, namely the clients' profiles, is transformed into expedient or useful information. This ability of the decision model gives rise to a deep discussion about the value and price of information.

Key words: Artificial Intelligence, Information, Inference, Knowledge, Creditworthiness.

1 Introduction

Information society has come. Economists and sociologists, among others, realize that information is the resource of the future. Hundreds of books were published recently on Knowledge Management (KM) and Information Management (IM), worldwide. Unfortunately there is no precise definition of what these concepts mean. Giving the gist of what we learned from dozens of publications: KM is managing knowledge and IM is managing information; even valid definitions of knowledge and information are missing. We quote two representative authors: "Information is expedient knowledge" [29]; "Knowledge is information in use, ..." [22]. We are confused whether knowledge is information or information is knowledge; so we hope for answers from great thinkers: "All knowledge is memory" (Hobbes); "To know what knowing and to know what doing, that is knowledge" (Confucius); "Denken ist die Erkenntnis durch Begriffe" (Kant). With the likeliest translation this reads "Thinking is knowledge or insight by concepts". All this wisdom seemingly does not create a useful definition of information and knowledge, so we should consult the exact sciences.

If there exists a precise definition of what information really is− and Information Theory provides such a definition − it could and should be a basis for a more

stringent terminology also among economists and sociologists, we feel. And perhaps Information Theory even admits a better understanding of what knowledge really is, too.

The present paper will try a cautious transfer of Information Theory to Artificial Intelligence (AI) thus permitting precise definitions within this concept. And even more: Economical or sociological problems expressable in AI-terms, might then find their respective information theoretical interpretation. This is a first step from syntactical to semantical and expedient information; an economical decision problem will show its relevance. From this first step towards Knowledge Management and Information Management will be a long and difficult way, of course.

In Section 2 Shannon's Theory of Communication and the axiomatic justification of entropy and information are sketched, in Section 3 these concepts are applied to Artificial Intelligence, Section 4 presents an information theoretical model of creditworthiness and in Section 5 the value of information is discussed. A conclusion completes this paper.

Parts of the following considerations are developed in a German publication [21], but with a different focus from the one in the present paper.

2 History of entropy and information

We met entropy for the first time in a physics lesson, when we learned that in a thermodynamic equilibrium a closed system always tends to increase entropy and that this physical magnitude measures the residual thermal energy which cannot be transformed into mechanical energy. The American engineer Claude

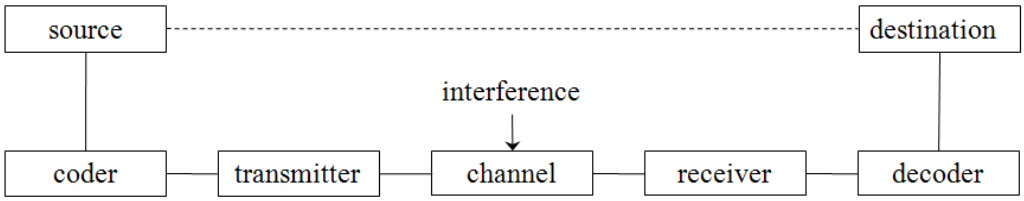


Fig. 1. A communication channel

Elwood Shannon (1916-2001) was responsible for the codification of messages between Roosevelt and Churchill during World War II and in 1948 he wrote down what he had learned about a "Mathematical Theory of Communication" [24]. The main subject of his work was to study the transmission of codified messages from a transmitter to a receiver via a channel and their decoding, see Figure 1. The channel may suffer from interference or not.

In this section we are interested in the transition of messages or symbols from a source to a destination, only, neglecting the technical part of transmitting coded signals, see again Figure 1. So we consider chains of symbols from a finite alphabet like for example a b c a a b c..., and our main purpose is to measure the average information in such a chain. There is a didactically good introduction to information and entropy written by Topsøe (1974). Following his reasoning, information is strongly related to the number of yes/no-questions necessary to eliminate uncertainty about the unknown arriving symbols. For the very special case of the alphabet $\sum = \{a, b, c\}$ and the symbols arriving with frequencies f_1, f_2, f_3, independently from each other, we develop the idea further. Assume a person A knows the arriving symbols and B does not. Then B could ask "is it a?". If the answer is "yes", the query is over. If it is "no" there could be a second question "is it b?". For either answer "yes" or "no" the query is over and the average number of questions is $f_1 \cdot 1 + f_2 \cdot 2 + f_3 \cdot 2$. Please verify that for f_1, f_2, f_3 equal 1/2, 1/4, 1/4 we get an average of 1,5 questions and for 3/4, 1/8, 1/8 it counts 1,25. Are these results the respective desired information or reduction of uncertainty? At least we doubt the result $(1\text{-}\varepsilon) \cdot 1 + \varepsilon/2 \cdot 2 + \varepsilon/2 \cdot 2 = 1 + \varepsilon$ for an arbitrarily small $\varepsilon > 0$. This would mean that the average information always exceeds 1 even for the case where the arrival of a is almost certain! Topsøe develops that building optimal queries for t-tuples of independent symbols $z_1 z_2 \ldots z_t$ from $\sum^t$ and calculating the *average* number of questions would be a more suitable approach. Then – after many definitions and lemmas, and letting t grow to infinity – he receives a result which is known as the first main theorem of information theory. We repeat it here.

Theorem 1 If for an alphabet $\sum = \{s_1, ..., s_n\}$ its elements realise in a repetitive process, each time with probabilities $P(s_1)$,...,$P(s_n)$ and independent from each other, then the average information with respect to these realisations is $H = -\sum_i P(s_i)\ ld\ P(s_i)$. This average information is equal to the average uncertainty inherent in the process. The arrival of a concrete s_i results in an information gain of $ldP(s_i)$.

In Theorem 1, ld is the logarithm with basis 2 and H is called *entropy*. Mind the fact that this entropy (as average uncertainty) and information (as average uncertainty reduction) have the same numerical value but are dual concepts.

The reader easily verifies that for $\sum = \{s_1, s_2\}$ and $P(s_1) = P(s_2) = 1/2$ the entropy H is 1; this unit has the dimension [bit]. Furthermore we see that for the above probabilities 1/2, 1/4, 1/4 we get exactly H= 1,5, whereas for 3/4, 1/8, 1/8 the estimated "information" of 1,25 differs from the exact 1,06 [bit]. For 1, 0, 0 the entropy vanishes, as we expected.

Shannon did not only study independent processes like in Theorem 1 but he also considered the intrinsic probabilistic dependency structure between the symbols in Markov chains. For this we consider a process, now generating m-words $z_1 \ldots z_m$ from $\sum^m$ rather than single symbols. This time $\sum$ might be the alphabet of the English language and each m-word a sequence of letters of length m,

including blanks e.g. Then the m-word entropy is

$$H_m = -\sum_{z_1 \ldots z_m} P(z_1 \ldots z_m) ldP(z_1 \ldots z_m),$$

being $P(z_1 \ldots z_m)$ the probabilities of such m-words. To study dependencies between letters it is necessary to look into the words. The factorization $P(z_1 \ldots z_m) = P(z_1) \cdot P(z_2|z_1) \ldots P(z_m|z_{m-1} \ldots z_1)$ into conditional probabilities permits a decomposition of H_m in accordance with (1).

$$\begin{aligned} H_m = &-\sum_{z_1} P(z_1) ldP(z_1) - \sum_{z_1} P(z_1) \sum_{z_2} P(z_2|z_1) ldP(z_2|z_1) - \ldots \\ &- \sum_{z_{m-1} \ldots z_1} P(z_{m-1} \ldots z_1) \sum_{z_m} P(z_m|z_{m-1} \ldots z_1) ldP(z_m|z_{m-1} \ldots z_1) \quad (1) \end{aligned}$$

Equation (1) often is written as $H_m = H_1 + H_{2|1} + \ldots + H_{m|m-1,\ldots,1}$, a sum of *conditioned entropies* and we have $H_1 \geqq H_{2|1} \geqq \ldots \geqq H_{m|m-1,\ldots,1}$, c.f. [15], p. 19.

If all letters in the m-words would occur with the same distribution and independently from each other, this would mean $H_m = H_1 + \ldots + H_1 = m \cdot H_1$. But in real texts of human languages, H_m is significantly smaller than $m \cdot H_1$ due to the probabilistic dependencies between the letters.

If the letters would be generated by a Markov-chain of order k, $k < m$, then the m-word entropy would become $H_m^k = H_1 + H_{2|1} + \ldots + (m-k) H_{k+1|k,\ldots,1}$, see [3], p. 97. The longer the actual memory k of the Markov-chain the higher the uncertainty reduction in the m-words, because $H_m^k \geqq H_m^l \geqq H_m$ for $l > k$.

Even for a modest $k = 2$ this reduction is significant, and in turn the symbols' dependencies are surprising. Shannon and Weaver [25], p. 54 give a nice example in which they simulate m-words given the conditional probabilities empirically collected from English texts. Such an m-word for $m = 102$ is the following:

IN NO LAT WHEY CRADICT FROUL BIR GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

The reader notices that this text does not make sense but nevertheless seems to be English. It reflects the intrinsic probabilistic conditional structure between letters. Such probabilistic conditional structures will occupy us in the Artificial Intelligence concept to be presented in Section 3.

Shannon's communication theory was pioneering, but it was a group of Russian mathematicians, which made it a mathematical theory. Jaglom and Jaglom [9], instead of focussing on communication focussed on experiments. More precisely, they studied the uncertainty about the unknown outcomes of experiments and its reduction = information when the outcomes realize.

Example 1 We make the composed experiment of first flipping a fair coin and then drawing a card from a deck of cards. The possible outcomes under consideration and their respective probabilities are

heads & red	heads & clubs	heads & spades	tails & red	tails & clubs	tails & spades
1/4	1/8	1/8	1/4	1/8	1/8

An easy calculation shows H=2,5 [bit]. As the experiment is separable – first the coin, then the card – the respective probabilities and conditioned probabilities are 1/2, 1/2 for the first part and 1/2, 1/4, 1/4 for the second. The entropies of either part we know already: 1 [bit] for the first and 1,5 [bit] for the second. So the total entropy is $H(1/2, 1/2) + 1/2\ H(1/2, 1/4, 1/4) + 1/2\ H(1/2, 1/4, 1/4) = 1 + 1/2 \cdot 1{,}5 + 1/2\ 1{,}5 = 2{,}5$ [bit].

The Russian mathematicians discovered that this separability of entropy is typical and that it is an essential property. There are four properties, which imply the form of H to be like in Theorem 1. To see this let v be possible outcomes of an experiment and $\cup v = \Omega$. Let furthermore $A_1,\ldots,A_k$ be any partition of Ω, i.e. $\cup A_i = \Omega$ and $A_i \cap A_j = \emptyset\ i \neq j$. Let P be a probability measure on Ω and ϕ a function of its probabilities.

H1 $\phi(P(v), v \in \Omega)$ is independent of the order of its arguments.

H2 ϕ is continuous in its arguments.

H3 $\phi(P(v), v \in \Omega) = \phi(P(A_1), \ldots, P(A_k)) + \sum_{i=1}^{k} P(A_i)\phi(P(v|A_i), v \in A_i).$

These three axioms reflect exactly what we expect the entropy of experiments to do. And even more: If we add

H4 $\phi(1/2, 1/2) = 1,$

then they are sufficient to determine the exact form of ϕ.

Theorem 2 If a function ϕ satisfies H1 – H4, it is necessarily of the form

$$\phi(P(v), v \in \Omega) = -\sum_{v} P(v) ld P(v).$$

The very costly proof goes back to Faddejew [4], p. 86 – 90.

This was a short story about the history of entropy and information, but do these concepts help to measure expedient knowledge and information in economic situations? This will be studied in the next sections.

3 Entropy and information in Artificial Intelligence

Shannon's model was basically dynamic in that a flow of symbols was considered. So the incoming information depends highly on the concrete symbols emitted

by the source. In a long run, however, the *average* information H looses the dynamic character and is a function of the symbols' probability distribution, only. Also Jaglom and Jalglom's experiments with unknown outcomes are of a mere static nature.

In this section we neither consider a flow of symbols nor outcomes of an experiment, but the essential pivot of the theory is a set of configurations, i.e. tuples of variables' values which describe the objects of a knowledge domain. Such configurations are similar to the m-words in the last section and as such have certain probabilities to be true in this domain. With an increasing number of such *configurations* the estimation of their probabilities becomes uncomfortable or even impossible. Therefore a language to communicate the probability distribution is necessary. The conditional structure in such a distribution then is considered the knowledge about the domain.

To develop this idea further, the reader is invited to regard a distribution in which nearly all conditioned probabilities are close to 1 or 0. If now a conditioning event becomes evident, respective conditioned events can be concluded to be nearly true or false. It must be the aim of any kind of knowledge acquisition to detect high degree dependencies between real world facts and to model them as high degree dependencies between events in a probability distribution.

So, the subject of this section is exactly this: Show how to make a system learn messages of a certain syntax, acquire knowledge and enable it to respond to questions – in an information theoretical and conditional environment. To do all these things efficiently, we need a good description of the knowledge domain, a language as communication tool between system and user, an inference mechanism to acquire knowledge and to derive facts from acquired knowledge, respectively. And we need adequate measures for the amount of acquired knowledge and for information flows. A system with this capacity consists of the following elements.

Let $V=\{V_1,\ldots,V_n\}$ be a finite set of finite-valued variables with attributes v_j of V_j. We often use mnemonic upper case names for the variables and lower case names for the attributes. A credit is GOod GO=yes/no (1/0), a client disposes of financial MEans ME=yes/no (1/0) are typical examples of variables and their respective attributes. Formulas of the type $V_j = v_j$ are literals. They are atomic propositions, which can be true (t) or false (f) under a certain interpretation. From such literals, elements of a propositional language L are formed by the junctors $\wedge$ (and), $\vee$ (or), $^-$ (not) and by parentheses; such elements are denoted by upper case letters A, B, $C\ldots$. Complete or simple conjuncts of literals we often write as unordered tuples such as $v = v_1 \ldots v_n$. V is the set of all complete conjuncts and $|\mathsf{V}|$ its cardinality. $|$ is the binary conditional operator. Formulas of the type $B|A$ are conditionals, GO=yes|ME=yes is a simple example. $B|A$ for a tautological A is equivalent to the unconditioned formula B. The set of all $B|A$ is the conditional propositional language $\mathsf{L}|\mathsf{L}$. $B|A$ is t(rue) if B and A are true, it is f(alse) for a false B and a true A, and it is undefined for a false A. So with true or false conditionals we can express (conditioned) facts about the domain,

like GO=yes|ME=yes [t] (is true), e.g. To improve the usefulness of such conditionals we allow probabilities instead of just t or f. These probabilities express the *degree* to which a conditional in the given domain is true. Such probabilistic conditionals or facts we write $B|A$ $[x]$, being x the respective probability.

If now we have a set of several such probabilistic facts or messages $\mathsf{R}=\{B_i|A_i$ $[x_i]$, $i=1,\ldots,I\}$, are they informative to the system and how to learn them?

The epistemic state of the system is a probability measure P on V with its total intrinsic probabilistic conditional structure, and this epistemic state must be built up from R.

Example 2 With the variables GO=yes/no, ME=yes/no and SUrety=yes/no we study four probability measures on V.

GO	**ME**	**SU**	P^0	P^1	P^2	P^3
yes	yes	yes	1/8	1/7	1/6	3/8
yes	yes	no	1/8	1/7	1/6	3/8
yes	no	yes	1/8	1/7	1/6	0
yes	no	no	1/8	1/7	1/6	0
no	yes	yes	1/8	0	0	0
no	yes	no	1/8	1/7	1/9	2/8
no	no	yes	1/8	1/7	1/9	0
no	no	no	1/8	1/7	1/9	0

In the first distribution, P^0, the prediction of GO= yes given any combination of ME and SU is always 0,5; the system is ignorant with respect to this question. The second distribution, P^1, knows that ME=yes∧SU=yes certainly implies GO=yes, and the third one, P^2, over and above attributes a 2/3 probability to GO=yes.

Table 1 - Epistemic states for a three-variables knowledge domain

Note that the respective entropies are 3 [bit], 2,81 [bit] and 2,78 [bit] for the three distributions P^0, P^1, P^2. The more conditional probabilistic structure in a distribution the lower entropy. Conditional structure is *knowledge* and $H(P^0) - H(P^i)$, $i = 1, 2$, measures such knowledge.

Now we explain how to put this knowledge into the system, starting from an ignorance representing uniform distribution P^0. To do so we need a function, which measures the information theoretical *distance* between two distributions. If a distribution P, for whatever reason, is changed to a distribution Q then the distance is

$$R(Q, P) = \sum_v Q(v) ld(Q(v)/P(v)).$$

R is called the *relative entropy* of Q with respect to P. R measures the overall change of conditional probabilistic structure from P to Q, for a detailed discussion confer [13], [16], [19], [26]. $K(Q||P)$, the well known Kullback−Leibler (KL) divergence is equal to $R(Q, P)$ [1]. One reason for the change from P to Q might be new messages in the form of probabilistic conditionals. So in Example 2, P^0 was first adapted to the message GO=yes|ME=yes∧SU=yes [1], yielding P^1. P^2 was the result of adapting P^0 to two facts, namely GO=yes|ME=yes∧

SU=yes [1] and GO=yes [2/3]. Mathematical calculations show that P^1 and P^2 are distributions of minimal relative entropy with respect to P^0, given the respective probabilistic conditionals, c.f. [17]. Minimizing relative entropy means best possible preserving the probabilistic structure in an epistemic state when adapting it to new messages [2], [8], [10].

There is an absolute different concept to transform the probability distribution or the epistemic state, respectively. Once we receive ad hoc knowledge about a special situation or a special *scenario*, this information will be imposed on the epistemic state only temporarily and then will be abandoned. Here again the relative entropy R is a suitable means to process this ad hoc knowledge, as shows the following example 3.

Example 3 P^2 from table 1 contains knowledge. To use this knowledge, now enter the ad hoc information that ME=yes [1]. Obviously the adaptation of P^2 to this information yields P^3, again shown in table 1; now the probability of GO=yes is 3/4. The system has concluded that under the given basic knowledge and also imposing ad hoc knowledge, an object's probability to show attribute GO=yes is 75 %. The system was never explicitly informed about this fact, the value 75 % had to be derived from P^2 and from the ad hoc knowledge. Please verify that the entropy in P^3 is 1,56 [bit].

This discussion gives rise to a general mathematical concept of knowledge acquisition, query and response, c.f. [20], [18]. First consider the knowledge acquisition:

$$P^* = \arg\min R(Q, P^0) \quad \text{s.t. } Q(B_i|A_i) = x_i, i = 1 \ldots I \tag{2}$$

P^* is the resulting distribution when adapting P^0 to all probabilistic facts $B_i|A_i$ $[x_i]$. If these facts were not valid in P^0 we get $P^* \neq P^0$, and the facts become *information* for the system. For an axiomatic justification of this concept c.f. also [12], [16], [26] and again [20]. All gathered information is *knowledge*. In Example 2 we got $P^*=P^1$ after learning one fact, and $P^*=P^2$ after learning two facts. This is what knowledge acquisition is concerned with. Now consider a query:

$$P^{**} = \arg\min R(Q, P^*) \quad \text{s.t.} \quad Q(F) = 1 \tag{3}$$

P^{**} is the resulting distribution when P^* undergoes a certain ad hoc situation F. In Example 3 we had $P^{**}=P^3$.

$$P^{**}(G), \text{ for any proposition } G \tag{4}$$

is the answer of the system to the question "How likely is G, given basic knowledge P^* and an ad hoc message F?". In Example 3 we calculated for $G \equiv$ GO=yes: $P^{**}(G) = 3/4$. $P^{**}(G)$ was *inferred* from P^* and F. This is what *inference* is concerned with.

The here described knowledge processing is sophisticated in that it has very

desirable properties [11]. So if Müller is a German and if all Germans are creditworthy then the system concludes Müller to be creditworthy, too. This is transitivity. If the system learns 80% of all Germans to be creditworthy then any *male* German, e.g., inherits this property, if no other information is available. This is called cautions monotony. If the system learns the Germans, older then 60, to be 95% creditworthy and then finds out that Germans in general only in 80% of the cases pay back their credits correctly, it nevertheless keeps its earlier conviction. This is called categorical specifity. The reader is invited to reflect these properties in view of the demands which human intellect must meet, to produce good survival strategies.

It remains to resume that $R(P^*, P^0) = H(P^0) - H(P^*)$ is the quantity of knowledge acquired by the system and $R(P^{**}, P^0) = H(P^0) - H(P^{**})$ is the knowledge amount in the situation that F is true. The respective equalities are obvious and their verification is left to the reader. All quantities measure in [bit]. If the system's knowledge increases by b [bit], it received an equal amount of information. The respective message was *informative*. Please verify that the epistemic states P^1, P^2 and P^3 in Examples 2 and 3 received 0,19, 0,22 and 1,44 [bit] of information, respectively, and hence dispose of an equal acquired amount of knowledge. The acquisition process from P^0 to P^*, the transformation of P^* to P^{**} and the entropies of the respective epistemic states are provided by an expert system shell called SPIRIT [27]. In the next section we build up a decision support model for a bank's consumer credit business, based on the hitherto developed theory.

4 Decision support for the credit business

A bank gives consumer credits to clients under certain conditions which are determined by the market rate of interest, the effective interest, and a service charge of 2%, e.g. As the market rate in EUROland is a low 2%, for an effective interest of 7% the bank realises a required rate of return of 14,66%, if the credit is paid back correctly within 4 years. For the easy calculations confer [23], p. 336-342. So this is a return of 1.466 for a 10.000 EURO credit. For a bad credit the loss is 8.614 EURO. This value comes from an estimated pay back rate lower than 20 %, which the bank usually collects only at the end of the credit's lifespan, here 4 years, and hence must be discounted at market interest. So the bank confronts the decision situation in Table 2.

	GO=yes	GO=no
LO=yes	1.466 EURO	-8.614 EURO
LO=no	-29 EURO	0 EURO

LO=yes/no stands for "loan the money or not" and GO=yes/no for a good or bad credit, like above.

Table 2 – Decision situation for the bank

The -29 EURO are estimated opportunity costs (c 2 % of 1.466 EURO of a refused credit for a good client. The recent trivial strategy of the bank was to concede all demanded credits and so for an average 88 % of good clients its average rate of return was a weak 256 EURO. Because the bank wants to go online it contracts a consultant to analyse the situation. We briefly repeat the analyst's reasoning.

A decider, absolutely uninformed about the percentage of good clients in the population, might assume a 50/50 share and this certainly favours the not loan strategy, as $0{,}5 \cdot 1.466 + 0{,}5 \cdot (-8.614) < 0{,}5 \cdot (-29) + 0{,}5 \cdot (0)$. For the observed 88 % of good credits, the respective numbers read $0{,}88 \cdot 1.466 + 0{,}12 \cdot (-8.614) > 0{,}88 \cdot (-29) + 0{,}12 \cdot (0)$, thus justifying the actual trivial loan strategy of the bank. As is well known, Laplace's daemon could predict exactly good and bad credits. Then for the 88 % of good clients he would perform an average $0{,}88 \cdot 1.466 = 1.290$ EURO return, much better than the poor 256 EURO from above. We are not Laplace's daemon, but a good prediction model might improve the decisions, too. Prediction models in the relevant literature are Scoring Models, Discriminant Analysis, Neural Networks etc. [5], [6], [7]. Here we prefer an AI-system based on the theory developed in the last section.

Index	P act	
0	0,88	GO=yes
1	0,70	SU | GO=yes
2	0,51	SU | GO=no
3	0,66	(IA ∧ KN) | GO=yes
4	0,39	(IA ∧ KN) | GO=no
5	0,10	(¬IA ∧ ¬KN) | GO=yes
6	0,35	(¬IA ∧ ¬KN) | GO=no
7	0,24	(IA ∧ ¬KN) | GO=yes
8	0,22	(IA ∧ ¬KN) | GO=no
9	0,15	(KN ∧ NB ∧ ME) | GO=yes
10	0,06	(KN ∧ NB ∧ ME) | GO=no
11	0,11	(KN ∧ ¬NB ∧ ME) | GO=yes
12	0,05	(KN ∧ ¬NB ∧ ME) | GO=no
13	0,20	(KN ∧ NB ∧ ¬ME) | GO=yes
14	0,16	(KN ∧ NB ∧ ¬ME) | GO=no
15	0,20	(KN ∧ ¬NB ∧ ¬ME) | GO=yes
16	0,16	(KN ∧ ¬NB ∧ ¬ME) | GO=no
17	0,18	(¬KN ∧ ¬NB ∧ ME) | GO=yes
18	0,21	(¬KN ∧ ¬NB ∧ ME) | GO=no
19	0,43	(IN ∧ ME) | GO=yes
20	0,34	(IN ∧ ME) | GO=no
21	0,25	(IN ∧ ¬ME) | GO=yes
22	0,19	(IN ∧ ¬ME) | GO=no
23	0,25	(¬IN ∧ ¬ME) | GO=yes
24	0,34	(¬IN ∧ ¬ME) | GO=no
25	0,59	JO | GO=yes
26	0,53	JO | GO=no
27	1,00	U=1466 | (LO=yes ∧ GO=yes)
28	1,00	U=-8614 | (LO=yes ∧ GO=no)
29	1,00	U=0 | (LO=no ∧ GO=no)
30	1,00	U=-29 | (LO=no ∧ GO=yes)

Fig. 2. Facts for the creditworthiness model

Each time a client applied for a credit we collected the following data: financial MEans available ME=yes/no, somebody offers SUrety SU=yes/no, INcome sufficient IN=yes/no, has a JOb for more than three years JO=yes/no, client is KNown to the bank KN=yes/no, No Bad earlier credits NB=yes/no, an Inquiry Agency gives a positive judgement IA= yes/no. The screenshot in Figure 2 shows the frequencies P act of the clients' property profiles for good and bad credits GO=yes/no, from 3000 historical data. It furthermore shows the decision variable LO=yes/no and an utility variable U which at any time calculates the expected monetary return, depending on the respective decision and the clients profile of attributes.

The model will now be applied to a control sample of again 3000 clients, also historical, for which we know the clients' profiles and the pay back modus GO=yes/no. Of course the bank wants to separate "good" from "bad" clients

with respect to its approximate break-even $\overline{p} = 0,852$ for which $\overline{p} \cdot 1.466 + (1 - \overline{p}) \cdot (-8.614) = \overline{p} \cdot (-29) + (1 - \overline{p}) \cdot 0$. The following Theorem 3 justifies the application of the separation mechanism derived from the first sample.

Theorem 3 Let P be any probability measure on the attribute space given in Figure 2, let m be an arbitrary profile of the clients' attributes. With $p = P(\text{GO=yes})$ we have the following proposition:
$P(\mathsf{m}|GO = yes) \gtreqless P(\mathsf{m}|GO = no)$ iff $P(GO = yes|\mathsf{m}) \gtreqless p$.

The proof of Theorem 3 is an immediate application of Bayes' theorem, because of space limits we omit it here. The theorem justifies the profiles as a separation criterion for GO=yes and GO=no for any P, and especially for such a P with $p = \overline{p}$, the break-even for the bank.

For each client from the control sample with profile m, the loan was given if $P^*(\text{GO= yes}|\mathsf{m}) > 0.852$ and was denied, otherwise. The model showed a good performance as it increased the average return from a former 256 EURO to now 515 EURO. In 1.988 cases it gave loans to clients with a good credit history and it correctly denied 297 loans for those with a bad history. The system failed 558 + 157 times, denying 558 credits for good and allowing 157 credits for bad histories. Summing over all respective returns (1.466 EURO, -8.614 EURO, -29 EURO, 0 EURO) the total return was 1.545.828 EURO. Dividing by 3000 yields 515 EURO.

That was the decision model and its performance for a 3000-person control sample, but what is the value of such a model?

5 Expedient information, its value and price

A superficial reasoning about the model's value would come to a fast conclusion: Its value is the bank's future return for an estimated number of clients and years. This mere monetary value concept suffers from a deeper theoretical justification. The value varies with the number of clients and even with the credit conditions. For changing markets and effective interests the accumulated return alters significantly.

The information theoretical concept of the system seems to be a better basis for its evaluation. There are two first information measures related to the system: The amount $R(P^*, P^0) = 3{,}55$ [bit] of knowledge acquired by adapting P^0 to all conditional facts as in Figure 2 of the last section, and the amount of the external factor information processed by the system. The external factor is the information about the attribute profiles, which clients must put at the system's disposal. Each time a client's profile is put into the system's epistemic state, entropy decreases significantly. The sum of all 3000 such information jumps amount to 16.748 [bit]. Neither measure is adequate for our purposes. In either case the consideration of absolutely irrelevant attributes with respect to the actual problem (hair colour, sex, colour of dress etc.) would cause equal or even

higher acquired knowledge and processed information, respectively, but nevertheless make the system a useless instrument. In such cases information would not be *expedient* or *useful*.

Laplace's daemon disposed of very useful knowledge, as the (hidden) attributes separated perfectly good from bad credits. There is an information theoretical function which measures this separation capacity, the *transinformation* T. The reader not familiar with this concept might study any textbook on information theory, like [14], e.g. For our purposes it is sufficient to develop that

$$T = H(P^*(GO = yes/no)) - \sum_{m} P^*(\mathsf{m}) H(P^*(GO = yes/no|\mathsf{m}))$$

is the uncertainty on GO= yes/no minus the conditioned entropy given all profiles m, T is always nonnegative. The conditioned entropy measures the average remaining uncertainty in GO in spite of known profiles. If it is high, creditworthiness does not depend on the profiles, if it is low the dependency is big. Thus in turn, T decreases with growing conditioned entropy. T is also equal to

$$\sum_{\mathsf{m}} P^*(\mathsf{m}) \cdot \sum_{GO=yes/no} R(P^*(GO|m), P^*(GO)).$$

The last expression shows T to be the weighted sum of relative entropies each of which measures the information theoretical distance of $P^*(GO|\mathsf{m})$ with respect to P^*(GO). The higher in average this distance the greater the influence of m over GO, all m. We calculated $T = 0{,}085$ [bit] for our creditworthiness example.

Laplace's daemon "explains all uncertainty about the creditworthiness away" and makes the transinformation maximum, in our case T=0,6 [bit], whereas the here built system has a performance of T=0,085 [bit] or 14,1 % of this benchmark. Each time the bank applies for a credit decision, the system transforms the external factor information of a client's attribute profile m into expedient information, $R(P^*(GO|\mathsf{m}), P^*(GO))$. A stream of applying clients generates a stream of such impacts and their weighted sum in average equals 0,085 [bit]. Expedient knowledge generates expedient information. Knowledge is a potential and does not use up.

The following Figure 3 shows the transformation from external information to expedient information. Expedient information is a precious resource, but what is a fair price? Imagine in a last step the system to attend the queries of a great number of banks, each time confronted with different credit levels, credit durations and interest rates. Each time it transmits a certain information quantity, making information a raw material or a resource for the credit business. The price of this resource is the result of the market equilibrium between supply and demand. Here the supply consists of all disposable creditworthiness prediction systems like Probabilistic Systems, Systems based on Discriminant Analysis or Neural Networks, etc. All banks, saving banks and other credit institutes demand such methods. We don't know which price will realise in such a market, of

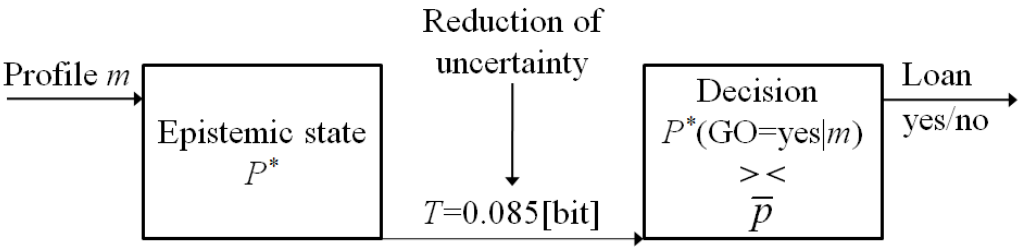


Fig. 3. From external information to expedient information

course. But this price will highly depend on each system's capability to transform external into expedient information. For the here developed AI-model this transformation process was shown to be measurable. Could this be a standard for the comparison of different prediction systems and could this even be a first step towards a more serious preoccupation with information as the most precious resource of the future?

6 Résumé

In this contribution, for the very specific situation of decision support for the consumer credit business the transformation process from incoming external information in form of the clients' profiles into expedient information concerning their creditworthiness was developed. This modern view makes information processing transparent and even measurable, thus permitting a theory-based evaluation of the raw material information. We certainly stand at the beginning of a new development, which hopefully overcomes the frequently meaningless statements concerning Information Management and Knowledge Management in recent publications.

References

1. Csiszàr, I. 1975. - I-Divergence Geometry of Probability Distributions and Minimization Problems. The Annals of Probability, Vol. 3, No.1, 148-158.
2. Cheeseman, P. (1983) - A method of computing generalized Bayesian probability values for expert systems. Proceedings IJCAI-83, Morgan Kaufmann, San Mateo, CA, 198-202.
3. Ebeling, W. & Freund, J. & Schweitzer, F. (1998) - Komplexe Strukturen: Entropie und Information. Teubner. Stuttgart.
4. Faddejew, D. K. (1961) - Arbeiten zur Informationstheorie. VEB Vol. 2, 2nd Edition. Berlin.
5. Fahrmeir, L. (2001) - Multivariate statistical modelling based on generalized linear models. Springer. New York.
6. Goldberg, D. E. (1989) - Genetic Algorithms in Search, Optimization, and Machine Learn-ing. Addison-Wesley. Reading.

7. Huberty, C. J. (1994) - Applied discriminant analysis. Wiley. New York.
8. Hunter, D. (1985) - Uncertain Reasoning using maximum entropy inference, 1st Conference on Uncertainty in Artificial Intelligence (UAI), 203-210.
9. Jaglom, A. & Jaglom, I.M. (1984) - Wahrscheinlichkeit und Information. Harri Deutsch. Frankfurt a.M.
10. Jaynes, E. T. (1983) - Papers on Probability, Statistics and Statistical Physics. Reidel, Dordrecht.
11. Kern-Isberner, G. (1997) - A logical sound method for uncertain reasoning with quantified conditionals. Proceedings 1st International Conference on Quantitative and Quantitative Practical Reasoning (ECSQARU-FAPR '97), 365-379.
12. Kern-Isberner, G. (1998) - Characterising the principle of minimum cross-entropy within a conditional-logical framework. Artificial Intelligence. Vol. 98, 169 - 208.
13. Kern-Isberner, G. (2001) - Conditionals in Nonmonotonic Reasoning and Belief Revision. Lecture Notes in Artificial Intelligence 2087. Springer. Berlin.
14. Mathar, R. (1996) - Informationstheorie: diskrete Modelle und Verfahren. Teubner, Stuttgart.
15. Meyer-Eppler, W. (1969) - Grundlagen und Anwendungen der Informationstheorie. Springer. Berlin.
16. Paris, J. B. & Vencowská (1990) - A note on the inevitability of maximum entropy. Int. J. of Approximate Reasoning 14, 183-223.
17. Rödder, W. & Meyer, C.-H. (1996) - Coherent Knowledge Processing at Maximum Entropy by SPIRIT. Proceedings of the Twelfth Conference Uncertainty in Artificial Intelligence, 470-476.
18. Rödder, W. (2000) - Conditional Logic and the Principle of Entropy. Artificial Intelligence 117, 83-106.
19. Rödder, W. (2001) - Knowledge Processing under Information Fidelity. Proc. IJCAI 2001 - Seventeenth International Joint Conference on Artificial Intelligence, 749-754.
20. Rödder, W. & Kern-Isberner, G. (2003) - From Information to Probability - An Axiomatic Approach. International Journal of Intelligent Systems 18/4, 383-403.
21. Rödder, W. & Reucher, E.(2005) - Vom menschlichen zum virtuellen Entscheider - ein Ansatz zur informationstheoretischen Leistungsbewertung; in: Mroß, M.; Thielmann-Holzmayer, C. (Eds.): Zeitgemäßes Personalmanagement, DUV - Gabler Edition Wissenschaft, Wiesbaden, 287-305.
22. Sallis, E. & Jones, D. (2002) - Knowledge Management in Education: Enhancing Learning and Education. Kogan Page Limited. London.
23. Schierenbeck, H. (1995) - Grundzüge der Betriebswirtschaftslehre. Oldenburg. München.
24. Shannon, C. E. (1948) - Mathematical Theory of Communication. The Bell System Technical Journal 27, 379-423 + 623-656.
25. Shannon, C. E. & Weaver, W. (1976) - Mathematische Grundlagen der Informationstheorie. Oldenbourg. München.
26. Shore, J. E. & Johnson, R. W. (1980) - Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross Entropy. IEEE Trans. Information Theory 26 (1), 26-37.
27. SPIRIT (2007) - http://www.xspirit.de (2007-05-31).
28. Topsoe, F. (1974) - Informationstheorie. Teuber. Stuttgart.
29. Wittmann, W. (1959) - Unternehmung und unvollständige Information. Westdeutscher Verlag. Köln.

A Probabilistic Approach to Data Fusion

Jens Fisseler[1] and Imre Fehér[1]

Department of Computer Science, FernUniversität in Hagen, 58084 Hagen, Germany,
Tel.: (+49) 2331 987-4294, Fax: (+49) 2331 987-4288,
jens.fisseler@fernuni-hagen.de, feherimi@gmail.com

Abstract. Data fusion is the process of combining data and information from two or more sources. It has its origin in market research, where it is used to combine data from different surveys. Most data fusion studies use statistical matching as their fusion algorithm, which has several drawbacks. Therefore, we propose a novel approach to data fusion, based on knowledge discovery and knowledge representation with probabilistic graphical models. We evaluate our approach on synthetic and real-world data, demonstrating its feasbility.

1 Introduction

Data fusion is the process of combining data and information from two or more sources. One of its application areas is market research, where it is used to combine data from different surveys. Ideally, one would conduct a survey with all questions of interest. But longer questionnaires lead to a lower response rate and increased bias, and also require more time and funds to plan and execute [3, 20]. Therefore, data fusion is used to combine the information gathered by two or more surveys, all of them having different questions and separate interviewee groups. In general, data fusion is a practical solution to make the information contained in readily available data sets amenable for joint analysis.

Most data fusion studies conducted in market research utilize *statistical matching* as their fusion algorithm [23]. Statistical matching uses a distance measure to find similar objects in the given data sets, which are then combined. In this paper, we propose an alternative approach to data fusion based on probabilistic models. We use a knowledge discovery algorithm to compute sets of probabilistic rules that model the dependencies between the variables in the data sets. These rule sets are then combined to build a joint probabilistic model of the data. We evaluate our approach on synthetic data and present the results of a real-world application.

The next section presents a short introduction to data fusion and further necessary background. Section 3 presents our novel data fusion process, which is evaluated in Sect. 4. Some concluding remarks are given in Section 5.

2 Background

2.1 Data Fusion

Throughout this paper, we are concerned with the problem of fusing the information contained in two data sets, $\mathcal{D}_A$ and $\mathcal{D}_B$. The statistical framework of data fusion we are concerned with [3] is based on the assumption that $\mathcal{D}_A$ and $\mathcal{D}_B$ are two samples of an unknown probability distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ over the random variabes $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$, $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ being pairwise disjoint. Furthermore, the samples of $\mathcal{D}_A$ have the values of $\boldsymbol{Y}$ missing, and the samples of $\mathcal{D}_B$ have $\boldsymbol{X}$ missing. The variables in $\boldsymbol{Z}$ are the *common variables* of $\mathcal{D}_A$ and $\mathcal{D}_B$, and data fusion assumes that the variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent given values for the variables in $\boldsymbol{Z}$

$$p(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = p(\boldsymbol{X} \mid \boldsymbol{Z}) p(\boldsymbol{Y} \mid \boldsymbol{Z}), \tag{1}$$

which is also written as $\boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$.

Most data fusion studies use *statistical matching* as their fusion algorithm [23]. Statistical matching assumes that one data set, called the *donor*, is used to provide the missing values for the other data set, the *recipient*. The statistical matching algorithm computes k best matching donor objects for every recipient object, utilizing some distance measure on the common variables. Standard measures such as the Euclidian or Hamming distance can be used, but sometimes the distance measure must be adjusted to the fusion process. For example, the common variables might contain so called *critical variables* (also called *cell* or *threshold variables*), for which the donor and recipient object must have the same values in order to be matched.

After the k best matching donor objects have been computed for the current recipient, the values for its missing variables must be calculated. This is done by summarizing the values of the k donor object for each variable missing in the recipient data set. For instance, the values of numerical variables can be summarized by their mean, whereas the values of categorical variables can be summarized by their mode.

Evaluating the quality of the data fusion is not a trivial problem [23]. Evaluation can be either internal or external, depending on the stage of the overall data analysis process at which the evaluation is performed. *Internal evaluation* takes place immediately after the data fusion and takes into account only the information available after the data fusion itself. *External evaluation* on the other hand utilizes information obtained during the other steps of the data analysis process, and thus can assess the appropriateness of the data fusion for the whole data analysis process.

Four levels of quality of a data fusion procedure can be defined [13]:

1. The marginal and joint distributions of the variables in the input data sets are preserved in the fused data.
2. In addition to the first level, the correlation structure of the variables is preserved.

3. In addition to level 2, the overall joint distribution is preserved in the fused data.
4. In addition to level 3, the true but unobserved values of all variables are preserved after fusion.

Note that levels 3 and 4 can only be validated for simulation studies where a given data set is split into several parts, providing the input for the data fusion procedure. Level 1 can always be validiated.

2.2 Probabilistic Conditional Logic

Representing and reasoning with (uncertain) knowledge is one of the main concerns of artificial intelligence. One way to represent and process uncertain knowledge is to use probabilistic methods, which, with the introduction of *probabilistic graphical models*, have seen increasing research interest during the last two decades [2, 16].

Directly representing a joint probability distribution is prohibitive for all but the smallest problems, because of the exponential growth of the memory needed for storing the distribution. Therefore, probabilistic graphical models utilize graphs to represent (in)dependencies between random variables, exploiting these to obtain a sparse representation of the joint probability distribution and to facilitate efficient reasoning. *Bayesian networks* (*BNs*) are perhaps the best known class of probabilistic graphical models. They use a directed acyclic graph to represent the dependencies between the random variables and parameterize it with a conditional probability distribution for each variable, thereby specifying a joint probability distribution. Despite their wide-spread use, BNs have some drawbacks. Their directed acyclic structure prohibits the representation of certain cyclic or mutual dependencies, and they require the specification of many (conditional) probabilities, which is especially troublesome in case a Bayesian network is constructed by an expert and not learned from data.

Another way to construct a probabilistic graphical model is to specify certain constraints, and compute an appropriate joint probability distribution that satisfies these constraints. In principle, there are many joint probability distributions satisfying a given set of constraints, but in order to make meaningful inferences one must choose a single "best" model. The *principle of maximum entropy* states that, of all the distributions satisfying the given constraints, one should choose the one with the largest entropy[1], because it is the least unbiased, the one with "maximum uncertainty" with respect to missing information. It can also be shown that the principle of maximum entropy is the unique correct method of inductive inference satisfying intuitive, commonsense requirements [15, 21].

The *probabilistic conditional logic* (*PCL*) [19] is a formalism to represent constraints on a joint probability distribution. Assume we are given a set $\boldsymbol{U} = \{V_1, \ldots, V_k\}$ of random variables V_i, each with a finite range $\mathcal{V}_i$. The atoms of

[1] The *entropy* of a discrete probability distribution P with sample space Ω is defined as $H(P) := -\sum_{\omega \in \Omega} p(\omega) \log_2 p(\omega)$.

PCL are of the form $V_i = v_i$, depicting that random variable V_i has taken the value $v_i \in \mathcal{V}_i$, and formulas are constructed using the usual logical connectives $\neg$, $\vee$, $\wedge$. The constraints expressible with PCL are *probabilistic facts* $\rho[x]$ and *probabilistic rules* $(\psi \mid \phi)[x]$, where ρ, ψ and ϕ are formulas built from literals and the usual logical connectives, and x is in $[0, 1]$ [2]. A probability distribution P with sample space $\Omega = \prod_{i=1}^{k} \mathcal{V}_i$ *represents* a probabilistic rule $(\psi \mid \phi)[x]$, written $P \models (\psi \mid \phi)[x]$, iff $p(\phi) > 0$ and $p(\phi \wedge \psi) = x \cdot p(\phi)$; it represents a set $\mathcal{R}$ of probabilistic rules, written $P \models \mathcal{R}$, iff it represents each probabilistic rule in $\mathcal{R}$.

The maximum entropy distribution $ME(\mathcal{R}) = P^* := \operatorname{argmax}_{P \models \mathcal{R}} H(P)$ representing a set of probabilistic rules $\mathcal{R} = \{(\phi_1 \mid \psi_1)[x_1], \ldots, (\phi_m \mid \psi_m)[x_m]\}$ can be depicted as

$$p^*(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_{j=1}^{m} \lambda_j f_j(\boldsymbol{x})\right) = \frac{1}{Z} \prod_{j=1}^{m} e^{\lambda_j f_j(\boldsymbol{x})}, \tag{2}$$

where $\lambda_j \in \mathbb{R}^{>0}$ are the weights depicting the influence of the *feature functions* $f_j : \Omega \to [-x_j, 1 - x_j]$, one feature function for each probabilistic rule, and $Z := \sum_{\boldsymbol{x} \in \Omega} \exp\left(\sum_{j=1}^{m} \lambda_j f_j(\boldsymbol{x})\right)$ is a normalization constant. Equation 2 is the log-linear model notation for *Markov networks*, so PCL is a formalism for specifying Markov networks, i.e. *undirected* graphical models [16]. The expert system shell SPIRIT [14,17] is based on PCL, enabling the user to enter a set of probabilistic rules and facts and to efficiently answer probabilistic queries, similar to an expert system shell using Bayesian networks.

2.3 Learning Probabilistic Rules from Data

The expert system shell SPIRIT, introduced in the previous section, can be used to build a probabilistic knowledge base by specifying a set of probabilistic rules. I.e., we are given a set of probabilistic rules and utilize the principle of maximum entropy to compute a probability distribution. Now assume we would – instead of a set of rules – be given an (empirical) probability distribution. Then we could ask the question which set of probabilistic rules would yield this probability distribution, again utilizing the principle of maximum entropy.

Computing probabilistic rules from empirical probability distributions is an example of *knowledge discovery from databases* (*KDD*) [4]. KDD is an interdisciplinary area of research, drawing – amongst others – on methods from machine learning, databases, statistics and knowledge aquisition for expert systems. As the interest on research in probabilistic graphical models and knowledge discovery in databases started to grow almost simultaneously, a lot of work has been done on developing methods for learning probabilistic models from data, especially Bayesian networks [9].

The development of a method for computing probabilistic rules from empirical data was based on the idea that, using the discovered rules, the maximum

[2] Probabilistic facts $\rho[x]$ can also be represented as $(\rho \mid \top)[x]$.

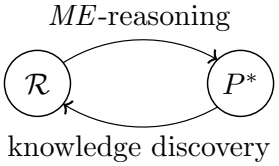


Fig. 1. Knowledge discovery by reversing knowledge representation

entropy approach to constructing a probabilistic model described in Sect. 2.2 should give the same probability distribution as the one the empirical data was sampled from. This way, knowledge discovery can be interpreted as being inverse to knowledge representation and reasoning, see Fig. 1.

Suppose we are given a set $\boldsymbol{U} = \{V_1, \ldots, V_k\}$ of random variables V_i, each with a finite range $\mathcal{V}_i$ (cf. Sect. 2.2). Given a sample $\tilde{P}$ of a probability distribution over $\boldsymbol{U}$ with sample space $\Omega = \prod_{i=1}^{k} \mathcal{V}_i$, we want to compute a set of probabilistic rules $\mathcal{R}$ such that $\tilde{P}$ is a sample of $ME(R)$. We assume that $\mathcal{R}$ consists of so-called *single-elementary rules*, i.e. rules with a single atom as their conclusion.

Recalling Equation 2, one can see that each rule $R \in \mathcal{R}$ has a certain amount of influence on the probability of every event of the *ME*-probability distribution $ME(R)$. By associating abstract symbols with this influence one can build a theory of conditional structures [11] which can be used to disentangle the complex joint influence of probabilistic rules. Using this theory, one can search for certain numerical relationships in an empirical probability distribution $\tilde{P}$ and use these relationships to compute a set of rules which is able to model $\tilde{P}$ by Equation 2. Details on this knowledge discovery approach are given in [11, 12]; here we will only give a brief overview of the algorithm.

- The algorithm starts with a set of single-elementary rules. In principle, one would choose each of the $\prod_{i=1}^{k} |\mathcal{V}_i|$ literals as the head, and all possible combinations of the literals of the remaining $k-1$ variables as the conclusions. As this results in an exponential number of rules, the events $\omega \in \Omega$ with zero probability are used to reduce the initial rule set.
- During the second step the algorithm searches for numerical relationships that can be used to reduce the initial set of rules. These relationships are depicted by even-length cycles in an undirected graph induced by Ω. As there is an exponential number of such cycles, a length restriction has to be imposed.
- After the numerical relationships have been computed, they are used to reduce the set of rules. This is done by conjoining or removing rules until all numerical relationships have been taken into account.

This algorithm has been implemented in CondorCKD [6, 7], which is a part of the larger Condor system [1]. CondorCKD can be used to compute the set of ME-optimal rules for given empirical probability distributions, which are assumed to reside in tabular form as CSV or ARFF files. The resulting rules

are interesting in themselves, but can also be used to construct a probabilistic model with SPIRIT, see Sect. 2.2 and the following sections.

3 Data Fusion with CondorCKD and SPIRIT

Given two empirical probability distributions $\tilde{P}_A(\boldsymbol{X}, \boldsymbol{Z})$ and $\tilde{P}_B(\boldsymbol{Y}, \boldsymbol{Z})$[3], we can use CondorCKD to learn sets of probabilistic rules $\mathcal{R}_A$ and $\mathcal{R}_B$ that are models for $\tilde{P}_A$ resp. $\tilde{P}_B$. Combining these rule sets yields a probabilistic model – a Markov network, cf. Sect. 2.2 – for the unknown joint probability distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. This Markov network has a corresponding graph structure $G = (\boldsymbol{U}, E)$ with one node for every variable in $\boldsymbol{U} = \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$. Two nodes S, T in $\boldsymbol{U}$, $S \neq T$, are connected by an edge iff there is at least one rule in $\mathcal{R}_A \cup \mathcal{R}_B$ that contains both S and T [14]. Because $\mathcal{R}_A$ and $\mathcal{R}_B$ are computed from $\tilde{P}_A(\boldsymbol{X}, \boldsymbol{Z})$ and $\tilde{P}_B(\boldsymbol{Y}, \boldsymbol{Z})$, G will contain no edges between variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$. The only way two variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ might be connected is by a path going through $\boldsymbol{Z}$, i.e. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *graphically separated by* $\boldsymbol{Z}$ *in* G, written as

$$\boldsymbol{X} \perp\!\!\!\perp_G \boldsymbol{Y} \mid \boldsymbol{Z}. \tag{3}$$

For Markov networks, Equation 3 implies the conditional independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ [16], written as

$$\boldsymbol{X} \perp\!\!\!\perp_G \boldsymbol{Y} \mid \boldsymbol{Z} \quad \Rightarrow \quad \boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y} \mid \boldsymbol{Z}.$$

Thus, if the conditional independence assumption (see Equation 1) is valid for two given data sets $\mathcal{D}_A$ and $\mathcal{D}_B$, constructing a probabilistic graphical model with CondorCKD and SPIRIT gives an adequate model for the unknown joint probability distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. The quality of this model depends on the conditional independence of $\tilde{P}_A(\boldsymbol{X}, \boldsymbol{Z})$ and $\tilde{P}_B(\boldsymbol{Y}, \boldsymbol{Z})$ given $\boldsymbol{Z}$. For a known joint probability distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ the conditional independence of the marginal distributions $P(\boldsymbol{X}, \boldsymbol{Z})$ and $P(\boldsymbol{Y}, \boldsymbol{Z})$ given $\boldsymbol{Z}$ can be measured with the *conditional mutual information*[4] [8]

$$\begin{aligned} I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) &= \mathbb{E}_P \left(\log_2 \frac{p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z})}{p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{y} \mid \boldsymbol{z})} \right) \\ &= H(\boldsymbol{X}, \boldsymbol{Z}) + H(\boldsymbol{Y}, \boldsymbol{Z}) - H(\boldsymbol{Z}) - H(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}), \end{aligned} \tag{4}$$

which is zero iff $\boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$.

$I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$ can of course only be calculated when the joint probability distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is known, which in general is not the case with data fusion problems. But the conditional mutual information can be used in experiments with a *known* joint distribution to assess the quality and validity of our proposed data fusion process.

Summarizing, our data fusion process consists of the following steps:

[3] These are defined by computing the relative frequencies of the different joint events or objects in $\mathcal{D}_A$ resp. $\mathcal{D}_B$.

[4] Note that $\mathbb{E}_P(f)$ denotes the *expected value* of function f with respect to the probability distribution P.

1. Compute two rule sets $\mathcal{R}_A$ and $\mathcal{R}_B$ for the given input data sets $\mathcal{D}_A$ and $\mathcal{D}_B$, using CondorCKD.
2. Build a model for the joint data by constructing a *ME*-probability distribution with SPIRIT, using $\mathcal{R}_A \cup \mathcal{R}_B$ as input.
3. Evaluate the quality of the data fusion process, using (at least) level 1 validation.

4 Experiments

In order to verify our data fusion process, we first use a synthetic data set and various partitionings of its variables to verify whether low conditional mutual information results in a higher quality of the data fusion. After that we demonstrate the applicability of our approach by fusing two real-world data sets.

4.1 Fusing synthetic data sets

We use a variant of the well-known *Léa Sombé* example [18, 22] as our synthetic data set. It has six binary variables, describing people in a fictional community:

S : Being a student
Y : Being young
G : Being single
P : Being a parent
M: Being married
C : Cohabiting

The dependencies between these attributes can be expressed by six rules:

(R1) 90% of all students are young: $(Y \mid S)[0.9]$
(R2) 80% of all young people are single, $(G \mid Y)[0.8]$
(R3) 70% of all single people are young, $(Y \mid G)[0.7]$
(R4) 30% of the young people are students, $(S \mid Y)[0.3]$
(R5) 90% of all students with children are married, $(M \mid S \wedge P)[0.9]$
(R6) 80% of the cohabiting people are young, $(Y \mid C)[0.8]$

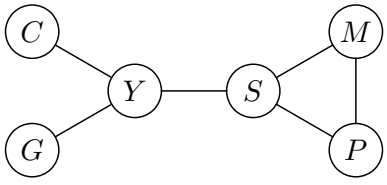


Fig. 2. Markov network of the *Léa Sombé* example

Using these rules we build a probabilistic model of the *Léa Sombé* example, whose Markov network is depicted in Fig. 2. In order to obtain a data set that

can be processed by CondorCKD we generate a sample with 50,000 elements. As we need two data sets for data fusion we have to partition the attribute set $\{S, Y, G, P, M, C\}$ in two non-disjoint sets and generate appropriate marginal samples.

Examining the probabilistic rules and the resulting Markov network it can be seen that there are two almost independent sets of attributes, $\{S, P, M\}$ and $\{Y, G, C\}$. These attribute sets are made dependent by the rules (R1) and (R4), which can be verified by inspecting the Markov network depicted in Fig. 2, where the attributes $\{M, P\}$ are graphically separated from $\{G, C\}$ by $\{S, Y\}$.

Based on these observations, we perform two kinds of partitionings. For the first kind of partitionings we select one of the six attributes as the common variable. Starting from the two sets of attributes, $\{S, P, M\}$ and $\{Y, G, C\}$, this results in six different partitionings, with the common variable being adjoined to the attribute set of which it is not an element. For the second kind of partitionings we select two attributes as the common variables, one from $\{S, P, M\}$ and one from $\{Y, G, C\}$, which results in nine different partitionings. As with the first kind of partitionings we adjoin each of the common variables to the attribute set of which it is not an element.

Table 1. Results of experiments with known joint probability distribution.

$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Z}$	$I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$	$D_{KL}(P \parallel P_{orig})$
$\{S, Y, P, M\}$	$\{Y, G, M, C\}$	$\{Y, M\}$	0.00039963	0.00100968
$\{S, Y, P, M\}$	$\{S, Y, G, C\}$	$\{S, Y\}$	0.00034244	0.00104120
$\{S, Y, P, M\}$	$\{Y, G, P, C\}$	$\{Y, P\}$	0.00041526	0.00130367
$\{S, G, P, M\}$	$\{S, Y, G, C\}$	$\{S, G\}$	0.00029772	0.00131082
$\{S, P, M\}$	$\{S, Y, G, C\}$	$\{S\}$	0.00040399	0.00136513
$\{S, Y, P, M\}$	$\{Y, G, C\}$	$\{Y\}$	0.00046692	0.00205626
$\{S, P, M, C\}$	$\{Y, G, P, C\}$	$\{P, C\}$	0.06343040	0.07862572
$\{S, P, M, C\}$	$\{Y, G, M, C\}$	$\{S, M\}$	0.06294772	0.08167080
$\{S, P, M, C\}$	$\{Y, G, C\}$	$\{C\}$	0.06385933	0.08491127
$\{S, P, M\}$	$\{Y, G, M, C\}$	$\{M\}$	0.06774234	0.08557503
$\{S, P, M\}$	$\{Y, G, P, C\}$	$\{P\}$	0.06824536	0.08558482
$\{S, G, P, M\}$	$\{Y, G, M, C\}$	$\{G, M\}$	0.06570083	0.08708857
$\{S, G, P, M\}$	$\{Y, G, C\}$	$\{G\}$	0.06662601	0.08939546
$\{S, G, P, M\}$	$\{Y, G, P, C\}$	$\{G, P\}$	0.06623336	0.09087545

Table 1 shows the 15 different partitionings and the results of the evaluation of the data fusion process. Columns "$\boldsymbol{X}$" and "$\boldsymbol{Y}$" depict the attributes of the two input data sets and column "$\boldsymbol{Z}$" shows their common variables. The fourth column, "$I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$", shows the conditional mutual information of $P(\boldsymbol{X}, \boldsymbol{Z})$ and $P(\boldsymbol{Y}, \boldsymbol{Z})$ given $\boldsymbol{Z}$, which we argue to be an indicator of the qualitify of the data fusion. The last column, "$D_{KL}(P \parallel P_{orig})$", shows the information diver-

gence[5] of the fused data set P and the original data P_{orig}. Note that Table 1 is sorted by the last column.

As can be seen, the 15 different partitionings can be grouped in two categories, based on the conditional mutual information and indicated by the horizontal rule in Table 1. The conditional mutual information of the partitionings in the first category is more than two orders of magnitude smaller of those partitionings in the second category. The information divergence between the fused data sets in the first category and the original data set is also more than one order of magnitude smaller than the information divergence of the fused data sets in the second category, as shown in the fifth column.

Because $\{S, Y\}$ graphically separates $\{M, P\}$ and $\{G, C\}$, cf. Fig. 2, partitioning into $\{S, Y, P, M\}$ and $\{S, Y, G, C\}$ should give the best data fusion result. Although it is only second best with respect to $D_{KL}(P \parallel P_{orig})$, which is probably due to sampling errors, it can be clearly seen that a low conditional mutual information results in good data fusion quality. I.e., our data fusion process gives good results in case the conditional independence assumption of data fusion, see Equation 1, is fulfilled.

4.2 Fusing real-world data

After evaluating our data fusion process on synthetic data, we now apply it on two real-world data sets. These originate from a Hungarian telecommunication company and are called $\mathcal{D}_{int}$ and $\mathcal{D}_{ext}$. $\mathcal{D}_{int}$ is a sample (sample size 10,000) from the data warehouse of the telecommunication company and contains the following variables:

Internet access? Binary variable, depicting whether the customer's household has internet access.

Minutes outbound Numerical variable, depicting the monthly minutes of outbound (landline) calls.

Minutes inbound Numerical variable, depicting the monthly minutes of inbound (landline) calls.

Invoice total Numerical variable, depicting the monthly invoice total.

Alternative carrier? Binary variable, depicting whether the customer has used an alternative telephone provider.

$\mathcal{D}_{ext}$ is the result of a survey among a representative sample (sample size 3,140) of all company customers. It contains the following variables:

Internet access? Binary variable, same as in $\mathcal{D}_{int}$.

Minutes outbound Numerical variable, same as in $\mathcal{D}_{int}$.

[5] The *information divergence*, also called *relative entropy* or*Kullback-Leibler divergence*, of two probability distributions P and Q with common sample space Ω is defined as $D_{KL}(P, Q) := \sum_{\omega \in \Omega} p(\omega) \log_2 \frac{p(\omega)}{q(\omega)}$. It measures the difference of information from P to Q, where Q is assumed to be the "true" probability distribution, and is zero iff P equals Q [8, 10].

Mobile phone? Binary variable, depicting whether any person in the customer's household owns a mobile.

Cable television? Binary variable, depicting whether the customer's household has cable TV.

ESOMAR Categorical variable, representing the socio-economic group of the customer according to ESOMAR criteria[6].

Table 2. Discretization criteria for the variables "Minutes inbound", "Minutes outbound" and "Invoice total"

Minutes inbound		*Minutes outbound*		*Invoice total*	
Range (min.)	*Category*	*Range* (min.)	*Category*	*Range* (HUF)	*Category*
$[0, 90)$	few	$[0, 130)$	few	$[0, 3600)$	small
$[90, 200)$	average	$[130, 250)$	average	$[3600, 4800)$	average
$[200, \infty)$	many	$[250, \infty)$	many	$[4800, \infty)$	large

$\mathcal{D}_{int}$ and $\mathcal{D}_{ext}$ shall be fused in order to plan and execute certain marketing activities for which information from both sources is needed. Budget and/or time constraints sometimes prohibit the collection of a larger survey, so this kind of data fusion is quite common and is used to enrich readily availabe data with information more costly to aquire.

As both data sets contain several numerical variables, these must be discretized first. This is done according to the criteria shown in Table 2.

Because there is no joint data set available, we cannot measure the conditional mutual information in order to get an indicator what quality can be expected of the data fusion. For larger, business critical data fusion projects one would initially conduct a survey containing *all* variables of interest in order to assess the validity of the conditional independence assumption. For our data sets we can only evaluate how well they agree on their common variables. Because neither data set is the "original" data, we compute both information divergences for $\mathcal{D}_{int}$ and $\mathcal{D}_{ext}$, marginalized on their common variables "Internet access?" and "Minutes outbound":

$$D_{KL}(\tilde{P}_{int} \,\|\, \tilde{P}_{ext}) = 0.0000008162553$$
$$D_{KL}(\tilde{P}_{ext} \,\|\, \tilde{P}_{int}) = 0.0000008164547$$

The information divergence between both data sets (marginalized on their common variables, "Internet access?" and "Minutes outbound") is negligible, although the $\mathcal{D}_{ext}$ is 9 months older and thus there might be some changes in customer behaviour represented in $\mathcal{D}_{int}$.

[6] `http://www.esomar.org/`

The data fusion process is the same as described in Sect. 3. We compute two sets of probabilistic rules $\mathcal{R}_{int}$ and $\mathcal{R}_{ext}$ from the given data sets, and combine these rule sets with SPIRIT. Validation of the data fusion process can only be done according to level 1 as described in Sect. 2.1, i.e. we can only evaluate how well the marginals of the joint model agree with the original data sets. For this purpose, we generate a sample (sample size 10,000) from the joint probabilistic model and compute the information divergences between the appropriate marginals and the original input data sets:

$$D_{KL}(P \parallel \tilde{P}_{int}) = 0.2704161$$
$$D_{KL}(P \parallel \tilde{P}_{ext}) = 0.125255$$

Although the information divergence between the common variables of $\mathcal{D}_{int}$ and $\mathcal{D}_{ext}$ is very small, the joint model has an information divergence larger than might be expected, at least in comparison to the results with the synthetic data. Part of this deviation is due to the fusion of the two probabilistic rule sets, as can be seen by computing the relative entropy between the reconstructed data sets and the original input data sets:

$$D_{KL}(P_{\mathcal{R}_{int}} \parallel \tilde{P}_{int}) = 0.05815166$$
$$D_{KL}(P_{\mathcal{R}_{ext}} \parallel \tilde{P}_{ext}) = 0.004139472$$

Further work is necessary to investigate what aspects of the data fusion process contribute to the larger deviation of the joint model from the original data sets, and whether this is due to the data fusion itself or because of other factors like improper common variables.

5 Conclusions and Further Work

We have presented a novel approach to data fusion which is based on probabilistic models. Using the knowledge discovery software CondorCKD, we compute probabilistic rules sets from the given input data sets, which are then combined to build a joint model for the data sets, using the expert system shell SPIRIT. We have evaluted our data fusion process on artificial and real-world data.

Whereas experiments with synthetic data sets yielded promising results, the evaluation of the data fusion experiments with real-world data sets was somewhat inconclusive. Therefore, further work is necessary to evaluate the proposed approach on more real-world data sets and compare its results to those of other data fusion algorithms.

Nevertheless, our data fusion process is interesting and useful in itself, as it offers an alternative approach to fuse data sets containing non-numeric, i.e. categorical, variables.

References

[1] Christoph Beierle and Gabriele Kern-Isberner. Modelling conditional knowledge discovery and belief revision by abstract state machines. In Egon Börger, Angelo

Gargantini, and Elvinia Riccobene, editors, *Abstract State Machines, Advances in Theory and Practice, 10th International Workshop, ASM 2003*, volume 2589 of *LNCS*, pages 186–203. Springer, 2003.

[2] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, 1999.

[3] Marcello D'Orazio, Marco Di Zio, and Mauro Scanu. *Statistical Matching: Theory and Practice*, chapter The Statistical Matching Problem. Wiley Series in Survey Methodology. Wiley, 2006.

[4] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press, 1996.

[5] Imre Fehér. Datenfusion von Datensätzen mit nominalen Variablen mit Hilfe von CONDOR und SPIRIT. Master's thesis, Department of Computer Science, FernUniversität in Hagen, Germany, 2007.

[6] J. Fisseler, G. Kern-Isberner, C. Beierle, A. Koch, and C. Müller. Algebraic knowledge discovery using Haskell. In *Practical Aspects of Declarative Languages, 9th International Symposium, PADL 2007*, Lecture Notes in Computer Science, Berlin, Heidelberg, New York, 2007. Springer. (to appear).

[7] Jens Fisseler, Gabriele Kern-Isberner, and Christoph Beierle. Learning uncertain rules with CONDORCKD. In David C. Wilson and Geoffrey C. J. Sutcliffe, editors, *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*. AAAI Press, 2007.

[8] Robert M. Gray. *Entropy and Information*. Springer, 1990.

[9] Michael I. Joran, editor. *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

[10] J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, 1992.

[11] Gabriele Kern-Isberner. Solving the inverse representation problem. In Werner Horn, editor, *ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence*, pages 581–585. IOS Press, 2000.

[12] Gabriele Kern-Isberner and Jens Fisseler. Knowledge discovery by reversing inductive knowledge representation. In D. Dubois, Ch. A. Welty, and M.-A. Williams, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004)*, pages 34–44. AAAI Press, 2004.

[13] Hans Kiesl and Susanne Rässler. How valid can data fusion be? Technical Report 200615, Institut für Arbeitsmarkt und Berufsforschung (IAB), Nürnberg (Institute for Employment Research, Nuremberg, Germany), 2006. Online available at `http://ideas.repec.org/p/iab/iabdpa/200615.html`.

[14] Carl-Heinz Meyer. *Korrektes Schließen bei unvollständiger Information*. PhD thesis, Department of Business Science, FernUniversität in Hagen, Germany, 1998.

[15] J. B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:183–223, 1990.

[16] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[17] W. Rödder, E. Reucher, and F. Kulmann. Features of the expert-system-shell SPIRIT. *Logic Journal of IGPL*, 14(3):483–500, 2006.

[18] Wilhelm Rödder and Gabriele Kern-Isberner. Léa sombé und entropie-optimale Informationsverarbeitung mit der Expertensystem-Shell SPIRIT. *OR Spektrum*, 19:41–46, 1997.

[19] Wilhelm Rödder and Gabriele Kern-Isberner. Representation and extraction of information by probabilistic logic. *Information Systems*, 21(8):637–652, 1997.

[20] Gilbert Saporta. Data fusion and data grafting. *Computational Statistics & Data Analysis*, 38:465–473, 2002.

[21] John E. Shore and Rodney W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.

[22] Léa Sombé. *Reasoning Under Incomplete Information in Artificial Intelligence: A Comparison of Formalisms Using a Single Example*. Wiley, 1990. Special Issue of the *International Journal of Intelligent Systems*, 5(4).

[23] Peter van der Putten, Joost N. Kok, and Amar Gupta. Why the information explosion can be bad for data mining, and how data fusion provides a way out. In Robert L. Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani, editors, *Proceedings of the Second SIAM International Conference on Data Mining*. SIAM, 2002.

On a Conditional Irrelevance Relation for Belief Functions based on the Operator of Composition

Radim Jiroušek

Institute of Information and Automation
Academy of Sciences of the Czech Republic,
and
Faculty of Management, University of Economics, Prague
radim@utia.cas.cz

Abstract. The paper presents one additional possibility how to define conditional independence relation for belief functions. The approach is based on the operator of composition originally designed for multidimensional model processing. Not to make confusion with the preceding definitions we call this relation *conditional irrelevance*. The main result of the paper, Block Irrelevance theorem, shows that this relation meets the semigraphoid axioms.

1 Introduction

Last years of the last century witnessed emergence of a new approach to efficient representation of multidimensional probability distributions. This approach, which is an alternative to Graphical Markov Modeling, is based on a simple idea: multidimensional distribution is *composed* from a system of low-dimensional (oligodimensional) distributions by repetitive application of a special operator of composition. This is also the reason why the models are called *compositional models*. In several papers, in which the properties of the operator were studied [3–5], it was shown (among others) that these models are, in a way, equivalent to Bayesian networks. Roughly speaking, *any multidimensional distribution representable by a Bayesian network can also be represented with approximately the same number of parameters (probabilities) in the form of a compositional models, and vice versa.*

Once (upon a time), after presenting a lecture on compositional models at some conference I was asked a question whether it would be possible to introduce analogous models also in the framework of belief functions. I have to confess that that time my answer was rather negative. Nevertheless, since then the question persistently came back to my subconscious until it originated a serious research, which resulted in the contribution to ISIPTA'07 ([6]), in which the operator of composition for belief functions was introduced. The discussions connected with writing that paper inspired me to make a rather provoking proposal: to introduce a new definition of a conditional independence relation in the framework of belief functions with the help of the operator of composition. Since there are many other definitions of this notion, we call the relation *conditional irrelevance*.

The goal of this paper is neither to show that this definition is the best one, nor to compare this definition with all the others. The goal of this paper is rather modest. To show that the notion of conditional irrelevance introduced with the help of the operator of composition intuitively corresponds to the properties expected and that also formally it meets the required characteristics, namely the semigraphoid properties.

Though the present paper is a contribution to belief function theory, we will not use the term of *belief function* any more in this paper. We are convinced that it will make the paper more legible for the reader when we will restrict our considerations to *basic belief assignments*, only. Therefore we will define a composition of basic assignments.

As said above, the paper is somehow connected with the contribution [6]. In fact it takes over from this paper not only the denotation and some assertions but also some other formulations (and one example).

The contribution is organized as follows. In Section 2 it summarizes basic notions, notation and introduce the operator of composition, which is consequently illustrated by examples in Section 3. Section 4 is devoted to the necessary properties of the operator and finally, in Section 5 we introduce the notion of conditional irrelevance and prove Block Irrelevance Theorem.

2 Notation

Consider a finite index set $N = \{1, 2, \ldots, n\}$ and finite sets $\{\mathbf{X}_i\}_{i \in N}$. In this text we will consider *multidimensional frame of discernment*

$$\Omega = \mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and its *subframes*. For $K \subset N$, $\mathbf{X}_K$ denotes a Cartesian product of those $\mathbf{X}_i$, for which $i \in K$:

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

A *projection* of $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1, i_2, \ldots, i_\ell\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_\ell}) \in \mathbf{X}_K.$$

Analogously, for $K \subset L \subseteq N$ and $A \subset \mathbf{X}_L$, $A^{\downarrow K}$ will denote a *projection* of A into $\mathbf{X}_K$:

$$A^{\downarrow K} = \{y \in \mathbf{X}_K | \exists x \in A : y = x^{\downarrow K}\}.$$

Let us remark that we do not exclude situations when $K = \emptyset$. In this case $A^{\downarrow \emptyset} = \emptyset$.

In addition to the projection, in this text we will need also the opposite operation which will be called extension. By an *extension* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ we will understand a set

$$A \otimes B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \ \ \& \ \ x^{\downarrow L} \in B\}.$$

Notice that if $A^{\downarrow K \cap L} \cap B^{\downarrow K \cap L} = \emptyset$ then also $A \otimes B = \emptyset$.

Consider a *basic* (*probability or belief*) *assignment* (or just assignment) m on $\mathbf{X}_N$, i.e.

$$m : \mathcal{P}(\mathbf{X}_N) \longrightarrow [0,1]$$

for which $\sum_{A \subseteq \mathbf{X}_N} m(A) = 1$. For each $K \subset N$ its *marginal basic assignment* is defined (for each $B \subseteq \mathbf{X}_K$):

$$m^{\downarrow K}(B) = \sum_{A \subseteq \mathbf{X}_N : A^{\downarrow K} = B} m(A).$$

Having two basic assignments m_1 and m_2 on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively (we assume that $K, L \subseteq N$), we say that these assignments are *projective* if

$$m_1^{\downarrow K \cap L} = m_2^{\downarrow K \cap L},$$

which occurs if and only if there exists a basic assignment m on $\mathbf{X}_{K \cup L}$ such that both m_1 and m_2 are marginal assignments of m.

Now, we are ready to introduce the operator of composition. Consider two sets $K, L \subset N$. At this moment we do not pose any restrictions on K and L; they may be but need not be disjoint, one may be subset of the other. We even admit that one or both of them are empty[1]. Let m_1 and m_2 be basic assignments on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively.

Our goal is to define new basic assignment, denoted $m_1 \triangleright m_2$, which will be defined on $\mathbf{X}_{K \cup L}$ and will contain all of the information contained in m_1 and as much as possible of information of m_2 (for the exact meaning see properties (iii) and (iv) of Lemma 1). The required property is met by the following definition.

Definition 1. For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ a *composition* $m_1 \triangleright m_2$ is defined for each $C \subseteq \mathbf{X}_{K \cup L}$ by one of the following expressions:

[**a**] if $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L}) > 0$ and $C = C^{\downarrow K} \otimes C^{\downarrow L}$ then

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L})};$$

[**b**] if $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L}) = 0$ and $C = C^{\downarrow K} \times \mathbf{X}_{L \setminus K}$ then

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

[**c**] in all other cases

$$(m_1 \triangleright m_2)(C) = 0.$$

[1] Notice that basic assignment m on $\mathbf{X}_\emptyset$ is defined $m(\emptyset) = 1$. Let us note that this is the only case where we accept $m(\emptyset) > 0$, otherwise $m(\emptyset) = 0$ according to the classical definitions of basic assignment and belief function, see [7].

Remark 1. The definition is inspired by the probabilistic operator of composition, which was defined in [3]. For probability distributions $\pi(x_K)$ and $\kappa(x_L)$ $(x_K \in \mathbf{X}_K, x_L \in \mathbf{X}_L)$ for which

$$\forall x \in \mathbf{X}_{K\cap L}\ (\kappa(x) = 0 \Longrightarrow \pi(x) = 0)$$

their composition $\pi \triangleright \kappa$ was defined by the formula:

$$\pi(x_K) \triangleright \kappa(x_L) = \frac{\pi(x_K)\kappa(x_L)}{\kappa(x_{K\cap L})}.$$

In case that there existed $x \in \mathbf{X}_{K\cap L}$ such that $\kappa(x) = 0$ and $\pi(x) > 0$ then the composition remained undefined.

From the point of view of this paper the most important issue concerning this probabilistic definition is that $\pi(x_K) \triangleright \kappa(x_L)$, if defined, is a probability distribution under which variables $X_{K\setminus L}$ and variables $X_{L\setminus K}$ are conditionally independent given variables $X_{K\cap L}$. ■

Remark 2. Why, in the definition of the operator of composition for basic assignments, do we have to distinguish three different situations [a], [b] and [c]?

Case [a] corresponds to those $C \subseteq \mathbf{X}_{K\cup L}$, which bear the information contained in m_1 and m_2 and resembles the definition from probability theory.

Case [b] is used in the situations where is a danger of a strict discord; the discord occurs when $m_2^{K\cap L}(C^{K\cap L}) = 0 < m_1^{K\cap L}(C^{K\cap L})$. In probability theory, in this situation the composition remained undefined. For belief functions we have a possibility to assign the respective mass $m_1(C^{\downarrow K})$ (regardless it is positive or not) to that C corresponding to the maximum ignorance.

Regarding the topic of this paper it is important to stress the point [c]. Namely, this point says that 0 is assigned to all the sets $C \subseteq \mathbf{X}_{K\cup L}$, which describe undesirable (conditional) dependence[2]. ■

Remark 3. Let us stress, for the reader familiar with the Dempster's rule of combination, that the introduced operator is something quite different.

First, Dempster's rule of combination is defined for two basic assignments defined on the same frame of discernment (there is no restriction regarding frames of discernments of arguments connected with the introduced operator of composition; nevertheless, composition of basic assignments defined on the same frame of discernment is uninteresting, because in this case the result is always the first argument - see the next Remark).

[2] Let us illustrate this property at this moment just by the simplest possible example. Consider a situation when a positive mass is assigned to set $\{(0,1),(1,0)\} \subset \mathbf{X}_{1,2}$. It says that we have a positive belief that variables X_1 and X_2 do not equal to each other. Therefore, we have a positive belief about their mutual relationship and thus we can hardly speak about their independence or irrelevance. The conditional version of this simple (unconditional) case will be discussed in Example 2

Moreover, Dempster's rule of combination (for $C \neq \emptyset$)

$$(m_1 \oplus m_2)(C) = \frac{\sum\limits_{A \cap B = C} m_1(A) \cdot m_2(B)}{1 - \sum\limits_{A \cap B = \emptyset} m_1(A) \cdot m_2(B)}$$

equals $(m_2 \oplus m_1)(C)$; the respective operator $\oplus$ is commutative, which is not the case for the operator $\triangleright$ - see also property (iii) of Lemma 1.

The reader should keep in mind that the operator of composition was designed for the situations when one has two basic assignments defined on different frames of discernment and wants to get a new basic assignments defined on a larger frame of discernment incorporating (as much as possible of) information contained in the original basic assignments. ■

Remark 4. Notice, what Definition 1 yields in the following two degenerate situations:

- if $K \cap L = \emptyset$ then $m_1 \triangleright m_2 = m_1 \cdot m_2$ (recall that $m_2^{\downarrow\emptyset}(\emptyset) = 1$) – for details regarding this situation see Example 1;
- if $K \supseteq L$ then $m_1 \triangleright m_2 = m_1$. ■

3 Examples

Example 1. Consider two basic assignments m_1 and m_2 on $\mathbf{X}_1 = \{a, \bar{a}\}$ and $\mathbf{X}_2 = \{b, \bar{b}\}$, respectively, which are specified in Table 1.[3] Since, in this case, m_1

Table 1. Basic assignments m_1 and m_2.

$A \subseteq \mathbf{X}_1$	$m_1(A)$
$\{a\}$	0.2
$\{\bar{a}\}$	0.3
$\{a\bar{a}\}$	0.5

$A \subseteq \mathbf{X}_2$	$m_2(A)$
$\{b\}$	0.6
$\{\bar{b}\}$	0
$\{a\bar{b}\}$	0.4

and m_2 are defined for disjoint sets of variables ($K \cap L$ is empty), composition simplifies to the expression

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow\{1\}}) \cdot m_2(C^{\downarrow\{2\}}),$$

which is to be understood exactly in the sense of Definition 1: for all C such that $C = C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}}$ it is defined by the product $m_1(C^{\downarrow\{1\}}) \cdot m_2(C^{\downarrow\{2\}})$, for all the other C it is 0 (see also Table 2).

Using Table 2, where the values of $m_1 \triangleright m_2$ are presented, the reader can easily check that $m_1 = (m_1 \triangleright m_2)^{\downarrow\{1\}}$, and since m_1 and m_2 are trivially projective also $m_2 = (m_1 \triangleright m_2)^{\downarrow\{2\}}$ (see Lemma 1 below). ■

[3] Let us note that, for the sake of simplicity, we use in examples $x_1 \ldots x_n$ instead of $(x_1, \ldots, x_n)$.

Table 2. Basic assignment $m_1 \triangleright m_2$ from Example 1.

$C \subseteq \mathbf{X}_{\{1,2\}}$	$C = C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}}$	$(m_1 \triangleright m_2)(C)$
$\{ab\}$	$\{a\} \otimes \{b\}$	0.12
$\{a\bar{b}\}$	$\{a\} \otimes \{\bar{b}\}$	0
$\{\bar{a}b\}$	$\{\bar{a}\} \otimes \{b\}$	0.18
$\{\bar{a}\bar{b}\}$	$\{\bar{a}\} \otimes \{\bar{b}\}$	0
$\{ab, a\bar{b}\}$	$\{a\} \otimes \mathbf{X}_2$	0.08
$\{ab, \bar{a}b\}$	$\mathbf{X}_1 \otimes \{b\}$	0.3
$\{ab, \bar{a}\bar{b}\}$		0
$\{a\bar{b}, \bar{a}b\}$		0
$\{a\bar{b}, \bar{a}\bar{b}\}$	$\mathbf{X}_1 \otimes \{\bar{b}\}$	0
$\{\bar{a}b, \bar{a}\bar{b}\}$	$\{\bar{a}\} \otimes \mathbf{X}_2$	0.12
$\{ab, a\bar{b}, \bar{a}b\}$		0
$\{ab, a\bar{b}, \bar{a}\bar{b}\}$		0
$\{ab, \bar{a}b, \bar{a}\bar{b}\}$		0
$\{a\bar{b}, \bar{a}b, \bar{a}\bar{b}\}$		0
$\{ab, a\bar{b}, \bar{a}b, \bar{a}\bar{b}\}$	$\mathbf{X}_1 \otimes \mathbf{X}_2$	0.2

Example 2. Consider three binary variables X_1, X_2, X_3 with $\mathbf{X}_1 = \{a, \bar{a}\}$, $\mathbf{X}_2 = \{b, \bar{b}\}$, $\mathbf{X}_1 = \{c, \bar{c}\}$, and two 2-dimensional basic assignments m_1 and m_2 as specified in Table 3.

Notice that these two assignments are not projective; for this see their one-dimensional marginals in Table 4. Therefore, because of property (iii) of Lemma 1 presented below $m_1 \triangleright m_2 \neq m_2 \triangleright m_1$.

To determine general 3-dimensional assignment (of binary variables) one has to specify 255 numbers, because $\mathbf{X}_{\{1,2,3\}}$ has $2^8 - 1 = 255$ nonempty subsets. However, when computing $m_1 \triangleright m_2$, most of these 255 values equal 0 because most of these subsets do not meet the condition[4] $C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$ and therefore the corresponding value of the assignment $m_1 \triangleright m_2$ is defined by the point [c] of the definition.

[4] It is not difficult to show that for binary variables there are exactly 99 nonempty subsets $C \subseteq \mathbf{X}_{\{1,2,3\}}$, for which $C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$:
There are only 3 different $C^{\downarrow\{1,2\}}$, for which $C^{\downarrow\{2\}} = \{b\}$; namely $\{ab\}, \{\bar{a}b\}$ and $\{ab, \bar{a}b\}$. Analogously, there are 3 different $C^{\downarrow\{2,3\}}$, for which $C^{\downarrow\{2\}} = \{b\}$. Therefore, there are only 9 sets C, for which $C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$ and $C^{\downarrow\{2\}} = \{b\}$. Analogously, there are 9 such sets with $C^{\downarrow\{2\}} = \{\bar{b}\}$.
In the same way one can show that there are 81 sets C, for which $C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$ and $C^{\downarrow\{2\}} = \{b, \bar{b}\}$. This is because there are 9 different $C^{\downarrow\{1,2\}}$, for which $C^{\downarrow\{2\}} = \{b, \bar{b}\}$, and also 9 sets $C^{\downarrow\{2,3\}}$, for which $C^{\downarrow\{2\}} = \{b, \bar{b}\}$.

Table 3. Basic assignments $m_1(x_{\{1,2\}})$ and $m_2(x_{\{2,3\}})$.

$C \subseteq \mathbf{X}_{\{1,2\}}$	$m_1(C)$
$\{ab\}$	0.1
$\{a\bar{b}\}$	0.5
$\{\bar{a}b\}$	0.2
$\{\bar{a}\bar{b}\}$	0
$\{ab, a\bar{b}\}$	0
$\{ab, \bar{a}b\}$	0
$\{ab, \bar{a}\bar{b}\}$	0
$\{a\bar{b}, \bar{a}b\}$	0
$\{a\bar{b}, \bar{a}\bar{b}\}$	0
$\{\bar{a}b, \bar{a}\bar{b}\}$	0
$\{ab, a\bar{b}, \bar{a}b\}$	0
$\{ab, a\bar{b}, \bar{a}\bar{b}\}$	0
$\{ab, \bar{a}b, \bar{a}\bar{b}\}$	0
$\{a\bar{b}, \bar{a}b, \bar{a}\bar{b}\}$	0
$\{ab, a\bar{b}, \bar{a}b, \bar{a}\bar{b}\}$	0.2

$C \subseteq \mathbf{X}_{\{2,3\}}$	$m_2(C)$
$\{bc\}$	0
$\{b\bar{c}\}$	0
$\{\bar{b}c\}$	0.3
$\{\bar{b}\bar{c}\}$	0.1
$\{bc, b\bar{c}\}$	0
$\{bc, \bar{b}c\}$	0
$\{bc, \bar{b}\bar{c}\}$	0.1
$\{b\bar{c}, \bar{b}c\}$	0
$\{b\bar{c}, \bar{b}\bar{c}\}$	0
$\{\bar{b}c, \bar{b}\bar{c}\}$	0.1
$\{bc, b\bar{c}, \bar{b}c\}$	0
$\{bc, b\bar{c}, \bar{b}\bar{c}\}$	0
$\{bc, \bar{b}c, \bar{b}\bar{c}\}$	0.3
$\{b\bar{c}, \bar{b}c, \bar{b}\bar{c}\}$	0
$\{bc, b\bar{c}, \bar{b}c, \bar{b}\bar{c}\}$	0.1

What are the subsets for which $C \neq C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$? For example, it is easy to show that all the sets of cardinality 7 belong to this category (hint: show that for any $C \subseteq \mathbf{X}_{\{1,2,3\}}$, for which $|C| = 7$, $C^{\downarrow\{1,2\}} = \mathbf{X}_{\{1,2\}}$ and $C^{\downarrow\{2,3\}} = \mathbf{X}_{\{2,3\}}$).

Table 4. One-dimensional marginal assignments $m_1^{\downarrow\{1\}}, m_1^{\downarrow\{2\}}$ and $m_2^{\downarrow\{2\}}, m_2^{\downarrow\{3\}}$.

$A \subseteq \mathbf{X}_1$	$m_1^{\downarrow\{1\}}(A)$
$\{a\}$	0.6
$\{\bar{a}\}$	0.2
$\{a, \bar{a}\}$	0.2

$A \subseteq \mathbf{X}_2$	$m_1^{\downarrow\{2\}}(A)$
$\{b\}$	0.3
$\{\bar{b}\}$	0.5
$\{b, \bar{b}\}$	0.2

$A \subseteq \mathbf{X}_2$	$m_2^{\downarrow\{2\}}(A)$
$\{b\}$	0
$\{\bar{b}\}$	0.5
$\{b, \bar{b}\}$	0.5

$A \subseteq \mathbf{X}_3$	$m_2^{\downarrow\{3\}}(A)$
$\{c\}$	0.3
$\{\bar{c}\}$	0.3
$\{c, \bar{c}\}$	0.4

Since all singletons (one-point-sets) meet the considered equality, all sets C, for which $C \neq C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$ must have at least two elements: an example is $\{abc, \bar{a}b\bar{c}\}$. As further examples may serve sets $\{ab\bar{c}, \bar{a}bc, \bar{a}b\bar{c}, a\bar{b}\bar{c}\}$ and $\{\bar{a}bc, a\bar{b}c, ab\bar{c}\}$. A common characteristics of all these sets is that assigning a positive belief to them one introduces a type of conditional relationship between X_1 and X_3 given (at least one) value of X_2.

Let us turn our attention back to computation of $m_1 \triangleright m_2$ for assignments of our example. For this, one immediately notices that point [b] of the definition is used whenever $C \subseteq \mathbf{X}_{\{1,2,3\}}$ is considered for which $C^{\downarrow\{2\}} = b$, since $m_2^{\downarrow\{2\}}(b) =$

0. In fact, we get only 8 subsets, for which the assignment is positive - see Table 5, where the first column bears the information, which point of the definition is used to compute the respective value.

Table 5. Basic assignment $m_1 \triangleright m_2$ for Example 2.

	$C \subseteq \mathbf{X}_{\{1,2,3\}}$	$C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$	$(m_1 \triangleright m_2)(C)$
[a]	$\{a\bar{b}c\}$	$\{a\bar{b}\} \otimes \{\bar{b}c\}$	0.3
[a]	$\{a\bar{b}\bar{c}\}$	$\{a\bar{b}\} \otimes \{\bar{b}\bar{c}\}$	0.1
[a]	$\{a\bar{b}c, a\bar{b}\bar{c}\}$	$\{a\bar{b}\} \otimes \{\bar{b}c, \bar{b}\bar{c}\}$	0.1
[b]	$\{abc, ab\bar{c}\}$	$\{ab\} \otimes \mathbf{X}_1$	0.1
[b]	$\{\bar{a}bc, \bar{a}b\bar{c}\}$	$\{\bar{a}b\} \otimes \mathbf{X}_1$	0.2
[a]	$\{abc, \bar{a}bc, a\bar{b}\bar{c}, \bar{a}\bar{b}\bar{c}\}$	$\mathbf{X}_{\{1,2\}} \otimes \{bc, \bar{b}\bar{c}\}$	0.04
[a]	$\{abc, \bar{a}bc, a\bar{b}c, \bar{a}\bar{b}c, a\bar{b}\bar{c}, \bar{a}\bar{b}\bar{c}\}$	$\mathbf{X}_{\{1,2\}} \otimes \{bc, \bar{b}c, \bar{b}\bar{c}\}$	0.12
[a]	$\{abc, \bar{a}bc, a\bar{b}c, \bar{a}\bar{b}c, ab\bar{c}, \bar{a}b\bar{c}, a\bar{b}\bar{c}, \bar{a}\bar{b}\bar{c}\}$	$\mathbf{X}_{\{1,2\}} \otimes \mathbf{X}_{\{2,3\}}$	0.04

■

4 Basic properties of composition

In this section we shall recollect three assertions proved in [6], which will be used in this paper.

Lemma 1. *For arbitrary two basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ the following properties hold true:*

(i) $m_1 \triangleright m_2$ *is a basic assignment on* $\mathbf{X}_{K \cup L}$.
(ii) $(m_1 \triangleright m_2)^{\downarrow K} = m_1$.
(iii) $m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K \cap L} = m_2^{\downarrow K \cap L}$.
(iv) *If* $K \subseteq L$ *then* $m_2^{\downarrow K} \triangleright m_2 = m_2$.

Realize that property (iii) of the preceding Lemma says that the operator is commutative if and only if it is applied to two projective basic assignments. Generally, it is neither commutative nor associative. Therefore the following assertion is of great importance.

Lemma 2. *Let m_1, m_2, m_3 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \mathbf{X}_{K_3}$, respectively. If $K_1 \supseteq (K_2 \cap K_3)$ then*

$$(m_1 \triangleright m_2) \triangleright m_3 = (m_1 \triangleright m_3) \triangleright m_2. \tag{1}$$

The next assertion, which is the last one borrowed from [6], expresses conditions under which marginalization of a composition is simple.

Lemma 3. *Let m_1, m_2 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}$, respectively. If $K_1 \cup K_2 \supseteq L \supseteq K_1 \cap K_2$ then*

$$(m_1 \triangleright m_2)^{\downarrow L} = m_1^{\downarrow K_1 \cap L} \triangleright m_2^{\downarrow K_2 \cap L}.$$

The last assertion of this section is original and therefore will be presented with its proof (it is a generalization of Lemma 6 from [6]).

Lemma 4. *Let m_1, m_2 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}$, respectively. If $K_1 \cup K_2 \supseteq L \supseteq K_2$ then*

$$m_1 \triangleright m_2 = m_1 \triangleright (m_1 \triangleright m_2)^{\downarrow L}.$$

Proof. Due to (ii) of Lemma 1 assignments m_1 and $(m_1 \triangleright m_2)^{\downarrow L}$ are projective and therefore (thanks to property (iii) of the same lemma) these arguments may be commuted

$$m_1 \triangleright (m_1 \triangleright m_2)^{\downarrow L} = (m_1 \triangleright m_2)^{\downarrow L} \triangleright m_1 = (m_1^{\downarrow K_1 \cap L} \triangleright m_2) \triangleright m_1,$$

where the second modification is justified by Lemma 3. The last expression meets the assumptions of Lemma 2 and therefore we can exchange second and third arguments, from which the required expression is got by application of property (iv) of Lemma 1:

$$(m_1^{\downarrow K_1 \cap K_2} \triangleright m_2) \triangleright m_1 = (m_1^{\downarrow K_1 \cap K_2} \triangleright m_1) \triangleright m_2 = m_1 \triangleright m_2.$$

□

5 Conditional irrelevance

Each multidimensional probability distribution defines on a set of its variables a ternary relation called *independence structure* [9–11]. Its knowledge enables us to decompose the distribution, which may further help with its efficient representation and processing. If it is known, for example, that for a probability distribution π variables X_I and X_J are conditionally independent when variables X_K are given (for I, J, K disjoint), then distribution $\pi(X_{I \cup J \cup K})$ is uniquely specified by its marginals $\pi(X_{I \cup K})$ and $\pi(X_{J \cup K})$ and can be expressed as

$$pi(X_{I \cup J \cup K}) = \frac{\pi(X_{I \cup K}) \cdot \pi(X_{J \cup K})}{\pi(X_K)}.$$

In this section we shall deal with similar properties for basic assignments. Since there have been introduced several notions of conditional independence in the literature and we do not know their relation to the notion introduced in this paper, we will call the newly introduced relation *conditional irrelevance*. In a way we proceed in an opposite direction than mentioned above. If the basic assignment can be decomposed, we will call the corresponding sets of variables conditionally irrelevant.

Definition 2. Consider an arbitrary basic assignment m on $\mathbf{X}_M$ and three disjoint subsets $I, J, K \subset M$ $(I \neq \emptyset \neq J)$. We say that for basic assignment m *variables* X_I *and variables* X_J *are conditionally irrelevant given variables* X_K (in symbol $I \perp\!\!\!\perp J|K\ [m]$) if

$$m^{\downarrow I\cup J\cup K} = m^{\downarrow I\cup K} \triangleright m^{\downarrow J\cup K}.$$

If $K = \emptyset$ we will also say that variables X_I and variables X_J are (unconditionally) irrelevant (for basic assignment m). For this special situation we will also use simplified notation $I \perp\!\!\!\perp J\ [m]$.

Remark 5. Notice that since $m^{\downarrow I\cup K}$ and $m^{\downarrow J\cup K}$ are projective, $m^{\downarrow I\cup K} \triangleright m^{\downarrow J\cup K} = m^{\downarrow J\cup K} \triangleright m^{\downarrow I\cup K}$ (due to property (iii) of Lemma 1) and therefore the conditional irrelevance relation is symmetric in the sense that $I \perp\!\!\!\perp J|K\ [m] = J \perp\!\!\!\perp I|K\ [m]$.

Let us now show that for the relation of conditional irrelevance the following *Block Independence* theorem holds true, which means (together with the statement from the previous Remark 5) that this relation meets the semigraphoid properties.

Theorem 1. *Let* I, J, K, L *be disjoint subsets of* M*, let* I, J, K *be nonempty. Then for any basic assignment* m *on* $\mathbf{X}_M$ *the following equivalence holds true:*

$$I \perp\!\!\!\perp J \cup K|L\ [m] \quad \Longleftrightarrow \quad (I \perp\!\!\!\perp J|L\ [m])\ \&\ (I \perp\!\!\!\perp K|L \cup J\ [m]).$$

Proof. Validity of $I \perp\!\!\!\perp J \cup K|L\ [m] \implies I \perp\!\!\!\perp J|L\ [m]$ follows immediately from application of Lemma 3 (it is applicable because $(I\cup L)\cap(J\cup K\cup L) \subset I\cup J\cup L$):

$$m^{\downarrow I\cup J\cup L} = (m^{\downarrow I\cup J\cup K\cup L})^{\downarrow I\cup J\cup L} = (m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L})^{\downarrow I\cup J\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup L}.$$

To prove $I \perp\!\!\!\perp J \cup K|L\ [m] \implies I \perp\!\!\!\perp K|J \cup L\ [m]$ we will use in the following computations only property (iii) of Lemma 1 and Lemma 4 (its application is possible because $I \cup J \cup K \cup L \supseteq I \cup J \cup L \supseteq I \cup L$):

$$\begin{aligned} m^{\downarrow I\cup J\cup K\cup L} &= m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L} = m^{\downarrow J\cup K\cup L} \triangleright m^{\downarrow I\cup L} \\ &= m^{\downarrow J\cup K\cup L} \triangleright (m^{\downarrow J\cup K\cup L} \triangleright m^{\downarrow I\cup L})^{\downarrow I\cup J\cup L} \\ &= m^{\downarrow J\cup K\cup L} \triangleright (m)^{\downarrow I\cup J\cup L} = m^{\downarrow I\cup J\cup L} \triangleright m^{\downarrow J\cup K\cup L}. \end{aligned}$$

In the previous two steps we have proved one side of the required equivalence. Now, let us assume that $(I \perp\!\!\!\perp J|L\ [m])\ \&\ (I \perp\!\!\!\perp K|L \cup J\ [m])$.

According to the definitions of these two conditional irrelevance relations we get

$$m^{\downarrow I\cup J\cup K\cup L} = m^{\downarrow I\cup J\cup L} \triangleright m^{\downarrow J\cup K\cup L} = (m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup L}) \triangleright m^{\downarrow J\cup K\cup L}.$$

The last expression can be further modified using successively property (iii) of Lemma 1, Lemma 2, property (iv) of Lemma 1 and eventually again property (iii) of Lemma 1:

$$\begin{aligned} (m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup L}) \triangleright m^{\downarrow J\cup K\cup L} &= (m^{\downarrow J\cup L} \triangleright m^{\downarrow I\cup L}) \triangleright m^{\downarrow J\cup K\cup L} \\ &= (m^{\downarrow J\cup L} \triangleright m^{\downarrow J\cup K\cup L}) \triangleright m^{\downarrow I\cup L} \\ &= (m^{\downarrow J\cup K\cup L}) \triangleright m^{\downarrow I\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L}. \quad \square \end{aligned}$$

Example 3. Let us consider the following simplified situation: Joan is considering whether to go for a walk or not. To make the decision she has two sources of information: weather forecast from the yesterday's newspaper and the view out of the window. The situation will be modeled by three variables:

- X_1: weather forecast with two values s, c meaning 'sunny' and 'changeable', respectively;
- X_2 corresponds to the current weather: value r means that it is raining, value d means it does not rain, it is dry;
- X_3 describing Joan's decision: values w, h corresponding to decision 'go for a walk' and 'stay at home', respectively.

So, we are considering 3-dimensional frame of discernment $\Omega = \{s, c\} \times \{r, d\} \times \{w, h\}$. Assume that our belief regarding her decision-making situation is represented by 3-dimensional basic assignment given in Table 6 (in the table only positive values of the considered assignment appear, all others equal 0).

Table 6. 3-dimensional basic assignment describing Joan's walk example.

$C \subseteq \mathbf{X}_{\{1,2,3\}}$	$C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$	$m(C)$
$\{srh\}$	$\{sr\} \otimes \{rh\}$	$0.0\bar{6}$
$\{sdw\}$	$\{sd\} \otimes \{dw\}$	0.24
$\{crh\}$	$\{cr\} \otimes \{rh\}$	$0.1\bar{3}$
$\{cdw\}$	$\{cd\} \otimes \{dw\}$	0.06
$\{srw, srh\}$	$\{sr\} \otimes \{rw, rh\}$	$0.0\bar{3}$
$\{sdw, sdh\}$	$\{sd\} \otimes \{dw, dh\}$	0.16
$\{crw, crh\}$	$\{cr\} \otimes \{rw, rh\}$	$0.0\bar{6}$
$\{cdw, cdh\}$	$\{cd\} \otimes \{dw, dh\}$	0.04
$\{srw, sdh, crw, cdh\}$	$\mathbf{X}_{\{1,2\}} \otimes \{rw, dh\}$	0.10
$\{srw, srh, sdw, sdh, crw, crh, cdw, cdh\}$	$\mathbf{X}_{\{1,2\}} \otimes \mathbf{X}_{\{2,3\}}$	0.10

The question we want to answer is whether the decision regarding the walk and weather forecast are conditionally irrelevant given the current weather, i.e. whether $\{1\} \perp\!\!\!\perp \{3\}|\{2\}$ $[m]$. The necessary condition for this conditional irrelevance relation is fulfilled because all the sets $C \subseteq \mathbf{X}_{\{1,2,3\}}$ which are assigned positive value meet the condition $C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$. If this was not the case then it would be out of question that $m = m^{\downarrow\{1,2\}} \triangleright m^{\downarrow\{2,3\}}$.

To be able to answer the question compute the necessary 2-dimensional marginal assignments (see Table 7) and 1-dimensional marginal $m^{\{2\}}$ (Table 8).

Table 7. Marginal basic assignments $m^{\downarrow\{1,2\}}$ and $m^{\downarrow\{2,3\}}$.

$C \subseteq \mathbf{X}_{\{1,2\}}$	$m^{\downarrow\{1,2\}}(C)$
$\{sr\}$	0.1
$\{sd\}$	0.4
$\{cr\}$	0.2
$\{cd\}$	0.1
$\{sr, sd\}$	0
$\{sr, cr\}$	0
$\{sr, cd\}$	0
$\{sd, cr\}$	0
$\{sd, cd\}$	0
$\{cr, cd\}$	0
$\{sr, sd, cr\}$	0
$\{sr, sd, cd\}$	0
$\{sr, cr, cd\}$	0
$\{sd, cr, cd\}$	0
$\{sr, sd, cr, cd\}$	0.2

$C \subseteq \mathbf{X}_{\{2,3\}}$	$m^{\downarrow\{2,3\}}(C)$
$\{rw\}$	0
$\{rh\}$	0.2
$\{dw\}$	0.3
$\{dh\}$	0
$\{rw, rh\}$	0.1
$\{rw, dw\}$	0
$\{rw, dh\}$	0.1
$\{rh, dw\}$	0
$\{rh, dh\}$	0
$\{dw, dh\}$	0.2
$\{rw, rh, dw\}$	0
$\{rw, rh, dh\}$	0
$\{rw, dw, dh\}$	0
$\{rh, dw, dh\}$	0
$\{rw, rh, dw, dh\}$	0.1

From this we can verify that really $m = m^{\downarrow\{1,2\}} \triangleright m^{\downarrow\{2,3\}}$. It is enough to verify it for 12 arguments from Table 6. For example, for $C = \{sdw, sdh\}$ we get (using point [a] of Definition 1):

$$(m^{\downarrow\{1,2\}} \triangleright m^{\downarrow\{2,3\}})(\{sdw, sdh\}) = \frac{m^{\downarrow\{1,2\}}(\{sd\}) \cdot m^{\downarrow\{2,3\}}(\{dw, dh\})}{m^{\downarrow\{2\}}(\{d\})}$$
$$= \frac{0.4 \cdot 0.2}{0.5} = 0.16.$$

Let us remark that all positive values of $m^{\downarrow\{1,2\}} \triangleright m^{\downarrow\{2,3\}}$ are computed according to point [a] of Definition 1. Point [b] is never used in this situation because, obviously, marginal assignments $m^{\downarrow\{1,2\}}$ and $m^{\downarrow\{2,3\}}$ must be projective. ■

Table 8. One-dimensional marginal assignment $m^{\downarrow\{2\}}$.

$C \subseteq \mathbf{X}_2$	$m^{\downarrow\{2\}}(C)$
$\{r\}$	0.3
$\{d\}$	0.5
$\{r, d\}$	0.2

6 Conclusions

We have introduced the operator of composition for basic belief assignments. Originally, the operator of composition was designed to construct and to compute with multidimensional probabilistic models. If we are getting into problems when coping with computational complexity of probabilistic models all the more problems necessarily appear when applying belief function models, for which there do not exist distribution functions; we have to represent them by set functions defined on the whole power set of the frame of discernment $\Omega = \mathbf{X}_N$. Therefore, whilst multidimensionality for probability distributions means hundreds and thousands, multidimensionality for belief functions means tens at maximum.

However constructing multidimensional models was not a topic of the paper. Based on the properties of the operator of composition we have introduced a new type of conditional independence relation, called in the paper conditional irrelevance. We have shown that this relation meets the requirements of Block Independence theorem, which is (for symmetric relations) equivalent to semigraphoid axioms. The main purpose was not to assert that this type of relation is the only one which should be studied, rather the opposite. We wanted to provoke questions concerning similarities and differences between individual approaches to define conditional independence/irrelevance for belief functions. At this moment we cannot answer these questions, we know very little about relation between the models described in this paper and other models such as [1, 2, 8], as well as about the relation between the compositional models developed for belief functions and those introduced in possibility theory [12, 13].

Acknowledgements

The research was partially supported by GA AV ČR under grants A2075302, and MŠMT under grants 1M0572 and 2C06019.

References

1. R. G. Almond. *Graphical Belief Modelling.* Chapman & Hall, London, 1995.
2. E. Ben Yaghlane, Ph. Smets, and K. Mellouli. Directed Evidential Networks with Conditional Belief functions. *Proc. of ECSQARU 2003*, LNAI 2711, Springer, pp. 291–305, 2003.
3. R. Jiroušek. Composition of probability measures on finite spaces. *Proc. of the 13th Conf. Uncertainty in Artificial Intelligence UAI'97*, (D. Geiger and P. P. Shenoy, eds.). Morgan Kaufmann Publ., San Francisco, California, pp. 274–281, 1997.
4. R. Jiroušek. Graph modelling without graphs. *Proc. of the 7th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'98*, (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, pp. 809–816, 1988.
5. R. Jiroušek. Marginalization in composed probabilistic models. *Proc. of the 16th Conf. Uncertainty in Artificial Intelligence UAI'00* (C. Boutilier and M. Goldszmidt eds.), Morgan Kaufmann Publ., San Francisco, California, pp. 301–308, 2000.

6. R. Jiroušek, J. Vejnarová and M. Daniel. Compositional Models of Belief Functions. In: *Proc. of the 5th Int. Symposium on Imprecise Probabilitis and Their Applications ISIPTA'07*, (G. de Cooman, J. Vejnarová, M. Zaffalon, eds.). Charles University Press, Praha, pp. 243-252, 2007.
7. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey, 1976.
8. P. P. Shenoy. Binary join trees for computing marginals in the Shenoy-Shafer architecture. *Int. J. of Approximate Reasoning*, **17**, (2-3), pp. 239–263, 1997.
9. M. Studený. On stochastic conditional independence: the problems of characterization and description. *Annals of Mathematics and Artificial Intelligence,* **35**, p. 323-341, 2002.
10. M. Studený. Other approaches to the description of conditional independence structures. In: *Highly Structured Stochastic Systems,* (Green P. J., Hjort N. L., Richardson S., eds.). Oxford University Press, New York, pp 106-108, 2003.
11. M. Studený. *Probabilistic Conditional Independence Structures.* Springer, London, 2005.
12. J. Vejnarová. Composition of possibility measures on finite spaces: preliminary results. In: *Proc. of Proceedings of 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'98,* (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, pp. 25–30, 1998.
13. J. Vejnarová. Possibilistic independence and operators of composition of possibility measures. In: M. Hušková , J. Á. Víšek, P. Lachout (eds.) *Prague Stochastics'98,* JČMF, pp. 575–580, 1998.
14. H. Xu, and Ph. Smets. Reasoning in Evidential Networks with Conditional Belief functions. *Int. J. of Approximate Reasoning*, **14**, (2-3), pp. 155–185, 1996.

Reason Maintenance and the Ramsey Test

Haythem O. Ismail

German University in Cairo
Department of Computer Science
haythem.ismail@guc.edu.eg

Abstract. The Ramsey test provides an intuitive link between conditionals and belief revision. How easy is it to incorporate a Ramsey-account of conditionals in a reason maintenance system? In this paper, it is shown that this is indeed possible, within a relevance-logical framework. In addition, it is shown that independently motivated requirements on reason maintenance systems allow us to gracefully circumvent Gärdenfors's triviality result.

1 Introduction

Frank Ramsey's so called "Ramsey test" provides an intuitive link between conditionals (sentences of the form 'If P then Q') and belief change. The test grounds the plausibility of conditionals in a process of belief change. In an often quoted excerpt from [1], Robert Stalnaker gives a procedural interpretation of the Ramsey test:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.

This procedure is particularly appealing for researchers in artificial intelligence (AI), who have long been interested in reasoning about conditionals [2, 3, for example]. What makes it even more appealing is that, in its crucial second step, it provides a characterization of conditionals based on the (familiar to AI) concept of belief revision. The AI and philosophical literature on belief revision seem to have originated from different concerns. On one hand, the AI researchers were primarily motivated by the implementation issues of supplementing a general reasoning system with the facility to maintain consistency and revise its beliefs. This gave rise to what are known as "reason maintenance" (or "truth maintenance") systems. [4, 5, for example]. On the other hand, the philosophers were keen to uncover a tight set of *rationality postulates* that govern the principles whereby a logical theory is to be revised [6]. These two attitudes have witnessed considerable convergence over the history of belief revision [7–9, for example].

To everyone's distress, however, Peter Gärdenfors [10] proved that the Ramsey test account of conditionals, together with some seemingly reasonable constraints on belief revision (three of the AGM postulates [6]), is inconsistent with a minimal set of harmless demands on a logical theory. Gärdenfors's result triggered considerable research attempting to save the intuitive interpretation of conditionals provided by the Ramsey test [11–17, for example].

In this paper, I attempt to do two things. First, I will argue that, if belief revision is interpreted in the context of an implemented reason maintenance system, Gärdenfors's triviality result is avoided. This comes as a consequence of rejecting some the AGM postulates, based on general demands on implemented reason maintenance systems that are independent of conditionals and Gärdenfors's result. Second, as a side effect, I shall outline a theory for reasoning about conditionals within an implemented knowledge representation and reasoning (KRR) system with a reason maintenance component. The discussion will primarily focus on the SNePS KRR system [18, 19] and its reason maintenance component SNeBR [5, 20, 9].

In Section 2, we review Gärdenfors's triviality theorem and, in Section 3, we examine previous attempts to rectify the damage it has wrought. Section 4 presents a reason maintenance system based on relevance logic, which is then extended to accommodate conditionals. Finally, Section 5 evaluates the system with respect to Gärdenfors's triviality result.

2 The Triviality Trap

What exactly did Gärdenfors discover? As pointed out in the introduction, he discovered that the Ramsey test is inconsistent with simple common demands on a logical system. In particular, Gärdenfors proved that the Ramsey test will introduce a contradiction into a belief set that contains none of three pair-wise contrary propositions. That such a belief set may exist is uncontroversial. Hence, the triviality result.

Nevertheless, Gärdenfors's proof is based on a mesh of background assumptions. These assumptions are primarily of two types: (i) assumptions on the belief revision process implied by the Ramsey test, and (ii) assumptions on what a "belief set" is. Attempts to circumvent the triviality result are based on dropping one or more of these assumptions. To get a deeper understanding of what exactly the problem is, and to appreciate previous approaches to solve it, we start by listing Gärdenfors's background assumptions.[1]

In what follows, $\mathcal{L}_0$ is a ground language of classical propositional or first-order logic including the absurd proposition $\perp$[2], $\mathcal{L}_1$ is the closure of $\mathcal{L}_0$ under

[1] This is not an exhaustive list. It is a list of those assumptions that are (so far) uncontroversial and/or relevant to my examination of previous work and my own proposal.

[2] It should be noted, however, that the triviality result was shown to be valid for a more general class of monotonic [21] and non-monotonic [22] logics.

the conditional connective $>$, $\mathbb{K}$ is a set of belief sets, $\mathcal{K} \in \mathbb{K}$ is a belief set, and $* : \mathbb{K} \times \mathcal{L}_1 \longrightarrow \mathbb{K}$ is a belief revision operator.

1. Assumptions on $\mathcal{K}$:
   ($A^{\mathcal{K}}1$) Belief sets are sets of sentences[3]: $\mathcal{K} \subseteq \mathcal{L}_1$.
   ($A^{\mathcal{K}}2$) Belief sets may include conditional sentences: For some $\mathcal{K} \in \mathbb{K}$, $\mathcal{K} \not\subseteq \mathcal{L}_0$.
   ($A^{\mathcal{K}}3$) Belief sets are deductively-closed: if $\phi \in \mathrm{Cn}_0(\mathcal{K})$ then $\phi \in \mathcal{K}$.[4]
2. Assumptions on belief revision:
   (A^*1) Success: $\phi \in \mathcal{K} * \phi$.
   (A^*2) Consistency: If $\bot \in \mathrm{Cn}_0(\mathcal{K} * \phi)$ then $\bot \in \mathrm{Cn}_0(\{\phi\})$.
   (A^*3) Expansion 1: $\mathrm{Cn}_0(\mathcal{K} * \phi) \subseteq \mathrm{Cn}_0(\mathcal{K} \cup \{\phi\})$.
   (A^*4) Expansion 2: If $\neg\phi \notin \mathrm{Cn}_0(\mathcal{K})$, then $\mathrm{Cn}_0(\mathcal{K} \cup \{\phi\}) \subseteq \mathrm{Cn}_0(\mathcal{K} * \phi)$.

From (A^*4), (A^*5) immediately follows:

(A^*5) Preservation: If $\neg\phi \notin \mathrm{Cn}_0(\mathcal{K})$, then $\mathrm{Cn}_0(\mathcal{K}) \subseteq \mathrm{Cn}_0(\mathcal{K} * \phi)$.

With these background assumptions, Gärdenfors [10] states the Ramsey test as follows.

(RT) $\phi > \psi \in \mathcal{K}$ if and only if $\psi \in \mathcal{K} * \phi$.

For ease of reference, let us break (RT) into two conditionals:

($RT1$) If $\phi > \psi \in \mathcal{K}$ then $\psi \in \mathcal{K} * \phi$.
($RT2$) If $\psi \in \mathcal{K} * \phi$ then $\phi > \psi \in \mathcal{K}$.

To simplify the proof of the triviality theorem, Gärdenfors first proves the following crucial lemma: The "monotonicity criterion".

(M) For all $\mathcal{K}, \mathcal{K}' \in \mathbb{K}$ and all $\phi \in \mathcal{L}_1$, if $\mathcal{K} \subseteq \mathcal{K}'$ then $\mathcal{K} * \phi \subseteq \mathcal{K}' * \phi$.

Proof.

1. $\mathcal{K} \subseteq \mathcal{K}'$	(Assumption)
2. $\psi \in \mathcal{K} * \phi$	(Assumption)
3. $\phi > \psi \in \mathcal{K}$	(2, ($RT2$))
4. $\phi > \psi \in \mathcal{K}'$	(1, 3)
5. $\psi \in \mathcal{K}' * \phi$	(4, ($RT1$))
6. $\mathcal{K} * \phi \subseteq \mathcal{K}' * \phi$	(2, 5)
7. (M)	(1, 6)

Before we present Gärdenfors's triviality result, we define the notion of non-triviality of a belief revision system.

[3] Gärdenfors [10] uses the term "proposition" instead of "sentence". Yet, he states that "belief sets are just theories in the standard logical sense" [10, p. 83].

[4] Cn_0 is classical (deductive) logical consequence. We may assume a natural deduction system, although only modus ponens and the deduction theorem are needed.

(NT) A belief revision system $\langle \mathcal{L}_1, \mathbb{K}, * \rangle$ is non-trivial if, for some $\mathcal{K} \in \mathbb{K}$ and $A, B, C \in \mathcal{L}_1$,

1. $\{\neg(A \wedge B), \neg(A \wedge C), \neg(B \wedge C)\} \subseteq \mathrm{Cn}_0(\mathcal{K})$; and
2. $\neg A \notin \mathrm{Cn}_0(\mathcal{K}), \neg B \notin \mathrm{Cn}_0(\mathcal{K})$, and $\neg C \notin \mathrm{Cn}_0(\mathcal{K})$.

Theorem 1. *There is no non-trivial belief revision system that satisfies* $(A^{\mathcal{K}}1)$–$(A^{\mathcal{K}}3)$, (A^*1)–(A^*4), *and RT.*

Proof.

1. $\neg A \notin \mathcal{K}$	(NT)
2. $\mathcal{K} * A = \mathrm{Cn}_0(\mathcal{K} \cup \{A\})$	$(1, A^{\mathcal{K}}3, A^*3, A^*4)$
3. $(B \vee C) \in (\mathcal{K} * A) * (B \vee C)$	(A^*1)
4. $\neg(B \vee C) \notin (\mathcal{K} * A) * (B \vee C)$	$(3,\ A^*2)$
5. $\neg C \notin (\mathcal{K} * A) * (B \vee C)$	(Assumption)
6. $\neg(A \vee B) \notin \mathcal{K}$	(NT)
7. $\mathcal{K} * (A \vee B) = \mathrm{Cn}_0(\mathcal{K} \cup \{A \vee B\})$	$(6, A^{\mathcal{K}}3, A^*3, A^*4)$
8. $\mathcal{K} * (A \vee B) \subseteq \mathcal{K} * A$	$(2,7,A^{\mathcal{K}}3, \mathrm{Cn}_0)$
9. $(\mathcal{K} * (A \vee B)) * (B \vee C) \subseteq (\mathcal{K} * A) * (B \vee C)$	$(8,M)$
10. $\neg C \notin (\mathcal{K} * (A \vee B)) * (B \vee C)$	(5,9)
11. $\neg(B \vee C) \notin \mathcal{K} * (A \vee B)$	$(7, NT, \mathrm{Cn}_0)$
12. $(\mathcal{K} * (A \vee B)) * (B \vee C) = \mathrm{Cn}_0(\mathcal{K} * (A \vee B) \cup \{B \vee C\})$	$(11, A^{\mathcal{K}}3, A^*3, A^*4)$
13.$(\mathcal{K} * (A \vee B)) * (B \vee C) = \mathrm{Cn}_0(\mathcal{K} \cup \{A \vee B, B \vee C\})$	$(7, 12)$
14. $(\mathcal{K} * (A \vee B)) * (B \vee C) = \mathrm{Cn}_0(\mathcal{K} \cup \{B\})$	$(13, NT)$
15. $\neg C \in \mathrm{Cn}_0(\mathcal{K} \cup \{B\}$	(NT)
16. $\neg C \in (\mathcal{K} * (A \vee B)) * (B \vee C)$	$(14, 15)$
17. $\perp$	$(10, 16)$
18. $(\neg C \notin (\mathcal{K} * A) * (B \vee C)) \supset \perp$	$(5, 17, \mathrm{Cn}_0)$
19. $(\neg B \notin (\mathcal{K} * A) * (B \vee C)) \supset \perp$	(Similarly, 5–17)
20. $\perp$	$(4, 18, 19)$

Having displayed the proof in detail, we can carefully analyze the different loopholes proposed in the literature. Each of the proposed loopholes identifies one or more of the background assumptions and/or the Ramsey test as the culprit.

3 Loopholes

We shall consider six proposals [12–17], each identifying a different set of background assumptions as the culprit, or a different way out of the triviality trap. Except for [14], all proposals attempt to invalidate the use of (M) in the proof. They all (*pace* Gärdenfors in [10, 11]) preserve the Ramsey test (some version of it, for that matter), and choose to reject M based on the background assumptions. Table 1 lists the culprits identified by each of the six proposals.

Rott [12] argues convincingly that (M) is indeed true, but only trivially so. Informally, once we admit conditionals into belief sets (as per $(A^{\mathcal{K}}2)$), no belief set can be a proper subset of another. From this general result, the invalidity

	$A^{\mathcal{K}}1$	$A^{\mathcal{K}}2$	$A^{\mathcal{K}}3$	A^*1	A^*2	A^*3	A^*4
Rott [12]						×	×
Hansson [13]			×			×	×
Arló Costa and Levi [14]		×					
Lindström and Rabinowicz [15]	×						
Grahne [16]							×
Giordano et àl [17]						×	×

Table 1. A comparison of six proposals to escape Gärdenfors's triviality result. The crosses indicate the culprits identified by each proposal.

of (A^*3) and (A^*4) (and hence (A^*5)) follows. The result is proved by Hansson [13], making use of the following property of proper subsets.

(PSS) $\mathcal{K} \in \mathbb{K}$ has a proper subset if and only if there are $\mathcal{K}_1 \subseteq \mathcal{K}$ and $\mathcal{K}_2 \subseteq \mathcal{K}$ such that $\mathcal{K}_1 \not\subseteq \mathcal{K}_2$ and $\mathcal{K}_2 \not\subseteq \mathcal{K}_1$.

To informally illustrate Hansson's proof, I will refer to an example based on one due to Darwiche and Pearl [23].

Example 1. A murder occurs. John and Mary are the prime suspects. Detective 1 finds evidence incriminating John ($\mathcal{K}_1 = \mathrm{Cn}_0(\{J\})$). Detective 2, on the other hand, finds evidence incriminating Mary ($\mathcal{K}_2 = \mathrm{Cn}_0(\{M\})$). Both detectives report to their supervisor ($\mathcal{K} = \mathrm{Cn}_0(\{J, M\})$). As long as we only consider sentences in $\mathcal{L}_0$, then it is clear that $\mathcal{K}_1, \mathcal{K}_2$ and $\mathcal{K}$ satisfy (PSS). However, intuitively, $\neg(J \wedge M) > (J \wedge \neg M) \in \mathcal{K}_1$.[5] Similarly, $\neg(J \wedge M) > (\neg J \wedge M) \in \mathcal{K}_2$. But these two conditionals are contradictory; they cannot both be in $\mathcal{K}$.

The above example illustrates the fundamental difficulty in finding three sets satisfying (PSS): incomplete information licences the belief in conditionals that lose their support in a more informed belief state. Inspecting Table 1, [12, 13, 17] take issue with (A^*3) and (A^*4). Most probably [12, 13, 17] would identify the main glitch in the triviality proof with step 7.[6] Though they agree on the culprit, each of these authors proposes a different way out of the triviality trap. Rott [12] informally considers employing a non-monotonic logic instead of Cn_0. Hansson [13] presents a detailed theory founded on belief base revision. (Hence,

[5] Hansson [13, p. 529] stresses that this last conclusion is based on "prephilosophical" intuitions, and (crucially) not on the Ramsey test. However, I find the acceptance of this prephilosophical intuition together with the Ramsey test (Hansson's position) a bit strange. For this certainly commits us to accepting certain properties of belief revision. In particular, we are thus committed to (A^*4)—the very assumption that Hansson rejects.

[6] Of course, by rejecting (A^*3) and (A^*4), any step that is licensed by either is flawed. Nevertheless, those authors seem to base most of their informal arguments on the invalidity of step 7.

his rejection of $(A^{\mathcal{K}}2)$.)[7] Giordano et àl [17] restrict (A^*3) and (A^*4) to the maximal $\mathcal{L}_0$-subset of $\mathcal{K}$.

Grahne [16] would probably identify step 12 as the main glitch. Grahne's rejection of (A^*4) is based, not on admitting conditionals into belief states (as per $(A^{\mathcal{K}}2)$), but on his very interpretation of the belief change operator appropriate for the interpretation of conditionals as per the Ramsey test. Instead of the classical AGM belief revision operator [6] , Grahne opts for Katsuno and Mendelzon's belief *update* operator [24]. Belief revision is appropriate for a change in belief signaled by acquiring information about a static world. Belief update, on the other hand, is needed when the change in belief is necessary due to a change in the world. When revising $\mathcal{K} * (A \vee B)$ with $B \vee C$ in step 12, we assume the world has not changed. That is, $A \vee B$ is still true. Thus, B immediately follows, since adding $B \vee C$ just gives us more specific information about which of the contraries A and B is indeed true. On the other hand, when updating $\mathcal{K} * (A \vee B)$ with $B \vee C$, we assume that the world has changed, and, thus, cannot assume that $A \vee B$ is till true. To our best knowledge, only $B \vee C$ is certain, but not B.

Arló Costa and Levi [14] (following a hard-line position of Levi's [25]) reject $A^{\mathcal{K}}2$. They argue that conditionals do not qualify as members of belief sets (or as truth-value bearers), but as representations of an agent's dispositions to change their beliefs. From the premise that this is Ramsey's own position, Gärdenfors's (RT) is disqualified as a formal rendering of the Ramsey test. Their proposal is to adopt a stratified theory, where there is a clear distinction between the belief set $\mathcal{K}$ (a subset of $\mathcal{L}_0$) and the set of sentences, $s(\mathcal{K}) \subseteq \mathcal{L}_1$, supported by $\mathcal{K}$. Only the latter may include conditionals. A stratified version of the Ramsey test may then be stated:

(SRT) For all $\phi, \psi \in \mathcal{L}_0$, $\phi > \psi \in s(\mathcal{K})$ if and only if $\psi \in \mathcal{K} * \phi$.

Although a stratified version of (M) may also be derived (with the antecedent being $s(\mathcal{K}) \subseteq s(\mathcal{K}')$), the proof of the triviality result will be blocked at step 7, since $\mathcal{K} \subseteq \mathcal{K}'$ does not entail $s(\mathcal{K}) \subseteq s(\mathcal{K}')$.

Lindström and Rabinowicz hold yet another position [15]. They identify $A^{\mathcal{K}}1$ as the sole culprit, and, instead of taking $\mathcal{K}$ to be a set of sentences, they assume it is a set of *propositions*. How does this assumption get us out of the triviality trap? The assumption by itself may not help if sentences and propositions stand in one-to-one correspondence. This is exactly what Lindström and Rabinowicz reject. They argue that conditionals are *context-sensitive*: the same conditional may express different propositions in different contexts.[8] On their account, a context is simply a belief set.[9] By adopting a context-sensitive version of the Ramsey test, the proof of (M) is blocked and the triviality result is avoided.

[7] A belief base $\mathcal{B} \subseteq \mathcal{K}$ is a set of beliefs such that $\mathrm{Cn}_0(\mathcal{B}) = \mathcal{K}$.

[8] It should be noted that the same position was skeptically considered by Gärdenfors himself [10, p. 91].

[9] Here, I am using the term "belief set" to refer to sets of propositions. Lindström and Rabinowicz [15] use the term "belief state", instead; they reserve "belief set" for sets of sentences.

Lindström and Rabinowicz's rendering of the Ramsey test could be presented as follows, where the semantics of $<$ depends on the context $\mathcal{K}$.

(CRT) $\phi >_{\mathcal{K}} \psi \in \mathcal{K}$ if and only if $\psi \in \mathcal{K} * \phi$.

The proof of (M) blocks in step 5, which cannot be proved since $>_{\mathcal{K}'}$ is needed in place of $>_{\mathcal{K}}$ in step 4.

In what follows, I will present an assumption-based reason maintenance system. The details of the system are based on assumptions that differ fundamentally from some of those underlying the triviality result. As it turns out, these assumptions, which are independently motivated by issues of rational agency and computational complexity, allow us to gracefully escape the triviality trap when the system is extended to accommodate conditionals.

4 Reason Maintenance and Conditionals

4.1 General Requirements on Reason Maintenance

Unlike belief revision theories, reason maintenance systems are required to take into account issues of bounded computational resources and availability. These issues motivate the following three requirements on reason maintenance systems.

RM1. Belief sets are not closed under logical consequence.
RM2. Paradoxes of implication are not tolerated.
RM3. Implicit inconsistencies are tolerated.

The motivation for **RM1** is clear; no realistic computational (or rational) reasoning system can be logically closed. While we may talk about the closure of a belief set to facilitate the analysis of its potential theorems, the belief set itself must be finite, and as small as possible for that matter. Clearly, **RM1** is at odds with $(A^{\mathcal{K}}3)$.

RM2 is particularly required to block the derivation of arbitrary sentences from contradictions. From the point of view of rational agency, it is clear that agents (notably humans) can accommodate contradictory beliefs without committing to logical absurdity. From the point of view of computational reasoning systems, a system should provide useful, sound inferences even in the presence of contradictions.

RM3 is probably the least obvious. However, once **RM2** is accepted, it is clear that the harmful effects of contradictions can be isolated. In addition, inconsistencies are only tolerated as long as they are only implicit, once a contradiction is explicitly derived (that is, added to the belief state), then consolidation is triggered.

In what follows, a reason maintenance system that satisfies the above requirements will be presented. The system is based on [5] and [20, 9]. It is implemented as SNeBR, the belief revision component of the SNePS knowledge representation and reasoning system [18, 19].

4.2 The Case of $\mathcal{L}_0$

First, let us consider a reason maintenance system for the language $\mathcal{L}_0$. It is clear that the classical Cn_0 does not observe **RM2**. SNePS logic is a version of Anderson and Belnap's *relevance logic* [26, 27]. A full exposition of relevance logic is not needed (and not possible) here. Suffice it to say, that relevance logic does observe **RM2**, and that it achieves this by keeping track of the history of derivations. (Thus, we seem to have an independent motivation for recording derivation traces, which is required by assumption-based reason maintenance.) In what follows, Cn_R denotes relevance logic consequence.

Definition 1 *A* ***support set*** *of a sentence $\phi \in \mathcal{L}_0$ is a set $s \subseteq \mathcal{L}_0$ such that $\phi \in \mathrm{Cn}_R(s)$. s is* ***minimal*** *if, for every $s' \subset s, \phi \notin \mathrm{Cn}_R(s')$.*

The reader should note that minimal support sets of a sentence ϕ are Hansson's *ϕ-kernels* [28].

Definition 2 *A* ***belief state*** *$\mathcal{S}$ is a quadruple $\langle \mathcal{K}, \mathcal{B}, \sigma, \preccurlyeq \rangle$, where:*

1. *$\mathcal{K} \subseteq \mathcal{L}_0$ is a* ***belief set****.*
2. *$\mathcal{B} \subseteq \mathcal{K}$, with $\mathcal{K} \subseteq \mathrm{Cn}_R(\mathcal{B})$, is a* ***belief base****. If $\phi \in \mathcal{B}$, then ϕ is a* ***base belief****.*
3. *$\sigma : \mathcal{K} \longrightarrow 2^{2^{\mathcal{B}}}$ is a* ***support function****, where each $s \in \sigma(\phi)$ is a minimal support set of ϕ. If $\phi \in \mathcal{B}$, then $\{\phi\} \in \sigma(\phi)$.*
4. *$\preccurlyeq \subseteq \mathcal{B} \times \mathcal{B}$ is a total pre-order on base beliefs.*[10]

On the intuitive interpretation of the above definition, base beliefs are beliefs that have independent standing. For example, they are the result of perception or interaction with another agent (possibly a human operator/user). Crucially, they are not in the belief state based *solely* on inference. The belief set $\mathcal{K}$ is not closed under Cn_R; it represents the set of sentences that are either base beliefs or that were *actually derived* from base beliefs.[11] This is in contrast to the logically-closed $\mathrm{Cn}_R(\mathcal{K})$ which is the set of sentences *derivable* from base beliefs.

The set $\sigma(\phi)$ is the family of minimal support sets that were actually used, or discovered, to derive ϕ. $\mathcal{B}$ may include minimal support sets of ϕ that are, nevertheless, not in $\sigma(\phi)$, if they are not yet discovered to derive ϕ. The total pre-order $\preccurlyeq$ represents a preference ordering over base beliefs. This ordering will be used when belief revision requires sacrificing a base belief; the least preferred will be the victim. I will refrain from making any commitments about the origins of this ordering. In particular, unlike standard epistemic entrenchment relations [29, for example], I am not assuming any logical basis for preference. For the purpose of this paper, the ordering is just given.[12]

[10] A total pre-order is a complete, reflexive, and transitive binary relation.

[11] Thus, in time, a belief state can evolve into a *different* belief state that share the same base.

[12] For future investigation, we may consider the possibility of moving $\preccurlyeq$ into the object language.

Belief revision in this system is a distant variant of Hansson's *semi-revision* [29]—a non-prioritized belief revision operator where success is not guaranteed. We first need to define a notion of *relevant expansion*.

Definition 3 *Let $\mathcal{S} = \langle \mathcal{K}, \mathcal{B}, \sigma, \preccurlyeq \rangle$ be a belief state. The **relevant expansion** of $\mathcal{S}$ with $\phi \in \mathcal{L}_0$ is a belief state $\mathcal{S} + \phi = \langle \mathcal{K}_{+\phi}, \mathcal{B}_{+\phi}, \sigma_{+\phi}, \preccurlyeq_{+\phi} \rangle$, with the following properties:*

(A^+1) *Success: $\mathcal{B}_{+\phi} = \mathcal{B} \cup \{\phi\}$.*
(A^+2) *Inclusion: $\mathcal{K} \subseteq \mathcal{K}_{+\phi}$.*
(A^+3) *Relevance: If $\psi \in \mathcal{K}_{+\phi} \backslash \mathcal{K}$, and $s \in \sigma_{+\phi}(\psi)$, then there is $s' \in \sigma_{+\phi}(\phi)$ such that $s' \subseteq s$.*
(A^+4) *Support update: If $\psi \in \mathcal{K}$ and $s \in \sigma_{+\phi}(\psi)$, then either $s \in \sigma(\psi)$ or there is $s' \in \sigma_{+\phi}(\phi)$ such that $s' \subseteq s$.*
(A^+5) *Order preservation: $\preccurlyeq_{+\phi}$ is the smallest total pre-order on $\mathcal{B}_{+\phi}$ satisfying*
   1. $\preccurlyeq \subseteq \preccurlyeq_{+\phi}$ and
   2. for every $\psi \in \mathcal{B}$, either $\phi \preccurlyeq \psi$ or $\psi \preccurlyeq \phi$.

Relevant expansion is simply assertion with forward inference. The belief state resulting from relevant expansion by ϕ will include ϕ and anything that follows from it. That all newly derived sentences indeed follow from ϕ is guaranteed by (A^+3), provided that ϕ was not derived in $\mathcal{K}$. In addition, old sentences may acquire new support only as a result of discovered derivations from ϕ ((A^+4)). It should be noted that, given certain constraints on Cn_R, the set $\mathcal{K}_{+\phi}$ is finite (provided that $\mathcal{K}$ is). (A^+5) makes the simplifying assumption that adding ϕ does not disturb the preference relations already established; ϕ simply gets added in some appropriate position in the $\preccurlyeq$-induced chain of equivalence classes.

Definition 4 *Let $\mathcal{S} = \langle \mathcal{K}, \mathcal{B}, \sigma, \preccurlyeq \rangle$ be a belief state. The **relevant revision** of $\mathcal{S}$ with $\phi \in \mathcal{L}_0$ is a belief state $\mathcal{S} \dot{+} \phi = \langle \mathcal{K}_{\dot{+}\phi}, \mathcal{B}_{\dot{+}\phi}, \sigma_{\dot{+}\phi}, \preccurlyeq_{\dot{+}\phi} \rangle$, with the following properties:*

($A^{\dot{+}}1$) *Base inclusion: $\mathcal{B}_{\dot{+}\phi} \subseteq \mathcal{B}_{+\phi}$.*
($A^{\dot{+}}2$) *Inclusion: $\mathcal{K}_{\dot{+}\phi} \subseteq \mathcal{K}_{+\phi}$.*
($A^{\dot{+}}3$) *Lumping: $\psi \in \mathcal{K}_{+\phi} \backslash \mathcal{K}_{\dot{+}\phi}$ if and only if, for every $s \in \sigma_{+\phi}(\psi)$, $s \not\subseteq \mathcal{B}_{\dot{+}\phi}$.*
($A^{\dot{+}}4$) *Preferential core-retainment: $\psi \in \mathcal{B}_{+\phi} \backslash \mathcal{B}_{\dot{+}\phi}$ if and only if $\bot \in \mathcal{K}_{+\phi}$ and $\psi \in \{x \mid \exists s \in \sigma_{+\phi}(\bot), x \in s, \text{ and } \forall y \in s, x \preccurlyeq_{+\phi} y\}$.*
($A^{\dot{+}}5$) *Support update: If $\psi \in \mathcal{K}_{\dot{+}\phi}$, then $\sigma_{\dot{+}\phi}(\psi)$ is the largest subset of $\sigma_{+\phi}(\psi)$ restricted to $\mathcal{B}_{\dot{+}\phi}$.*
($A^{\dot{+}}6$) *Order preservation: $\preccurlyeq_{\dot{+}\phi}$ is the largest subset of $\preccurlyeq_{+\phi}$ restricted to $\mathcal{B}_{\dot{+}\phi}$.*

Thus, relevant revision is assertion with forward inference followed by consolidation [29]. As a result of consolidation, some base beliefs might be retracted in case relevant expansion with ϕ results in a contradiction.[13] ($A^{\dot{+}}1$) captures this

[13] Technically, the contradiction need not be supported by ϕ. However, in practice, a reason maintenance system should not tolerate explicit contradiction (SNeBR does not). Thus, prior to relevant revision with ϕ, we may assume that no explicit contradiction was around, and, thus, that ϕ is somehow responsible for discovering/introducing a contradiction.

intuition. Since belief sets are not the logical closures of their bases, ($A^{\dot{+}}2$) does not necessarily follow from ($A^{\dot{+}}1$). It is needed to indicate that relevant revision does not result in derivations that are not accounted for by relevant expansion. ($A^{\dot{+}}3$) makes sure that only sentence that are still supported are believable.[14]

($A^{\dot{+}}4$) guarantees that base beliefs that are evicted to retain (explicit) consistency indeed must be evicted. In addition, if a choice is possible, base beliefs that are least preferred are chosen for eviction. Note that, according to the above definition, this selection strategy is *skeptical*; that is, if multiple least preferred beliefs exist, all are evicted. This strategy, however, is only adopted here to simplify the exposition, and nothing relevant depends on it.

As a simple corollary, it follows from ($A^{\dot{+}}3$) and ($A^{\dot{+}}4$) that the resulting belief state is not known to be inconsistent:

($A^{\dot{+}}7$) Contradiction ignorance: $\bot \notin \mathcal{K}_{\dot{+}\phi}$

4.3 The Case of $\mathcal{L}_1$

To extend the reason maintenance system presented above to $\mathcal{L}_1$, a number of superficial alterations of the definitions are needed. The important point, however, is to devise an extension of Cn_R that accommodates conditionals. Following [26], I am assuming a natural deduction system. Adding the connective $>$ to the language, we need two inference rules—one for elimination and one for introduction. First, a piece of notation.

Definition 5 *Let $\mathcal{S} = \langle \mathcal{K}, \mathcal{B}, \sigma, \preccurlyeq \rangle$ be a belief state. The **hypothetical expansion** of $\mathcal{S}$ with $\phi \in \mathcal{L}_1$ is a belief state $\mathcal{S} \mp \phi = \langle \mathcal{K}_{\mp\phi}, \mathcal{B}_{\mp\phi}, \sigma_{\mp\phi}, \preccurlyeq_{\mp\phi} \rangle$ where*

1. $\mathcal{B}_{\mp\phi} = \mathcal{B} \cup \{\phi\}$;
2. $\sigma_{\mp\phi}(\phi) = \{\{\phi\}\}$; *and*
3. *for every $\psi \in \mathcal{B}$, $\psi \preccurlyeq_{\mp\phi} \phi$ and $\phi \not\preccurlyeq_{\mp\phi} \psi$*

Hypothetical expansion (re)introduces ϕ into the belief state with independent standing as a most preferred belief. It is similar to the *do* operator of Pearl [31] in that it detaches ϕ from its derivational history (causal history, in the case Pearl). We now define the elimination and introduction rules for $>$ as follows.

($> E$) If $\phi, \phi > \psi \in \mathcal{K}$, then ψ may be added to $\mathcal{K}$, with $\sigma(\psi) = \{s_\phi \cup s_> \mid \langle s_\phi, s_> \rangle \in \sigma(\phi) \times \sigma(\phi > \psi)\}$.

($> I$) If $\psi \in (\mathcal{K}_{\mp\phi})_{\dot{+}\phi}$, then $\phi > \psi$ may be added to $\mathcal{K}$ provided that $\sigma(\phi > \psi) = \{s \backslash \{\phi\} \mid s \in (\sigma_{\mp\phi})_{\dot{+}\phi}(\psi) \backslash \sigma_{\mp\phi}(\psi)\}$ is not empty.

The elimination rule ($> E$) is a direct extension of Anderson and Belnap's rule for the elimination of material implication [26]. The introduction rule ($> I$) is also an extension of Anderson and Belnap's rule for the introduction of material implication. The extension in this case is, by no means, direct though.

[14] By "lumping", I'm referring to the lumping operation of Kratzer [30], whereby certain propositions either stay together or go together.

$(> I)$ is actually the right-to-left direction of the Ramsey test $(RT2)$, within the context of relevance logic. In simple English, the rule describes a procedure whereby one may decide whether to believe in the conditional $\phi > \psi$:

1. Hypothetically expand the belief state with ϕ.
2. Perform relevant forward inference to derive all sentences that could be derived from ϕ.
3. Consolidate the resulting belief state, giving ϕ highest preference.
4. If ψ is in the resulting belief state, accept $\phi > \psi$.

In addition to deciding on whether to accept $\phi > \psi$, we also compute its support sets along the way. The relevance of this derivation is guaranteed by two measures. The first is the hypothetical expansion step. The reason why we need this is that we need to make sure that any derivation of ψ following relevant expansion with ϕ follows from ϕ itself, not merely from its supports. The second is the procedure used to compute $\sigma(\phi > \psi)$: We only consider support sets that were added as a result of relevant expansion with ϕ. This eliminates cases where a conditional is only accepted as a result of its consequent being already in the belief set. The final removal of ϕ from the sets of supports is inherited from Anderson and Belnap's rule for material implication introduction.

By adding $(> E)$ and $(> I)$ to our repertoire of inference rules, we define an extension $\mathrm{Cn}_{R>}$ of Cn_R for relevant conditional consequence. All the definitions of Section 4.2 may now be extended to $\mathcal{L}_0$ by replacing each occurrence of $\mathcal{L}_0$ by $\mathcal{L}_1$, and each occurrence of Cn_R by $\mathrm{Cn}_{R>}$. This may be considered overly permissive by many scholars. For, now, we allow two things that are traditionally not allowed.

We allow, *pace* Hansson [13], (i) belief bases to include conditional sentences and (ii) belief states to be revised with conditional sentences. The justification for this is the same: we view conditionals to possibly have independent standing. For a rational agent or knowledge representation and reasoning system, this is actually reasonable. The following example of a "useful counterfactual" is due to Costello and McCarthy [2, p. 1].

(1) If another car had come over the hill when you passed that car, there would have been a head-on collision

A natural context in which (1) may be uttered is one in which the speaker is teaching someone an important safety rule of car driving. Most probably, the person receiving this utterance does not have enough experience to have concluded it themselves. The person can learn from (1), though. And they can do that without ever trying to achieve its antecedents. Thus, it seems that in similar cases, the only reasonable thing to assume is that the conditional is a base belief that probably could not have been derived without external help.

5 The Trap Reentered

What can we say now about the Ramsey test and Gärdenfors's triviality result? First, as $(> I)$ indicates, only one direction of the Ramsey test is adopted; we

use it as a rule of introduction for conditionals. A direct reversal of $(> I)$ would yield.

(RRT1) If $\phi > \psi$ may be added to $\mathcal{K}$ with
$\sigma(\phi > \psi) = \{s \backslash \{\phi\} \mid s \in (\sigma_{\mp\phi})_{\dot{+}\phi}(\psi) \backslash \sigma_{\mp\phi}(\psi)\}$, then $\psi \in (\mathcal{K}_{\mp\phi})_{\dot{+}\phi}$ and $\{s \backslash \{\phi\} \mid s \in (\sigma_{\mp\phi})_{\dot{+}\phi}(\psi) \backslash \sigma_{\mp\phi}(\psi)\}$ is not empty.

This does not look right. For what does it mean to say that $\phi > \psi$ *may be* added to $\mathcal{K}$? Note that it does not mean that $\phi > \psi$ is *in* $\mathcal{K}$, for $\mathcal{K}$ is not logically closed. The primary way, within our system, we can make sure that $\phi > \psi$ *may be* added to $\mathcal{K}$ is probably by using $(> I)$. But, then, *(RRT1)* will not be very useful; it is only telling us something that we already know about $\mathrm{Cn}_{R>}$.

So maybe we can simply replace "may be added to $\mathcal{K}$" with "is in $\mathcal{K}$". But, then, the first antecedent $(\phi > \psi \in \mathcal{K})$ would imply the second consequent $(\sigma(\phi > \psi) \neq \varnothing)$, and the second antecedent (the definition of $\sigma(\phi > \psi)$) would imply the first consequent $(\psi \in (\mathcal{K}_{\mp\phi})_{\dot{+}\phi})$. Again, this version of *RRT1* does not seem useful.

Finally, we can simply drop all mention of supports and state that $\phi > \psi \in \mathcal{K}$ if and only if $\psi \in (\mathcal{K}_{\mp\phi})_{\dot{+}\phi}$. This is far from being the converse of $(> I)$. In addition, I do not see how useful it may be, at least from the point of view of rational agency or a KRR system. If anything, it may save us some time by caching one of the results of hypothetically revising with ϕ (we can also easily reconstruct the supports of ψ in the resulting belief state). This, however, is not unproblematic. For even assuming $\phi > \psi \in \mathcal{K}$, it need not be the case that $\psi \in (\mathcal{K}_{\mp\phi})_{\dot{+}\phi}$. This may happen, for example, if deriving ψ using $(> E)$ may result in a contradiction. Considering the following variant of Example 1 may clarify this point (where $\phi = M$ and $\psi = \neg J$).

Example 2. The supervisor believes (probably by default) that only one person committed the murder: $\{J > \neg M, M > \neg J\}$. Detective 1 reports evidence incriminating John. This results in adding the beliefs J and $\neg M$ to the supervisor's belief set. Now detective 2 reports evidence incriminating Mary. At this point, the supervisor needs to do some consolidation. Assuming that the evidence provided by both detectives is highly reliable, the supervisor would disbelieve $J > \neg M$, rendering $\neg M$ no longer supported. In addition, note that $M > \neg J$ also needs to be removed, to block the derivation of $\neg J$. Thus, when faced with strong evidence to the contrary, the supervisor gives up the belief in a single murderer.

Now, even if we assume some reasonable converse of $(> I)$, the triviality proof will not go through. Any step in the proof that depends on closure $(A^{\mathcal{K}}3)$, success (A^*1), or consistency (A^*2) will be invalid. In addition, since revisions are now based on $\mathrm{Cn}_{R>}$, step 8 is obviously invalid. For, even if we admit closure, $\mathrm{Cn}_{R>}(\mathcal{K} \cup \{A \vee B\}) \not\subseteq \mathrm{Cn}_{R>}(\mathcal{K} \cup \{A\})$ (since $A \vee B \notin \mathrm{Cn}_{R>}(\{A\})$).

In addition, Hansson [13, p. 531–532] argued that even if two belief sets are identical, base-revising them may yield different results. Thus any step in the

proof that relies on the equality of revising two identical belief sets will be invalid (for example, step 13).

Example 3 [13, p. 531–532]. Let $\mathcal{S}_1$ be a belief state with $\mathcal{B}_1 = \{p\}$. Consider, $\mathcal{S}_2 = \mathcal{S}_1 \dot{+} q$ and $\mathcal{S}_3 = \mathcal{S}_1 \dot{+} p \Leftrightarrow q$. Clearly, $\mathcal{B}_2 = \{p, q\}$, $\mathcal{B}_3 = \{p, p \Leftrightarrow q\}$, and $\mathrm{Cn}_{R>}(\mathcal{B}_2) = \mathrm{Cn}_{R>}(\mathcal{B}_3)$. In particular, if $p \Leftrightarrow q \in \mathcal{K}_2$ then $\sigma_2(p \Leftrightarrow q) = \{\{p, q\}\}$. similarly, if $q \in \mathcal{K}_3$, then $\sigma_3(q) = \{\{p, p \Leftrightarrow q\}\}$. Now consider $\mathcal{S}_2 \dot{+} \neg p$ and $\mathcal{S}_3 \dot{+} \neg p$, assuming $p \preccurlyeq \neg p$ in both $\mathcal{S}_2$ and $\mathcal{S}_3$. Given Definition 4, q is in the first belief set and $\neg q$ is derivable in the second.

Finally, similar to [15], the proof of (M) will be blocked due to the context-sensitivity of conditionals. In our system, context-sensitivity is defined by the dependence of $> I$ (and its presumed converse) on sets of supports. Even with closure, success, and consistency reinstated, similar to [15], the proof of (M) will be blocked at step 5:

1. $\mathcal{K} \subseteq \mathcal{K}'$ (Assumption)
  2. $\psi \in (\mathcal{K}_{\mp\phi})_{\dot{+}\phi}$ and $\{s \backslash \{\phi\} \mid s \in (\sigma_{\mp\phi})_{\dot{+}\phi}(\psi) \backslash \sigma_{\mp\phi}(\psi)\} \neq \varnothing$ (Assumption)
    3. $\phi > \psi \in \mathcal{K}$
      with $\sigma(\phi > \psi) = \{s \backslash \{\phi\} \mid s \in (\sigma_{\mp\phi})_{\dot{+}\phi}(\psi) \backslash \sigma_{\mp\phi}(\psi)\}$ (2, $(> I)$)
    4. $\phi > \psi \in \mathcal{K}'$
      with $\sigma'(\phi > \psi) = \{s \backslash \{\phi\} \mid s \in (\sigma_{\mp\phi})_{\dot{+}\phi}(\psi) \backslash \sigma_{\mp\phi}(\psi)\}$ (1, 3)

At this point we are stuck; we cannot prove $\psi \in (\mathcal{K}'_{\mp\phi})_{\dot{+}\phi}$ since this requires σ', and not σ, in the definition of the support of $\phi > \psi$. For instance, in Example 2, it is clear that the set of supports of $M > \neg J$ in $\mathcal{K} = \{J > \neg M, M > \neg J\}$ is different from that in $\mathcal{K}' = \mathcal{K} \cup \{J\}$, where in the former it is simply $\{\{M > \neg J\}\}$ and in the latter it is the empty sets.

6 Conclusions

Based on [5, 9], I have presented a reason maintenance system with an underlying relevance logic. I have shown how such a system could be extended to accommodate a relevance-logical account of conditionals. One direction of the Ramsey test is used as a conditional-introduction inference rule. The rule defines the set of supports of the derived conditional in such a way that effects context-sensitivity as per [15]. Given independently motivated assumptions on the underlying logic, the belief state, and the belief revision operator, Gärdenfors's triviality result is avoided.

The system presented here is similar to that of [13] in its use of base revision. It is different, however, in allowing conditional sentences to be members of belief bases and candidates for expansion and revision. Compared to the system of [16], the revision-based approach presented here is a practical alternative to the update-based approach of [16]. It is my conviction that, ultimately, both revision and update are needed for reasoning about conditionals. In particular,

the relation between indicative conditionals and belief revision on one hand, and subjunctive conditionals and belief update on the other hand, remains to be investigated.

References

1. Stalnaker, R.: A theory of conditionals. In Rescher, N., ed.: Studies in Logical Theory. Number 2 in American Philosophical Quarterly, Monograph Series. Basil Blackwell Publishers, Oxford (1968) 98–112
2. Costello, T., McCarthy, J.: Useful counterfactuals. Linköping Electronic Articles in Computer and Information Science **4**(12) (1999)
3. Ortiz, C.: Explanatory update theory: Applications of counterfactual reasoning to causation. Artificial Intelligence **108**(1–2) (1999) 125–178
4. Doyle, J.: A truth maintenance system. Artificial Intelligence **12**(3) (1979) 231–272
5. Martins, J., Shapiro, S.C.: A model for belief revision. Artificial Intelligence **35**(1) (1988) 25–79
6. Alchourron, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: Partial meet contraction and revision functions. The Journal of Symbolic Logic **50**(2) (1985) 510–530
7. Nebel, B.: A knowledge level analysis of belief revision. In Brachman, R., Levesque, H., Reiter, R., eds.: Principles of Knowledge Representation and Reasoning: Proceedings of the First International Conference (KR'89), Toronto, ON, Morgan Kaufmann (1989) 301–311
8. Williams, M.A.: Transmutations of knowledge systems. In Doyle, J., Sandewall, E., Torasso, P., eds.: Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference (KR'94), San Francisco, CA, Morgan Kaufmann (1994) 619–629
9. Johnson, F.L., Shapiro, S.C.: Dependency-directed reconsideration: Belief base optimization for truth maintenance systems. In: Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05). (2005) 313–320
10. Gärdenfors, P.: Belief revisions and the Ramsey test for conditionals. The Philosophical Review **95**(1) (1986) 81–93
11. Gärdenfors, P.: Variations on the Ramsey test: More triviality results. Studia Logica **46** (1987) 321–327
12. Rott, H.: Conditionals and theory change: Revisions, expansions, and additions. Synthese **81** (1989) 91–113
13. Hansson, S.O.: In defence of the Ramsey test. The Journal of Philosophy **89**(10) (1992) 522–540
14. Arló Costa, H., Levi, I.: Two notions of epistemic validity: Epistemic models for Ramsey's conditionals. Synthese **109** (1996) 217–262
15. Lindström, S., Rabinowicz, W.: The Ramsey test revisited. In Grocco, G., Fariñas del Cerro, L., Herzig, A., eds.: Conditionals: From philosophy to computer science. Oxford University Press, Oxford, UK (1996) 147–191
16. Grahne, G.: Updates and counterfactuals. Journal of Logic and Computation **8**(1) (1998) 87–117
17. Giordano, L., Gliozzi, V., Olivetti, N.: Belief revision and the Ramsey test: A solution. In: AI*IA 2001: Advances in Artificial Intelligence : 7th Congress of the Italian Association for Artificial Intelligence. Number 2175 in LNAI, Berlin/Heidelberg, Springer-Verlag (2001) 165–175

18. Shapiro, S.C., Rapaport, W.J.: SNePS considered as a fully intensional propositional semantic network. In Cercone, N., McCalla, G., eds.: The Knowledge Frontier. Springer-Verlag, New York (1987) 263–315
19. Shapiro, S.C.: SNePS: A logic for natural language understanding and commonsense reasoning. In Iwańska, Ł.M., Shapiro, S.C., eds.: Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language. AAAI Press/The MIT Press, Menlo Park, CA (2000) 175–195
20. Johnson, F.L., Shapiro, S.C.: Automatic belief revision in SNePS. In Baral, C., Truszczynski, M., eds.: Proceedings of the 8th International Workshop on Non-Monotonic Reasoning (NMR2000). (2000) Unpaginated, 5 pages.
21. Segerberger, K.: A note on an impossibility theorem of Gärdenfors. Noûs **23**(3) (1989) 351–354
22. Makinson, D.: The Gärdenfors impossibility theorem in non-monotonic contexts. Studia Logica **49**(1) (1990) 1–6
23. Darwiche, A., Pearl, J.: On the logic of iterated belief revision. Artificial Intelligence **89** (1997) 1–29
24. Katsuno, H., Mendelzon, A.: On the difference between updating a knowledge base and revising it. In: Principles of Knowledge Representation and Reasoning: Proceedings of the second International Conference (KR'91). (1991) 387–394
25. Levi, I.: Iteration of conditionals and the Ramsey test. Synthese **76** (1988) 49–81
26. Anderson, A., Belnap, N.: Entailment. Volume I. Princeton University Press, Princeton, NJ (1975)
27. Shapiro, S.C.: Relevance logic in computer science. In Anderson, A., Belnap, N., Dunn, M., eds.: Entailment. Volume II. Princeton University Press, Princeton, NJ (1992) 553–563
28. Hansson, S.O.: Kernel contraction. Journal of Symbolic Logic **59**(3) (1994) 845–859
29. Hansson, S.O.: A survey of non-prioritized belief revision. Erkenntnis **50**(2–3) (1999) 413–427
30. Kratzer, A.: An investigation into the lumps of thougth. Linguistics and Philosophy **12** (1989) 607–653
31. Pearl, J.: Causality. Cambridge University Press, New York (2000)

Subjective Models and Multi-agent Static Belief Revision

Guillaume Aucher*

Universit Paul Sabatier – University of Otago
IRIT – Équipe LILaC
118 route de Narbonne
F-31062 Toulouse cedex 9 (France)
aucher@irit.fr

Abstract. We generalize the AGM belief revision theory to the multi-agent case. To do so, we first generalize the semantics of the single-agent case, based on the notion of interpretation, to the multi-agent case. We also compare this generalized semantics, based on the notion of subjective model, to the standard semantics of multi-agent epistemic logic. Then we show that, thanks to the shape of our new semantics, all the results of the AGM framework transfer to the multi-agent case. Afterwards we investigate some postulates specific to our multi-agent setting. Finally, we give an example of revision operator that fulfills one of these new postulates and give an example of revision on a concrete example.

Note 1. An extended version of this paper with full proofs can be found at the following address: ftp://ftp.irit.fr/pub/IRIT/LILAC/rap-IRIT-RR-2007-20-EN.pdf.

1 Introduction

AGM belief revision theory [1] has been designed for a single agent. It seems natural to extend it to the multi-agent case. In this case the agent at stake (that we call Y like You), in his/her representation of the surrounding world, will have to deal not only with facts about the world but also with how the other agents perceive the surrounding world. So, we will have to extend or generalize the single agent semantics in order to take into account this multi-agent aspect.

Besides, we have to be careful about what kind of multi-agent belief revision we study. Indeed, we must make a distinction (also made in [4]) between multi-agent *dynamic* belief revision and multi-agent *static* (or *private*) belief revision. On the one hand, dynamic belief revision occurs after (or during) an event took place that changes the original situation and the other agents' beliefs. This is the case of a public announcement for example that often affects and changes all the

* I thank my PhD supervisors Hans van Ditmarsch and Andreas Herzig for useful comments and discussions. I also thank Jérôme Lang and an anonymous referee for their comments on this paper.

agents' beliefs. On the other hand, static belief revision occurs when the agent Y learns some piece of information about the original situation but this original situation and the other agents' beliefs do not actually change. Typically, this is the case when Y learns *privately* (from an external source for example) some piece of information (possibly epistemic) about the original situation, the other agents not being aware of anything. In this case, the other agents' beliefs clearly do not change. For example, suppose *y*ou (Y) believe p, and agent j believes p (and perhaps even that p is common belief of Y and j); when a third external agent privately tells you that $\neg p$ then you still believe that j believes p (and that j believes that p is common belief). The multi-agent dynamic case has received a lot of attention in what is commonly called dynamic epistemic logic ([5], [11], or [2] for instance). In this paper we study the static case; this static aspect enables us to use the standard methods of AGM belief revision theory.

The paper is organized as follows. In Section 2, we recall belief revision theory in the line of [8]. In Section 3, we first introduce the notions of multi-agent possible worlds and subjective models in order to adequately represent agent Y's perception of the surrounding world. Then we generalize the results of Section 2 to the multi-agent case. In Section 4, we investigate some additional rationality postulates specific to our multi-agent approach. Finally in Section 5, we give an example of revision operator and an application of this operator to a concrete example.

2 The Single Agent Case: the AGM Approach

In this paper Φ is a *finite* set of propositional letters and L the propositional language defined over Φ. We prefer to follow the knowledge base approach of [8] as it is easier to handle by computers. As argued by Katsuno and Mendelzon, because Φ is finite, a belief set K can be equivalently represented by a mere propositional formula ψ. Then $\varphi \in K$ iff $\psi \rightarrow \varphi$.

Lemma 1. *[8] Let * be a revision operator on knowledge sets and $\circ$ its corresponding operator on belief bases (i.e. $\psi \circ \mu$ implies φ iff $\varphi \in Cn(\psi) * \mu$). Then * satisfies the 8 AGM postulates $(K*1)-(K*8)$ iff $\circ$ satisfies the postulates $(R1)-(R6)$ below:*

(R1) $\psi \circ \mu \rightarrow \mu$.
(R2) *if* $\psi \wedge \mu$ *is satisfiable, then* $\psi \circ \mu \leftrightarrow \psi \wedge \mu$.
(R3) *If* μ *is satisfiable, then* $\psi \circ \mu$ *is also satisfiable.*
(R4) *If* $\psi_1 \leftrightarrow \psi_2$ *and* $\mu_1 \leftrightarrow \mu_2$, *then* $\psi_1 \circ \mu_1 \leftrightarrow \psi_2 \circ \mu_2$.
(R5) $(\psi \circ \mu) \wedge \varphi \rightarrow \psi \circ (\mu \wedge \varphi)$.
(R6) *If* $(\psi \circ \mu) \wedge \varphi$ *is satisfiable, then* $\psi \circ (\mu \wedge \varphi) \rightarrow (\psi \circ \mu) \wedge \varphi$.

Let $\mathcal{I}$ be the set of all the interpretations of the finite propositional language L. $Mod(\psi)$ denotes the set of all the interpretations that make ψ true. Let $\mathcal{M}$ be a set of interpretations of L. $form(\mathcal{M})$ denotes a formula whose set of models is equal to $\mathcal{M}$.

A *pre-order* $\leq$ over $\mathcal{I}$ is a reflexive and transitive relation on $\mathcal{I}$. A pre-order is *total* if for every $I, J \in \mathcal{I}$, either $I \leq J$ or $J \leq I$. Consider a function that assigns to each propositional formula ψ a pre-order $\leq_\psi$ over $\mathcal{I}$. We say this assignment is *faithful* if the following three conditions hold:

1. If $I, I' \in Mod(\psi)$, then $I <_\psi I'$ does not hold.
2. If $I \in Mod(\psi)$ and $I' \notin Mod(\psi)$, then $I <_\psi I'$ holds.
3. If $\psi \leftrightarrow \varphi$, then $\leq_\psi = \leq_\varphi$.

Let $\mathcal{M}$ be a subset of $\mathcal{I}$. An interpretation I is minimal in $\mathcal{M}$ with respect to $\leq_\psi$ if $I \in \mathcal{M}$ and there is no $I' \in \mathcal{M}$ such that $I' <_\psi I$. Let

$$Min(\mathcal{M}, \leq_\psi) := \{I; I \text{ is minimal in } \mathcal{M} \text{ with respect to } \leq_\psi\}$$

Theorem 1. *[8] Revision operator $\circ$ satisfies conditions $(R1) - (R6)$ iff there exists a faithful assignment that maps each knowledge base ψ to a total pre-order $\leq_\psi$ such that $Mod(\psi \circ \mu) = Min(Mod(\mu), \leq_\psi)$.*

3 The Multi-agent Case

3.1 Some Technical Preliminaries

In the sequel, G is a fixed set of agents such that $Y \in G$. An epistemic model M is a tuple $M = (W, \{R_j; j \in G\}, val)$ where W is a set of worlds, R_j are accessibility relations indexed by agents $j \in G$ and val is a function that assigns to each $w \in W$ a subset of Φ. A $KD45_G$ epistemic model is an epistemic model whose accessibility relations are serial, transitive and euclidean. Classically an epistemic model M is given with an actual world w_a: (M, w_a). We define $R_j(w)$ by $R_j(w) := \{v; wR_jv\}$ and say that M is generated from w if $W = (\bigcup_{j \in G} R_j)^*(w)$. Finally $|M|$ is the number of worlds in M. The epistemic language is defined by $\mathcal{L} : \varphi := \bot | p | \neg\varphi | \varphi \wedge \varphi | B_j\varphi | C_{G_1}\varphi$, $j \in G, G_1 \subseteq G, p \in \Phi$.

Its semantics is defined as usual by: $M, w \models p$ iff $p \in val(w)$; $M, w \models B_j\varphi$ iff for all $v \in R_j(w)$ $M, v \models \varphi$; $M, w \models C_{G_1}\varphi$ iff for all $v \in (\bigcup_{j \in G_1} R_j)^*(w)$ $M, v \models \varphi$. We now recall the definition of a bisimulation.

Definition 1. *Let Z be a relation between two finite epistemic models $M = (W, \{R_j; j \in G\}, val)$ and $M' = (W', \{R'_j; j \in G\}, val')$. We define the property of Z being a bisimulation in w and w', noted $Z : M, w \leftrightarrows M', w'$ as follows.*

1. *If wZw' then $val(w) = val'(w')$;*
2. *if wZw' and $v \in R_j(w)$ then there exists $v' \in R_j(w')$ such that vZv';*
3. *if wZw' and $v' \in R_j(w'$ then there exists $v \in R_j(w)$ such that vZv'.*

We can define bisimilarity between M, w and M', w', noted $M, w \leftrightarrows M', w'$ by $M, w \leftrightarrows M', w'$ iff there is a relation Z such that $Z : M, w \leftrightarrows M', w'$. It can be shown (in case M and M' are finite) that $M, w \leftrightarrows M', w'$ iff for all $\varphi \in \mathcal{L}$, $M, w \models \varphi$ iff $M', w' \models \varphi$. So, intuitively, two epistemic models are bisimilar if they contain the same information.

Proposition 1. *[3][10] Let M be a finite epistemic model and $w \in M$. Then there is an epistemic formula $\delta_M(w)$ (involving common knowledge) such that*

1. $M, w \models \delta_M(w)$
2. *For every finite epistemic model M' and world $w' \in M'$, if $M', w' \models \delta_M(w)$ then $M, w \leftrightarrows M', w'$.*

This proposition tells us that a finite model can be completely characterized (modulo bisimulation) by an epistemic formula. It will be very useful to prove that the results of the single agent case transfer to the multi-agent case[1].

3.2 Possible World Versus Multi-agent Possible World

The notion of multi-agent possible world In the AGM framework, one considers a single agent Y. The possible worlds introduced are supposed to represent how the agent Y perceives the surrounding world. Because she is the only agent, these possible worlds deal only with propositional facts about the surrounding world. Now, because we suppose that there are other agents than agent Y, a possible world for Y in that case should also deal with how the other agents perceive the surrounding world. These "multi-agent" possible worlds should then not only deal with propositional facts but also with epistemic facts. So to represent a multi-agent possible world we need to introduce a modal structure to our possible worlds. We do so as follows.

Definition 2 (multi-agent possible world).
A multi-agent possible world (M, w) is a finite epistemic model $M = (W, \{R_j; j \in G\}, val)$ such that for all j, R_j is serial, transitive and euclidean, and

- *$R_Y(w) = \{w\}$;*
- *there is no v and $j \neq Y$ such that $w \in R_j(v)$.*

Our definition is defined in such a way that in case Y is the only agent then a multi-agent possible world boils down to an interpretation. The first condition ensures us that in case Y assumes that she is in the multi-agent possible world (M, w) then for her w is the only possible world. The second condition will be explained in the next paragraph. Note that if we remove the constraints on the accessibility relations (seriality, euclidicity and transitivity) the results in this paper are still valid. We prefer to keep them because we find them more intuitive to model the notion of belief.

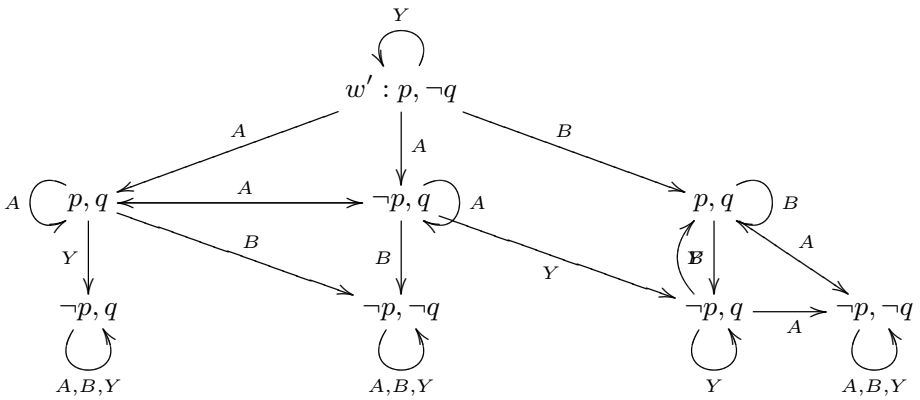
Definition 3 (subjective model). *A subjective model is a finite set of multi-agent possible worlds.*

[1] Note that van Benthem, in [10], already mentioned that this proposition could be used in belief revision theory

a (single-agent) possible world:

$$w : p, \neg q$$

a multi-agent possible world:



Note that in the single-agent case, a subjective model boils down to a (non-empty) set of interpretations, so represents a belief set. Intuitively, a subjective model is the formal model that agent Y has "in her head" and that represents how she perceives the surrounding world. This interpretation of our formalism differs from epistemic models (M, w_a), usually encountered in epistemic logic, which are supposed to represent objectively and from an external point of view how all the agents perceive the actual world w_a. We can nevertheless draw a *formal* parallel between the two formalisms (keeping in mind that their interpretation is different).

From subjective models to epistemic models, and vice versa

Let $\{(M_1, w_1), \ldots, (M_n, w_n)\}$ be a subjective model. The epistemic model associated to $\{(M_1, w_1), \ldots, (M_n, w_n)\}$ is the $KD45_G$ epistemic model $M = (W, R_j, val)$ defined as follows. $W := W_1 \cup \ldots \cup W_n$; $R_j := R_j^1 \cup \ldots \cup R_j^n$ for $j \neq Y$; $R_Y := R_Y^1 \cup \ldots \cup R_Y^n \cup \{(w_i, w_k); i, k = 1 \ldots n\}$; and $val(w) := val_i(w)$ if $w \in W_i$. We can now motivate the second item of Definition 2. Indeed, if this item was not fulfilled then it might be possible that j's beliefs about Y's beliefs (for some $j \neq Y$) might be different in the subjective model and in the associated epistemic model, due to the creation of new links between the multi-agent possible worlds.

Example 1. In Figure 1 is represented the subjective model $\{(M_1, w), (M_2, v)\}$ and in Figure 2 is represented an epistemic model bisimilar to the epistemic model associated to $\{(M_1, w), (M_2, v)\}$.

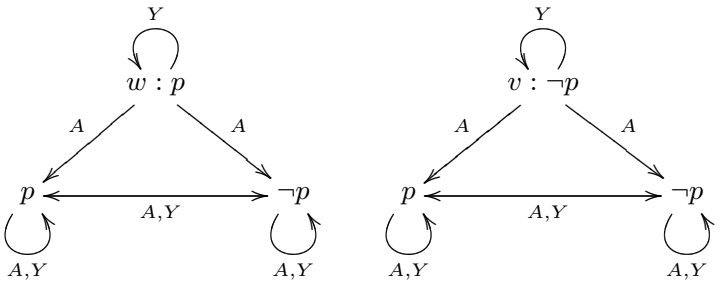


Fig. 1. A subjective model : multi-agent possible world (M_1, w) (left) and multi-agent possible world (M_2, v) (right)

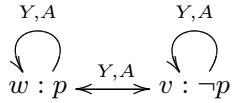


Fig. 2. (Epistemic model bisimilar to) the epistemic model associated to $\{(M_1, w), (M_2, v)\}$

Hence a subjective model can be represented equivalently by a $KD45_G$ epistemic model[2]. The other way round, one can easily show that any epistemic model that is generated from $R_Y(w_a)$, where w_a is the actual world, can be represented equivalently by a subjective model. So it turns out that epistemic models can model the same things as subjective models do. But the shape of our semantics, based on the notion of multi-agent possible world, allows to generalize easily concepts and methods from AGM belief revision theory, as we will now see.

3.3 The Multi-agent Generalization of the AGM Approach

In the multi-agent case like in the single-agent case, it does not make any sense to revise by formulas dealing with what the agent Y believes or considers possible. Indeed, due to the fact that positive and negative introspection are valid in KD45, Y already knows all she believes and all she disbelieves. So we restrict the epistemic language to a fragment that we call $\mathcal{L}_{j\neq Y}$ defined as follows.

Definition 4. $\mathcal{L}_{j\neq Y} : \varphi := \bot|p|B_j\psi|C_{G_1}\psi|\varphi \wedge \varphi|\neg\varphi$, $\psi \in \mathcal{L}$, $j \neq Y$, $Y \notin G_1 \subseteq G$

We can then apply with some slight modifications the procedure spelled out for the single agent case in Section 2.

First the postulates for multi-agent belief revision are identical to the ones spelled out in Lemma 1 but this time ψ, μ and φ belong to $\mathcal{L}_{j\neq Y}$. Now we define

[2] This equivalence could be easily specified formally by stating that for all i and $\varphi \in \mathcal{L}_{j\neq Y}$, $M, w_i \models \varphi$ iff $M_i, w_i \models \varphi$ (see Definition 5 below).

$\mathcal{I}_G$ to be the set of all multi-agent possible worlds modulo bisimulation, i.e. we pick the smallest multi-agent possible world among each class of bisimilarly indistinguishable multi-agent possible worlds. We define $Mod(\psi)$ by $Mod(\psi) := \{(M,w); (M,w) \in \mathcal{I}_G \text{ and } M,w \models \psi\}$. Let $\mathcal{M}$ be a subjective model. Thanks to Proposition 1 we can easily prove that

Fact (*) there is a formula $form(\mathcal{M}) \in \mathcal{L}_{j \neq Y}$ such that $Mod(form(\mathcal{M})) = \mathcal{M}$.

We then get the multi-agent generalization of Theorem 1 by replacing interpretations I by multi-agent possible worlds (M,w). The proof is completely similar to the single agent case and relies heavily on the fact (*) proved thanks to Proposition 1.

Theorem 2. *Revision operator $\circ$ on $\mathcal{L}_{j \neq Y}$ satisfies conditions $(R1) - (R6)$ iff there exists a faithful assignment that maps each knowledge base ψ to a total pre-order $\leq_\psi$ such that $Mod(\psi \circ \mu) = Min(Mod(\mu), \leq_\psi)$.*

Remark 1 (important). We have picked only one of the theorems of [8] but in fact all the theorems present in [8] transfer to the multi-agent case. It includes in particular the theorem about $\leq_l$ being a partial order instead of a total order.

In summary, the concept of subjective model allows for a straightforward transfer of the AGM framework and results.

4 Some Considerations Specific to our Multi-agent Approach

In this section we are going to investigate some multi-agent rationality postulates. Indeed, because we add a multi-agent structure to our possible worlds, it is natural to study how the other agents' beliefs evolve during a revision process.

As said in the introduction, multi-agent static revision amounts to make a private announcement to Y, the other agents not being aware of anything. So, in particular, the beliefs of the agents who are not concerned by the private announcement should not change. We first need to define formally the agents who are concerned by a formula.

4.1 On the kind of information a formula is concerned about

First note that an input may not only concern agents but also the objective state of nature, i.e. *p*ropositional *f*acts, that we note pf. For example, the formula $p \wedge B_j B_i \neg p$ concerns agent j's beliefs but also propositional facts (namely p). Besides, a formula cannot be about Y's beliefs because $\varphi \in \mathcal{L}_{j \neq Y}$ by assumption. So what an input is about includes propositional facts but excludes agent Y's beliefs. This leads us to the following definition.

Definition 5. $C := (G \cup \{pf\}) - \{Y\}$.

We define by induction the agents who are concerned by a formula as follows:

– $C(p) := pf$; $C(B_j\varphi) := \{j\}$; $C(C_{G_1}\varphi) := G_1$;
– $C(\neg\varphi) := C(\varphi)$; $C(\varphi \wedge \varphi') := C(\varphi) \cup C(\varphi')$.

For example, $C(p \vee (q \wedge B_j B_i r) \wedge B_k r)) = \{pf, j, k\}$, and $C(B_i p \vee B_j B_k \neg p) = \{i, j\}$. We then define a language $\mathcal{L}_{C_0}$ whose formulas concern only agents in C_0, and possibly propositional facts if $pf \in C_0$.

Definition 6. *Let $C_0 \subseteq C$. We define the language $\mathcal{L}_{C_0}$ as follows.*

$\varphi :=\bot \mid A \mid B_j\psi \mid C_{G_1}\varphi \mid \varphi \wedge \varphi \mid \neg\varphi$, $j \in C_0, G_1 \subseteq C_0$, $\psi \in \mathcal{L}$,

with $A = \Phi$ if $pf \in C_0$ and $A = \emptyset$ otherwise.

Now we define a notion supposed to tell us whether two pointed and finite epistemic models contain the same information about some agents' beliefs and possibly about propositional facts.

Definition 7. *Let $C_0 \subseteq C$. We say that (M, w) and (M', w') are C_0-bisimilar, noted $M, w \leftrightarrows_{C_0} M', w'$, iff*

– *if $pf \in C_0$ then $val(w) = val(w')$ and*
– *for all $j_0 \in C_0$,*
  if $v \in R_{j_0}(w)$ then there is $v' \in R_{j_0}(w')$ such that $M, v \leftrightarrows M', v'$,
  if $v' \in R_{j_0}(w')$ then there is $v \in R_{j_0}(w)$ such that $M, v \leftrightarrows M', v'$.

Proposition 2. *Let $C_0 \subseteq C$. Then $M, w \leftrightarrows_{C_0} M', w'$ iff for all $\varphi \in \mathcal{L}_{C_0}$, $M, w \models \varphi$ iff $M', w' \models \varphi$.*

Proposition 2 ensures us that the notion we just defined captures what we wanted. We then have a counterpart of Proposition 1.

Proposition 3. *Let $C_0 \subseteq C$, let M be a finite epistemic models and $w \in M$. Then there is $\delta_M^{C_0}(w)$ such that*

1. *$M, w \models \delta_M^{C_0}(w)$*
2. *for every finite epistemic model M' and world $w' \in M'$, if $M', w' \models \delta_M^{C_0}$ then $M, w \leftrightarrows_{C_0} M', w'$*

Definition 8. *Let $\mathcal{M}$ and $\mathcal{M}'$ be two sets of multi-agent possible worlds, we set $\mathcal{M} \leftrightarrows_{C_0} \mathcal{M}'$ iff for all $(M, w) \in \mathcal{M}$ there is $(M', w') \in \mathcal{M}'$ such that $M, w \leftrightarrows_{C_0} M', w'$, and for all $(M', w') \in \mathcal{M}'$ there is $(M, w) \in \mathcal{M}$ such that $M, w \leftrightarrows_{C_0} M', w'$.*

4.2 Some Postulates Specific to our Multi-agent Approach

As we said before, static revision amounts to a private announcement to Y, the other agents not being aware of anything. So, in particular, Y's beliefs about the beliefs of the agents who are not concerned by the formula should not change. This can be captured by the following postulate:

(R7) Let $\varphi, \varphi' \in \mathcal{L}_{j \neq Y}$ such that $C(\varphi) \cap C(\varphi') = \emptyset$.
If $\psi \rightarrow \varphi'$ then $\psi \circ \varphi \rightarrow \varphi'$

This postulate is the multi-agent version of Parikh and Chopra's postulate [6]. The example of the introduction illustrates this postulate: there $\varphi = \neg p$ and $\varphi' = B_j p \wedge B_j C_G p$. Now the semantic counterpart of $(R7)$:

Proposition 4. *Revision operator $\circ$ satisfies $(R7)$ iff for all $\varphi \in \mathcal{L}_{j \neq Y}$, for all $(M', w') \in Mod(\psi \circ \varphi)$ there is $(M, w) \in Mod(\psi)$ such that $M, w \leftrightarrows_{C'} M', w'$, with $C' := C - C(\varphi)$.*

Another interesting postulate is the following.

(R8) Let $\varphi, \varphi' \in \mathcal{L}_{j \neq Y}$ such that $C(\varphi) \cap C(\varphi') = \emptyset$.
If $\psi \wedge \varphi'$ is satisfiable then $(\psi \circ \varphi) \wedge \varphi'$ is satisfiable.

And the semantic counterpart:

Proposition 5. *Revision operator $\circ$ satisfies $(R8)$ iff for all $\varphi \in \mathcal{L}_{j \neq Y}$, for all $(M, w) \in Mod(\psi)$ there is $(M', w') \in Mod(\psi \circ \varphi)$ such that $M, w \leftrightarrows_{C'} M', w'$, with $C' := C - C(\varphi)$.*

Note that $(R8)$ is the converse of $(R7)$. Unlike $(R7)$, $(R8)$ is not really suitable for revision because all the worlds representing Y's epistemic state "survive" revision process if $(R8)$ is fulfilled. This is not the case in general because new information can discard some previous possibilities. This is however the case for update where we apply the update process to each world independently (see [7] for an in depth). So $(R8)$ is more suitable for an update operation.

In fact $(R8)$ can be seen as the multi-agent counterpart of the propositional update postulate $(U8)$: consider $\psi := B_i p \vee B_j p$ and $\varphi := \neg B_i p$. Then the revised formula is $\psi \circ \varphi = B_j p \wedge \neg B_i p$ according to postulate $(R2)$. But according to postulate $(R8)$, after the revision $\neg B_j p$ should be satisfiable because $\psi \wedge \neg B_j p$ was satisfiable.

Postulates $(R7)$ and $(R8)$ together are equivalent to: for all $\varphi, \varphi' \in \mathcal{L}_{j \neq Y}$ such that $C(\varphi) \cap C(\varphi') = \emptyset$, $\psi \rightarrow \varphi'$ iff $\psi \circ \varphi \rightarrow \varphi'$. Then

Proposition 6. *Revision operator $\circ$ satisfies $(R7)$ and $(R8)$ iff for all $\varphi \in \mathcal{L}_{j \neq Y}$, $Mod(\psi) \leftrightarrows_{C'} Mod(\psi \circ \varphi)$, with $C' := C - C(\varphi)$.*

5 An Example of Revision Operator

In this section we propose a revision operator based on a degree of similarity between multi-agent possible worlds defined very much in the same way as it is done in [9].

5.1 Mathematical Preliminaries

The structure $\mathcal{F}$ used to measure the degree of similarity between multi-agent possible worlds We are going to define a degree of similarity between multi-agent possible worlds. This degree of similarity will be measured by an integer or an infinite number. Ordinals are not commutative for addition so we resort to a fragment of hyperreal numbers. Hyperreal numbers are an extension of real numbers with infinite numbers and infinitely small numbers. We consider first the set $S_0 := \mathbb{N} \cup \{\infty\}$ of all natural numbers together with an arbitrary infinite number ∞. The fragment $\mathcal{F}$ we consider is simply the closure of S_0 under addition: $\mathcal{F} := \{x + y; x \in S_0 \text{ and } y \in S_0\}$. The order $<$ on $\mathcal{F}$ is just the restriction of the order for hyperreals to $\mathcal{F}$. So for instance we have $2 < \infty$, $\infty < \infty + \infty$,...

n-bisimulation Our definition of n-bisimulation is a slight modification of the definition of n-bisimulation in [3].

Definition 9. *Let Z be a relation between worlds of two finite epistemic models M and M'. We recursively define the property of Z being a n-bisimulation in w and w', noted $Z : M, w \leftrightarrows_n M', w'$:*

1. *$Z : M, w \leftrightarrows_0 M', w'$ iff wZw' and $val(w) \neq val(w')$;*
2. *$Z : M, w \leftrightarrows_1 M', w'$ iff wZw' and $val(w) = val(w')$;*
3. *For all $n \geq 1$ $Z : M, w \leftrightarrows_{n+1} M', w'$ iff wZw' and $val(w) = val(w')$ and for all $j \in G$,*
   - *for all $v \in R_j(w)$ there is $v' \in R_j(w')$ such that $Z : M, v \leftrightarrows_n M', v'$.*
   - *for all $v' \in R_j(w')$ there is $v \in R_j(w)$ such that $Z : M, v \leftrightarrows_n M', v'$.*

The usual definition of Z being a bisimulation corresponds to $Z : M, w \leftrightarrows_n M', w'$ for all $n \in \mathbb{N}^*$. It will be noted here $Z : M, w \leftrightarrows_\infty M', w'$. Now we can define n-bisimilarity between w and w', noted $M, w \leftrightarrows_n M', w'$ by $M, w \leftrightarrows_n M', w'$ iff there exists a relation Z such that $Z : M, w \leftrightarrows_n M', w'$.

Two worlds being n-bisimilar (with $n \geq 1$) intuitively means that they have the same modal structure up to modal depth $n - 1$, and thus they satisfy the same formulas of degree at most $n - 1$.

5.2 Definition of the Revision Operator

First we are going to define a degree of similarity d between two multi-agent possible worlds that will allow for a lexicographic order.

Definition 10. *Let (M, w) and (M', w') be two multi-agent possible worlds and $\mathcal{M}$ and $\mathcal{M}'$ be two sets of multi-agent possible worlds.*
Let $v \in M$, $v' \in M'$ and $n := max\{|M|; |M'|\}$,

- *$\delta(v, v') := max\{k \in \mathcal{F}; M, v \leftrightarrows_k M', v'\}$;*
- *$\delta(S, S') := max\{\delta(v, v'); v \in S \text{ and } v' \in S'\}$;*

– $d((M,w),(M',w')) :=$
$(\delta(w,w'), \sum_{j\in G} \delta(R_j(w),R_j(w')),.., \sum_{\substack{j_1,..,j_n\in G \\ j_i\neq j_{i+1}}} \delta(R_{j_1}\circ..\circ R_{j_n}(w), R_{j_1}\circ..\circ R_{j_n}(w')))$;
– $d(\mathcal{M},\mathcal{M}') := max\{d((M,w),(M',w')); (M,w)\in\mathcal{M}, (M',w')\in\mathcal{M}'\}$.

$\delta(v,v')$ measures a degree of similarity between the worlds v and v'. Note that $0 \leq \delta(v,v') \leq \infty$ for all v and v'. If $\delta(v,v') = \infty$ then the worlds v and v' are bisimilar by definition. So their degree of similarity is the highest possible. If $\delta(v,v') = 0$, that is $M, v \leftrightarrows_0 M', v'$ then their degree of similarity is the lowest possible because they differ even on propositional facts. $d((M,w),(M',w'))$ is a tuple which represents by how much two multi-agent possible worlds are similar relatively to their respective modal depth. Note that for a given modal depth we only compare the degree of similarity of worlds which have the same history (i.e. they are all accessed from w and w' by the same sequence of accessibility relations $R_{j_1},\ldots,R_{j_k}$). Doing so, in our comparison we stick very much to the modal structure of both multi-agent possible worlds. Besides we take the sum of their degree of similarity for every possible history in order to give the same importance to these different possible histories. Finally, the tuple is of size $n = max\{|M|,|M'|\}$ in order to reach all the worlds of both models.

When comparing multi-agent possible worlds, we would like to give priority to the similarity of worlds of low modal depth rather than to the similarity of worlds of high modal depth. This can be achieved by defining a lexicographic ordering between tuples.

Definition 11. *Let $(l_1,\ldots,l_n)\in\mathcal{F}^n$ and $(l'_1,\ldots,l'_m)\in\mathcal{F}^m$. We set $(l_1,\ldots,l_n) <_l (l'_1,\ldots,l'_m)$ iff*

– *$l_1 < l'_1$ or*
– *for all $i<j\leq min\{m,n\}$ $l_i = l'_i$ and $l_j < l'_j$ or*
– *for all $i\leq min\{m,n\}$ $l_i = l'_i$ and $n<m$.*

Then we set $\leq_l := <_l \cup\{(L,L),((\infty,L),(\infty,L'));$ L and L' are arbitrary tuples $\}$.

Now we can define the revision operator.

Definition 12. *Let $\psi\in\mathcal{L}_{j\neq Y}$. We assign to ψ a total pre-order $\leq_\psi$ on multi-agent possible worlds defined as follows:*

$(M,w)\leq_\psi (M',w')$ iff $d(Mod(\psi),(M,w)) \geq_l d(Mod(\psi),(M',w'))$.

The revision operator $\circ$ associated to this pre-order $\leq_\psi$ is defined semantically by $Mod(\psi\circ\mu) := Min(Mod(\mu),\leq_\psi)$.

We finally have the following nice property.

Proposition 7. *The assignment defined in Definition 12 is a faithful assignment. Therefore the operator $\circ$ defined in Definition 12 satisfies the postulates $(R1)-(R6)$. Besides $\circ$ satisfies also $(R7)$.*

5.3 Concrete Example

Consider the subjective model defined in Figure 1. Proposition 8 below tells us that $\{(M_1', w'); (M_2', v')\}$ displayed in Figure 3 is the result of the revision of the subjective model $\{(M_1, w); (M_2, v)\}$ by the formula $B_A p$.

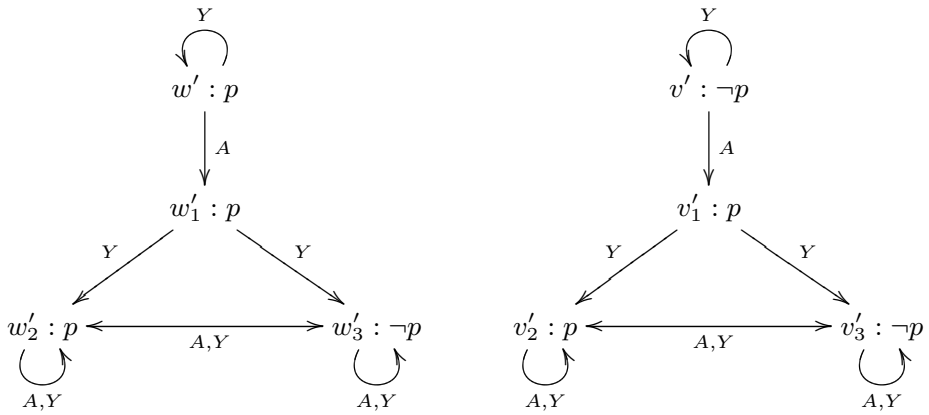


Fig. 3. Multi-agent possible worlds (M_1', w') and (M_2', v')

Proposition 8. $Mod(form(\{(M_1, w); (M_2, v)\}) * B_A p) = \{(M_1', w'); (M_2', v')\}$

The epistemic model associated to $\{(M_1', w'); (M_2', v')\}$ is bisimilar to the epistemic model depicted in Figure 4. If we compare this model with the original model of Figure 2 we observe that agent Y's beliefs about A's beliefs have indeed changed; but agent A's beliefs about Y's beliefs have not changed. This is what we should expect after a private announcement of $B_A p$ to Y because agent A is not aware of this announcement (see the introduction).

6 Conclusion

We have proposed a semantics to adequately represent the agent Y's perception of the surrounding world in a multi-agent setting and have connected this semantics with (standard) epistemic models. This semantics generalizes the single agent one of AGM belief revision theory. Then Proposition 1 has enabled us to also generalize easily the results of AGM belief revision theory to the multi-agent case. Finally, we have studied two additional multi-agent postulates and we have given an example of revision operator that satisfies one of these multi-agent postulates.

The power of our approach is that it generalizes all the results of the AGM belief revision theory to the multi-agent case. In fact, if we consider in particular that there are no other agents than Y then our approach boils down to classical AGM belief revision.

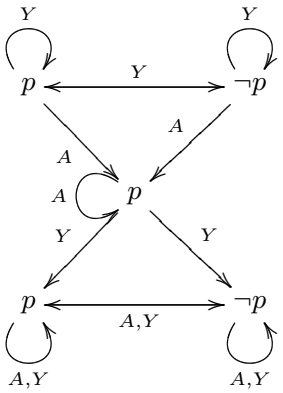


Fig. 4. Epistemic model bisimilar to the epistemic model associated to $\{(M'_1, w'); (M'_2, v')\}$

It would be interesting to investigate other multi-agent postulates and other distances over multi-agent possible worlds.[3] Another line of research would be to study multi-agent update as we have started in Section 4.2. Indeed, the results of [8] about propositional update transfer to the multi-agent case as well.

References

1. Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *J. Symb. Log.*, 50(2):510–530, 1985.
2. Guillaume Aucher. A combined system for update logic and belief revision. In *PRIMA 2004*, volume 3371 of *LNCS*, pages 1–17. Springer, 2004.
3. Philippe Balbiani and Andreas Herzig. Talkin'bout Kripke models. Accepted to Hylo'07, 2007.
4. Alexandru Baltag and Mehrnoosh Sadrzadeh. The algebra of multi-agent dynamic belief revision. *Electr. Notes Theor. Comput. Sci.*, 157(4):37–56, 2006.
5. Alexandru Baltag and Sonja Smets. Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electr. Notes Theor. Comput. Sci.*, 165:5–21, 2006.
6. Samir Chopra and Rohit Parikh. An inconsistency tolerant model for belief representation and belief revision. In *IJCAI*, pages 192–199, 1999.
7. Hirofumi Katsuno and Alberto O. Mendelzon. On the difference between updating a knowledge base and revising it. In *KR*, pages 387–394, 1991.
8. Hirofumi Katsuno and Alberto O. Mendelzon. Propositional knowledge base revision and minimal change. *Artif. Intell.*, 52(3):263–294, 1992.
9. Daniel J. Lehmann, Menachem Magidor, and Karl Schlechta. Distance semantics for belief revision. *J. Symb. Log.*, 66(1):295–317, 2001.

[3] In that respect our definition of $\delta(S, S')$ in Definition 10 could be criticized because a unique couple, namely the maximal one, is taken to represent a whole set of tuples. Likewise, our lexicographic ordering $\leq_l$ could be defined differently.

10. J. van Benthem. "One is a Lonely Number": logic and communication. In *Logic Colloquium'02*. ASL & A.K. Peters, 2006.
11. Hans van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147:229–275, 2005.

What You Should Believe: Obligations and Beliefs

Guido Boella[1], Célia da Costa Pereira[2], Gabriella Pigozzi[3], Andrea Tettamanzi[2], and Leendert van der Torre[3]

[1] Università degli Studi di Torino, Dipartimento di Informatica
10149, Torino, C.so Svizzera 185, Italy
guido@di.unito.it
[2] Università degli Studi di Milano, Dipartimento di Tecnologie dell'Informazione
26013, Crema, via Bramante 65, Italy
{pereira,andrea.tettamanzi}@dti.unimi.it
[3] Université du Luxembourg, Computer Science and Communication
L-1359, Luxembourg, rue Richard Coudenhove - Kalergi 6, Luxembourg
{gabriella.pigozzi,leon.vandertorre}@uni.lu

Abstract. This paper presents and discusses a novel approach to indeterministic belief revision. An indeterministic belief revision operator assumes that, when an agent is confronted with a new piece of information, it can revise its belief base in more than one way. We define a rational agent not only in terms of what it believes, as often assumed in belief revision, but also of what it ought or is obliged to do. Hence, we propose that the agent's goals play a role in the choice of (possibly) one of the several available revision options. Properties of the new belief revision mechanism are also investigated.

Keywords. Rational agents, indeterministic belief revision, qualitative decision theory.

1 Introduction

Norms and obligations are increasingly being introduced in Multiagent Systems, in particular to meet the coordination needs of open systems where heterogeneous agents interact with each other. Witness the numerous papers presented at conferences and the organization of workshops like NorMas and COIN in the last years. Introducing norms raise the issue, however, of the interaction between obligations and other mental attitudes like beliefs, goals, and intentions. While the relation between obligations and motivational attitudes is being studied [4,6,5,12,11,19,20,3,10,16], the relation between beliefs and obligations is still unclear. In this paper we study the role of obligations in the task of revising the agent's beliefs under the light of new information. Revising the beliefs can lead to a situation where a choice among different alternatives cannot be made on the basis of the available information. However, obligations and other motivational attitudes can lead a rational agent to choose among the equally likely alternatives, in order not to lose precious opportunities.

For example, suppose that you are a politician who is subject to the obligation to reduce deficit, for example due to a decision of the EU or the IMF, and you believe that

A) Blocking enrollment leads to a decrease in spending
B) A decrease of investment in infrastructures leads to a decrease in spending
C) A decrease in spending leads to a reduction of deficit

Therefore, your plan to meet your obligation is either to block the enrollment, or to decrease investment in infrastructures.

Now, suppose that someone very trustworthy and well-reputed convinces you that blocking enrollment does not lead to a reduction of deficit. Beliefs A and C cannot hold together anymore, and you have to give up one of them.

If you give up A, you still have another possibility to reduce the deficit because you can decrease spending by decreasing investment in infrastructures. However, if you give up C, you do not have any possibility to achieve the reduction of deficit. Indeed,

1. Let us first assume that A is factually wrong, whereas C is true. If you choose to retain (wrong) belief A and to reject C, you will do nothing and you will not succeed in reducing deficit. But, had you kept your belief in C and rejected A, you could have decreased investment in infrastructures in order to decrease spending, and therefore you could have met your obligation to reduce deficit. To conclude, by choosing to maintain A, you risk to miss an opportunity to meet your obligation.
2. Let us now assume that A is actually true and C is wrong. If you choose to keep (wrong) belief C, you will decrease spending, but you will not achieve the goal of reducing deficit. However, even if you had chosen the right revision, i.e., to retain A and reject C, there was no way for you to achieve your goal of reducing deficit. To conclude, by choosing C (wrong), you believed you could achieve a goal when you could not, so you will be disappointed for trying in vain, but at least you tried.

The moral of the story is that, if you are interested only in meeting your obligation (and there are no other goals relevant for you), choosing to maintain C — *even when it is factually wrong*, but you do not know whether it is false or not — is the only rational choice. This is because, independently of C being right or wrong, by choosing that belief you will be better off. Moreover, in one situation — the former — you will be better off if you choose C than if you choose A. Summarizing, you should drop A, because that way, you keep all possibilities to achieve your goal open.

We can formalize the above example, by defining the following atomic propositions:

b blocking enrollment;
s decrease spending;
d reduce deficit;
i decrease investment in infrastructures.

The belief base before being convinced that blocking enrollment does not lead to a reduction of deficit ($\neg(b \supset d)$) would contain the three formulas $b \supset s$, $i \supset s$, and $s \supset d$. You have to, first of all, reduce deficit, d, and, if possible, not decrease investment in infrastructures, $\neg i$. Adding $\neg(b \supset d)$ to your beliefs would make them inconsistent. Therefore, you have to revise your beliefs by giving up either $b \supset s$ or $s \supset d$. The choice you make may depend on the obligations you can meet in the alternatives: if you give up $b \supset s$, your plan will be to decrease investment in infrastructures, so you will not achieve $\neg i$, but might succeed in achieving d; if you give up $s \supset d$, your plan will

be to do nothing, so you will certainly not achieve d, but you will fulfill $\neg i$. Depending on the punishment you have for violating your obligation to achieve d or $\neg i$, you could prefer one or the other alternative.

We use the deficit reduction example as a running example throughout the paper.

The choice among belief bases is distinct from other decision problems, due to the possibility of wishful thinking. Consider for example that you have to block the enrollment (b) to decrease spending, and that this obligation is more important than the obligation of reducing deficit (d). What will you do? At least in a naive approach, you could reason by cases as follows. Assume you choose $b \supset s$: in that case you believe that accomplishing the obligation of blocking enrollment leads to a decrease in spending. Assume you choose $s \supset d$: in that case you believe you will achieve the goal of reducing deficit. Since b is more important than d, you choose $b \supset s$.

Instead, the idea of this paper is inspired by the notion of *conventional wisdom* (CW) as introduced by economist John Kenneth Galbraith:

> We associate truth with convenience, with what most closely accords with self-interest and personal well-being. ([14])

That is, CW consists of "ideas that are convenient, appealing". This is the rationale for keeping them. One basic brick of CW could then be the fact that some ideas are maintained because they maximize the goals that the agents (believe) they can achieve. This work may be seen as an initial attempt to formally capture the concept of a CW agent. In the following we provide a logical framework that models how a CW agent revises its beliefs under its obligations.

The paper is structured as follows. In Section 2 we introduce the aim of this paper, the used methodology and particular challenges encountered. In Section 3 we introduce the agent theory we use in our approach, and in Section 4 we introduce an indeterministic belief change operator in this agent theory. In Section 5 we define the choice among beliefs as a decision problem in the agent theory. Section 6 concludes.

2 Aim, Methodology and Challenges

The research problem of this paper is to develop a formal model to reason about the kind of choices among belief bases discussed in the previous section, and to generalize the example above in case of additional beliefs, multiple goals with different importance, conditional obligations, a way to take violated goals into account, and so on.

We use a combination of the framework of belief revision together with a qualitative decision theory. Classical approaches to belief revision assume that, when an agent revises its belief base in view of new input, the outcome is well-determined. This picture, however, is not realistic. When an agent revises its beliefs in the light of some new fact, it often has more than one available alternative. Approaches to belief revision that do not stipulate the existence of a single revision option are called *indeterministic* [18,24]. In this paper we suggest that one possible policy an agent can use in order to choose among available alternatives is to check the effect of the different revisions on the agent's set of goals.

Moreover, for the qualitative decision theory we are inspired by agent theories such as the BOID architecture [4,6,5,12,11,19,20,3,10], the framework of goal generation in 3APL as developed by van Riemsdijk and colleagues [26], and [8]. In particular, our agent model is based on one of the versions of 3APL, because the belief base in the mental state of a 3APL agent is a consistent set of propositional sentences, just like in the framework of belief revision. However, we do not care about how goals are generated and how their achievability (plan existence) is established. That is because we do not include "goal-adoption rules" or "practical reasoning rules" representing which action to choose in a particular state. We assume that there is a planning module, which would take a set of goals, actions, and an initial world state representation in input and produce a solution plan in output. This planning module might rely on the well-known graphplan algorithm [2], or any other propositional AI planner: as in object-oriented programming, we *encapsulate* the planner within a well-defined interface and overlook the implementation details of how a solution plan is found. This is in line with, on one hand, the BOID architecture [4], where the planning component is kept separate from the remainder of agent deliberation, and, on the other hand with the works of Móra and colleagues describing the relationship between propositional planning algorithms and the process of means-end reasoning in BDI agents. In these works, [21,22], it is shown how the mental state of an agent can be mapped to the STRIPS [13] notation forth and back. This relation has been done on an abstract BDI interpreter named X-BDI [27,7] and augmented with graphplan.

In other words, we model the choice among belief bases essentially as a decision problem, that is, as a choice among a set of alternatives. We do not use classical decision theory (utility function, probability distribution, and the decision rule to maximize expected utility), but a qualitative version based on maximizing achieved goals and minimizing violated goals in an abstract agent theory (see e.g. [9] for various approaches to formalize the decision process of what an agent should do), because such qualitative decision theories include beliefs and therefore are easier to combine with the theory of belief revision. However, what precisely are the alternatives?

An indeterministic belief revision operator associates multiple revision options to a belief base that turns out to be inconsistent as a consequence of a new piece of information. Our revision mechanism selects the revision alternative that allows the agent to maximize its achievable goals. However, it will not always be possible to select exactly one revision alternative. For example, there may be one most important goal set but two revision alternatives that lead the agent to achieve it. In this case, the two belief revision candidates are said to be equivalent. In Section 5.3 we will provide conditions under which a revision for a CW agent is deterministic, that is, when our revision operator can select exactly one revision alternative.

Besides the issue of wishful thinking, another complicating factor when choosing among belief bases in the context of conditional obligation rules, is that a maximization of goals may lead to a meta-goal to derive obligations by choosing revisions where you believe that the condition is true and the obligation applies. However, deriving goals by itself does not have to be desirable. In contrast, it may even be argued that fewer goals are better than more goals, as you risk to violate goals and become unhappy (as in Buddhism). We therefore also take goal violations into account.

3 An Abstract Agent Theory

In this section, we represent the formalism which is used throughout the paper.

3.1 A Brief Introduction to AI Planning and Agent Theory

Any agent, be it biological or artificial, must possess knowledge of the environment it operates in, in the form of, e.g., *beliefs*. Furthermore, a necessary condition for an entity to be an *agent* is that it acts. We shall call the factors that motivate an agent to act *obligations*. For artificial agents, obligations may be the purposes an agent was created for.

Obligations are necessary, not sufficient, conditions for action. When an obligation is met by other conditions that make it possible for an agent to act, that obligation becomes a *goal*.

The reasoning side of acting is known as practical reasoning or deliberation, which may include *planning*. Planning is a process that chooses and organizes actions by anticipating their expected effects with the purpose of achieving as good as possible some pre-stated objectives or goals.

The objective of our formalism is to analyze, not to develop, agent systems. More precisely, our agent must single out the set of goals to be given as input to a traditional planner. That is because the intentions of the agent are not considered. We merely consider beliefs (knowledge the agent has about the world states), obligations (or motivations) and relations (obligation-adopting rules) defining how the obligation base will change with the acquisition of new beliefs and/or new obligations. The goal generation process that underlies this work is very much in line with the work carried out in [25] on *oversubscription planning problems*, in which the main objective is to find the maximal set of desires to be reached in a given period and with a limited quantity of resources, and with goal generation in the BOID architecture [4].

3.2 Beliefs, Obligations, and Goals

The basic components of our language are *beliefs* and *obligations*. Beliefs are represented by means of a *belief base*. A belief base is a finite and consistent set of propositional formulas describing the information the agent has about the world and internal information. Obligations are represented by means of an *obligation base*. An obligation base consists of a set of propositional formulas which represent the situations the agent has to achieve. However, unlike the belief base, an obligation base may be inconsistent, e.g., $\{p, \neg p\}$.

Definition 1 (Belief Base B and Obligation Base O) *Let $\mathcal{L}$ be a propositional language with $\top$ a tautology, and the logical connectives $\wedge$ and $\neg$ with the usual meaning. The agent's belief base B is a consistent finite set such that $B \subseteq \mathcal{L}$. B can also be represented as the conjunction of its propositional formulas. The agent's obligation base is a possibly inconsistent finite set of sentences denoted by O, with $O \subseteq \mathcal{L}$.*

We define two modal operators Bel and Obl such that, for any formula ϕ of $\mathcal{L}$, Belϕ means that ϕ is believed whereas Oblϕ means that the agent has obligation ϕ. Since the belief and obligation bases of an agent are completely separated, there is no need to nest the operators Bel and Obl.

Definition 2 (Obligation-Adoption Rule) *An obligation-adoption rule is a triple* $\langle \phi, \psi, \tau \rangle \in \mathcal{L} \times \mathcal{L} \times \mathcal{L}$ *whose meaning is: if* Belϕ *and* Oblψ*, then* τ *will be adopted as an obligation as well.*

The set of obligation-adoption rules is denoted by R. If $\exists \phi', \psi', \tau'$ such that $\phi \leftrightarrow \phi', \psi \leftrightarrow \psi', \tau \leftrightarrow \tau'$, then $\langle \phi', \psi', \tau' \rangle \in R$.

Goals, in contrast to obligations, are represented by consistent obligation sets. There are various ways to generate candidate goal sets from the obligation adoption rules, as discussed in the remainder of this section.

Definition 3 (Candidate Goal Set) *A candidate goal set is a consistent subset of* O.

3.3 Mental State Representation

We assume that an agent is equipped with three components:

- belief base $B \subseteq \mathcal{L}$;
- obligation base: $O \subseteq \mathcal{L}$;
- obligation-adoption rule set R.

The mental state $\mathcal{S}$ of an agent is completely described by a triple $\mathcal{S} = \langle B, O, R \rangle$. In addition, we assume that each agent can be described using a problem-dependent function $\mathcal{V}$, a goal selection function G, and a belief revision operator $*$, as discussed below.

In our deficit reduction example, we have:

$$B = \{\neg(b \wedge \neg s), \neg(i \wedge \neg s), \neg(s \wedge \neg d)\},$$
$$O = \{d, \neg i\},$$
$$R = \{\langle \top, \top, d \rangle; \langle \top, \top, \neg i \rangle\}.$$

The semantics we adopt for the belief and obligation operators are standard.

Definition 4 (Semantics of Bel **operator)** *Let* $\phi \in \mathcal{L}$*,* Bel$\phi \Leftrightarrow B \models \phi$.

Definition 5 (Semantics of Obl **operator)** *Let* $\phi \in \mathcal{L}$*,* Obl$\phi \Leftrightarrow \exists$ *a maximal consistent subset* $O' \subseteq O$ *such that* $O' \models \phi$.

We expect a rational agent to try and manipulate its surrounding environment to fulfill its goals. In general, given a problem, not all goals are *achievable*, i.e. it is not always possible to construct a plan for each goal. The goals which are not achievable or those which are not chosen to be achieved are called *violated goals*. Hence, we assume a problem-dependent function $\mathcal{V}$ that, given a belief base B and a goal set $O' \subseteq O$, returns a set of couples $\langle O^a, O^v \rangle$, where O^a is a maximal subset of achievable goals and O^v is the subset of violated goals and is such that $O^v = O' \setminus O^a$. Intuitively, by considering violated goals we can take into account, when comparing candidate goal sets, what we lose from not achieving certain goals.

3.4 Comparing Goals and Sets of Goals

The aim of this section is to illustrate a qualitative method for goal comparison in the agent theory. More precisely, we define a qualitative way in which an agent can choose among different sets of candidate goals. Indeed, from an obligation base O, several candidate goal sets O_i, $1 \leq i \leq n$, may be derived. How can an agent choose among all the possible O_i? It is unrealistic to assume that for a rational agent all goals have the same priority. We use the notion of importance of obligations to represent how relevant each goal is for the agent depending, for instance, on the punishment for violating the obligations. The idea is that a rational agent tries to choose a set of candidate goals which contains the greatest number of achievable goals (or the least number of violated goals).

We assume we dispose of a total order $\succeq$ over an agent's obligations. In the example, you have to reduce, in the first place, deficit and, if possible, you should not decrease investments in infrastructures. Therefore, d is more important than $\neg i$, in symbols $d \succeq \neg i$.

The $\succeq$ relation can be extended from goals to sets of goals. We have that a goal set O_1 is more important than another one O_2 if, considering only the goals occurring in either set, the most important goals are in O_1 or the least important goals are in O_2. Note that $\succeq$ is connected and therefore a total pre-order, i.e., we always have $O_1 \succeq O_2$ or $O_2 \succeq O_1$.

Definition 6 (Equivalent Goals)
A goal ϕ_1 is said equivalent to a goal ϕ_2, noted $\phi_1 \approx \phi_2$, if and only if ϕ_1 and ϕ_2 are equally important, i.e. $\phi_1 \succeq \phi_2$ and $\phi_2 \succeq \phi_1$.

Definition 7 (Difference Goal set Operator)
Let O_1 and O_2 be two sets of goals. The difference based on the equivalence between goals in O_1 and in O_2 noted $O_1 \setminus_{\approx} O_2$, is defined as follow:

$$O_1 \setminus_{\approx} O_2 = \{\phi_1 \in O_1 | \neg \exists \phi_2 \in O_2 \text{ such that } \phi_1 \approx \phi_2\}$$

Definition 8 (Relative Importance of Sets of Goals)
Let $O_1' = O_1 \setminus_{\approx} O_2$ and $O_2' = O_2 \setminus_{\approx} O_1$. The goal set O_1 is at least as important as O_2, denoted $O_1 \succeq O_2$ iff

$$O_2' = \emptyset \text{ or } \exists \phi_1 \in O_1', \forall \phi_2 \in O_2' \, \phi_1 \succeq \phi_2.$$

In our example, it is easy to verify that $\{d, \neg i\} \succ \{d\} \succ \{\neg i\} \succ \emptyset$. However, we also need to be able to compare the mutual exclusive subsets (achievable and violated goals) of the considered candidate goal, as defined below.

3.5 Comparing Couples of Goal Sets

We propose two methods to compare couples of goal sets.

3.5.1 The *Direct* Comparison $\succeq_D$

Given the $\succeq_D$ criterion, a couple of goal sets $\langle O_1^a, O_1^v\rangle$ is at least as important as the couple $\langle O_2^a, O_2^v\rangle$, noted $\langle O_1^a, O_1^v\rangle \succeq_D \langle O_2^a, O_2^v\rangle$ iff $O_1^a \succeq O_2^a$ and $O_1^v \preceq O_2^v$.
$\succeq_D$ is reflexive and transitive but partial. $\langle O_1^a, O_1^v\rangle$ is strictly more important than $\langle O_2^a, O_2^v\rangle$ in two cases:

1. $O_1^a \succeq O_2^a$ and $O_1^v \prec O_2^v$, or
2. $O_1^a \succ O_2^a$ and $O_1^v \preceq O_2^v$.

They are indifferent when $O_1^a = O_2^a$ and $O_1^v = O_2^v$. In all the other cases, they are not comparable.

3.5.2 The *Lexical* Comparison $\succeq_{Lex}$

Given the $\succeq_{Lex}$ criterion, a couple of goal sets $\langle O_1^a, O_1^v\rangle$ is at least as important as the couple $\langle O_2^a, O_2^v\rangle$ (noted $\langle O_1^a, O_1^v\rangle \succeq_{Lex} \langle O_2^a, O_2^v\rangle$) iff $O_1^a = O_2^a$ and $O_1^v = O_2^v$; or there exists a $\phi \in \mathcal{L}$ such that:

1. $\forall\, \phi' \succeq \phi$, the two couples are indifferent, i.e., one of the following possibilities holds:
   a) $\phi' \in O_1^a \cap O_2^a$;
   b) $\phi' \notin O_1^a \cup O_1^v$ and $\phi' \notin O_2^a \cup O_2^v$;
   c) $\phi' \in O_1^v \cap O_2^v$.
2. Either of the following holds:
   a) $\phi \in O_1^a \setminus O_2^a$;
   b) $\phi \in O_2^v \setminus O_1^v$.

$\succeq_{Lex}$ is reflexive, transitive, but partial.

3.6 Defining the Goal Set Selection Function

In general, given a set of obligations O, there may be many possible candidate goal sets. A rational agent in state $\mathcal{S} = \langle B, O, R\rangle$ will select one precise candidate goal set O' which consists of the most important couple of achievable and violated goals.

Let us call G the function which maps a state $\mathcal{S}$ into the goal set selected by a rational agent in state $\mathcal{S}$. G is such that $G(\mathcal{S}) = O'$.

4 Situating the Problem: Indeterministic Belief Change

> "Most models of belief change are deterministic. Clearly, this is not a realistic feature, but it makes the models much simpler and easier to handle, not least from a computational point of view. In indeterministic belief change, the subjection of a specified belief base to a specified input has more than one admissible outcome.
> Indeterministic operators can be constructed as sets of deterministic operations. Hence, given n deterministic revision operators $*_1, *_2, \ldots, *_n$, $* = \{*_1, *_2, \ldots, *_n\}$ can be used as an indeterministic operator." [17]

Let us consider a belief base B and a new belief β. The revision of B in light of β is simply:

$$B * \beta \in \{B *_1 \beta, B *_2 \beta, \dots B *_n \beta\}. \tag{1}$$

More precisely, revising the belief base B with the indeterministic operator $*$ in light of the new belief β leads to one of the n belief revision results:

$$B * \beta \in \{B^1_\beta, B^2_\beta, \dots B^n_\beta\}, \tag{2}$$

where B^i_β is the i-th possible belief revision result.

Applying the operator $*$ is then equivalent to applying one of the virtual operators $*_i$ contained in its definition. While the rationality of an agent does not suggest any criterion to prefer one revision over the others, a defining feature of a CW agent is that it will choose which revision to adopt based on the consequence of that choice. One important consequence is the set of goals the agent will decide to pursue.

In our deficit reduction example, $\beta = \mathrm{Bel}(b \wedge \neg d)$, and

$$B * \beta \in \left\{ \begin{array}{l} B^1_\beta = \{b \wedge \neg d, \neg(s \wedge \neg d), \neg(i \wedge \neg s)\}, \\ B^2_\beta = \{b \wedge \neg d, \neg(b \wedge \neg s), \neg(i \wedge \neg s)\} \end{array} \right\}. \tag{3}$$

In the next section we propose some possible ways to tackle the problem of choosing one of the revision options.

5 Belief Revision as a Decision Problem

By considering an indeterministic belief revision, we admit $B * \beta$ to have more than one possible result. In this case, the agent must select (possibly) one among all possible revisions. Among the possible criteria for selection, one is to choose the belief revision operator for which the goal set selection function returns the most important goal set. In other words, selecting the revision amounts to solve an optimization problem.

5.1 Indeterministic State Change

The indeterminism of belief revision influences the obligation-updating process. In fact, the belief revision operator is just a part of the state-change operator, which is indeterministic as well, as a consequence of the indeterminism of belief revision. Therefore, $\mathcal{S}_\beta \in \{\mathcal{S}^1_\beta, \mathcal{S}^2_\beta, \dots, \mathcal{S}^n_\beta\}$, where $\mathcal{S}^i_\beta = \langle B^i_\beta, O^i_\beta, R\rangle$.

Which goal set is selected by an agent depends on G:

$$G(\mathcal{S}_\beta) \in \{G(\mathcal{S}^1_\beta), G(\mathcal{S}^2_\beta), \dots, G(\mathcal{S}^n_\beta)\}. \tag{4}$$

In the example, $G(\mathcal{S}_\beta) \in \{G(\mathcal{S}^1_\beta), G(\mathcal{S}^2_\beta)\}$, where $G(\mathcal{S}^1_\beta) = \{d\}$ and $G(\mathcal{S}^2_\beta) = \{\neg i\}$. The following table summarizes the possibilities the agent may face when choosing between the two alternative revisions.

reality $\rightarrow$ $\downarrow$ beliefs	$\not\models b \supset s$ $\models s \supset d$	$\models b \supset s$ $\not\models s \supset d$
B^1_β decrease investment in infrastructures	d is achieved $\neg i$ is not achieved	no obligation is met
B^2_β do nothing	d is not achieved $\neg i$ is achieved	

A traditional rational agent could not choose one of the $G(\mathcal{S}^i_\beta)$ because they are incomparable. Now, for a CW agent,

$$G(\mathcal{S}_\beta) \in \mathrm{I}\{G(\mathcal{S}^1_\beta), G(\mathcal{S}^2_\beta), \ldots, G(\mathcal{S}^n_\beta)\}, \tag{5}$$

where $\mathrm{I}(S)$ denotes the most important set of S defined as follows:

Definition 9 (Important Set I**)** *Given two sets S and X such that $S \subseteq X$, and given an importance relation $\succeq$ over X, the most important set of S is*

$$\mathrm{I}(S) = \{x \in S : \neg\exists x' \in S, x' \succ x\}. \tag{6}$$

5.2 Choosing a Revision

Choosing the most important revision option is not a trivial operation. We can distinguish two situations:

- there is just one most important goal set O', but more than one alternative options leads to O';
- there is no unique most important goal set; that is, there are different goal sets $O_1, \ldots, O_m$, none of which is strictly more important than the others, i.e., for all $i, j \in \{1, \ldots, m\}$, $O_i \succeq O_j$.

Definition 10 (Equivalent Belief Revision Candidates) *A belief revision candidate B^1_β is equivalent to another belief revision candidate B^2_β (denoted by $B^1_\beta \approx B^2_\beta$), if and only if $G(\mathcal{S}^1_\beta) \succeq G(\mathcal{S}^2_\beta)$ and $G(\mathcal{S}^2_\beta) \succeq G(\mathcal{S}^1_\beta)$.*

It is easy to verify that $\approx$ is a standard equivalence relation, i.e., reflexive, symmetric, and transitive.

The choice of which revision outcome to adopt may thus be deterministic or indeterministic. It is indeterministic in the two cases presented above. More precisely, the choice depends on the importance relations over the goal sets, which determine the equivalence between revision candidates:

- if $\|\mathrm{I}\{G(\mathcal{S}^1_\beta), G(\mathcal{S}^2_\beta), \ldots, G(\mathcal{S}^n_\beta)\}\| = 1$, i.e., the equivalent class of an important belief revision is a singleton and, if there is no i, j such that $G(\mathcal{S}^i_\beta) = G(\mathcal{S}^j_\beta)$, the choice of the belief operator is obviously deterministic;
- if $\|\mathrm{I}\{G(\mathcal{S}^1_\beta), G(\mathcal{S}^2_\beta), \ldots, G(\mathcal{S}^n_\beta)\}\| = 1$, and there is at least a couple i, j such that $G(\mathcal{S}^i_\beta) = G(\mathcal{S}^j_\beta)$, the choice is indeterministic, but also indifferent;
- if $\|\mathrm{I}\{G(\mathcal{S}^1_\beta), G(\mathcal{S}^2_\beta), \ldots, G(\mathcal{S}^n_\beta)\}\| > 1$, the choice is indeterministic.

It is important to notice that an agent that has to choose between $G(\mathcal{S}^i_\beta)$ and $G(\mathcal{S}^j_\beta)$ is in a different situation than an agent that has to randomly choose among a number of competing revisions. The reason is that a random choice is hardly a rational option. But, when an agent must choose between two revision options, it knows that, no matter which revision it chooses, the outcome does not change. In such a context, a random choice becomes a rational option.

Proposition 1 *Let $*$ be an indeterministic belief operator, and n be the number of possible belief revisions candidate. We have:*

$$1 \leq \|\mathrm{I}\{G(\mathcal{S}^1_\beta), G(\mathcal{S}^2_\beta), \ldots, G(\mathcal{S}^n_\beta)\}\| \leq n.$$

5.3 Conditions for Determinism of a CW Agent

Traditional indeterministic belief revision approaches allow for the result of belief revision to be indeterminate in the sense that there may be many possible revision alternatives that are equally rational. Our proposal builds on the idea that what an agent wishes to achieve can play a role in the choice of which beliefs to reject and which beliefs to retain. The example we have been using in this paper also tries to capture the intuition that an agent who behaves in this manner is rational. Our richer model can distinguish one revision alternative from the other depending on the effect that each option has on the agent's goal set. Hence, under certain conditions, the choice among several revision alternatives can be reduced to one. This is what we want to investigate now, that is we want to investigate the conditions under which a revision for a CW agent is deterministic even if an indetermistic revision operator is used, i.e., $\|\mathrm{I}\{G(\mathcal{S}^i_\beta)\}_{i=1,\ldots}\| = 1$ and, for all i, j, $G(\mathcal{S}^i_\beta) \neq G(\mathcal{S}^j_\beta)$.

Observation 1 *$B * \beta$ is deterministic in state $\mathcal{S} = \langle B, O, R\rangle$, iff no two alternative revisions are equivalent, i.e., for all i, j, $B^i_\beta \not\approx B^j_\beta$.*

Proposition 2 *A sufficient condition for no two alternative revisions, B^i_β and B^j_β, being equivalent is that*

1. *for all i, j, $G(\mathcal{S}^i_\beta) \neq G(\mathcal{S}^j_\beta)$;*
2. *the importance relation on goals is strict, i.e., for all $\phi, \phi' \in G(\mathcal{S}_\beta)$, $\phi \neq \phi'$, $\phi \succeq \phi' \Rightarrow \phi' \not\succeq \phi$.*

Proof: From Hypothesis 1 and 2, by applying Definition 8, we obtain $B^i_\beta \not\approx B^j_\beta$. Therefore, no two alternative revisions can be equivalent. □

6 Conclusions

A new framework, inspired by the concept of *conventional wisdom*, aiming at dealing with indeterminism in belief revision has been proposed. While a traditional agent would not be able to choose among multiple revision candidates in indeterministic belief revision, a CW agent evaluates the effects the different revision options have on its goals and selects the revision which maximizes its achievable goals. Fundamental definitions and properties of such belief revision mechanism have been given.

References

1. Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *J. Symb. Log.*, 50(2):510–530, 1985.
2. Avrim Blum and Merrick L. Furst. Fast planning through planning graph analysis. *Artif. Intell.*, 90(1-2):281–300, 1997.
3. G. Boella, J. Hulstijn, and L. van der Torre. Interaction in normative multi-agent systems. *Electronic Notes in Theoretical Computer Science*, 141(5):135–162, 2005.
4. J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly Journal*, 2(3–4):428–447, 2002.
5. J. Broersen, M. Dastani, and L. van der Torre. Realistic desires. *Journal of Applied Non-Classical Logics*, 12(2):287–308, 2002.
6. J. Broersen, M. Dastani, and L. van der Torre. Beliefs, obligations, intentions and desires as components in an agent architecture. *International Journal of Intelligent Systems*, 20:9:893–919, 2005.
7. Michael da Costa Móra, José Gabriel Pereira Lopes, Rosa Maria Vicari, and Helder Coelho. Bdi models and systems: Bridging the gap. In *ATAL*, pages 11–27, 1998.
8. C. da Costa Pereira and A. Tettamanzi. Towards a framework for goal revision. In Wim Vanhoof Pierre-Yves Schobbens and Gabriel Schwanen, editors, *BNAIC-06, Proceedings of the 18th Belgium-Netherlands Conference on Artificial Intelligence*, pages 99–106. University of Namur, 2006.
9. M. Dastani, J. Hulstijn, and L. van der Torre. How to decide what to do? *European Journal of Operational Research*, 160(3):762–784, february 2005.
10. M. Dastani and L. van der Torre. Specifying the merging of desires into goals in the context of beliefs. In *Proceedings of in Information and Communication Technology (EurAsia ICT 2002)*, LNCS 2510, pages 824–831. Springer, 2002.
11. M. Dastani and L. van der Torre. Games for cognitive agents. In *Proceedings of JELIA04*, LNAI 3229, pages 5–17. 2004.
12. M. Dastani and L. van der Torre. What is a normative goal? towards goal-based normative agent architectures regulated agent-based systems. LNAI 2934, pages 210–227. Springer, 2004.
13. Richard Fikes and Nils J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artif. Intell.*, 2(3/4):189–208, 1971.
14. John K. Galbraith. *The Affluent Society*. Houghton Mifflin, Boston, 1958.
15. P. Gärdenfors. The dynamics of belief systems: Foundations vs. coherence. *Revue Internationale de Philosophie*, 1989.
16. Guido Governatori, Antonino Rotolo, and Vineet Padmanabhan. The cost of social agents. In *AAMAS*, pages 513–520, 2006.
17. Sven Ove Hansson. Logic of belief revision. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2006.
18. S. Lindström and W. Rabinowicz. Epistemic entrenchment with incomparabilities and relational belief revision. In A. Fuhrmann and M. Morreau, editors, *The Logic of Theory Change*, pages 93–126. 1991.
19. D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.
20. D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.
21. Felipe Rech Meneguzzi, Avelino Francisco Zorzo, and Michael da Costa Móra. Mapping mental states into propositional planning. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems*. ACM Press, 2004.

22. Felipe Rech Meneguzzi, Avelino Francisco Zorzo, and Michael Da Costa Móra. Propositional planning in BDI agents. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, pages 58–63, Nicosia, Cyprus, 2004. ACM Press.
23. Bernhard Nebel. A knowledge level analysis of belief revision. In R. Brachman, H. J. Levesque, and R. Reiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 1st International Conference*, pages 301–311, San Mateo, 1989. Morgan Kaufmann.
24. Erik J. Olsson. Lindström and Rabinowicz on relational belief revision. In T. Ronnow-Rasmussen, B. Petersson, J. Josefsson, and D. Egonsson, editors, *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*. 2007.
25. David E. Smith. Choosing objectives in over-subscription planning. In Shlomo Zilberstein, Jana Koehler, and Sven Koenig, editors, *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS 2004)*, pages 393–401, Whistler, British Columbia, Canada, June 3–7 2004. AAAI.
26. M. Birna van Riemsdijk. *Cognitive Agent Programming: A Semantic Approach*. PhD thesis, University of Utrecht, 2006.
27. A.O. Zamberlam, L.M.M. Giraffa, and C.M. Móra. X-bdi: uma ferramenta para programao de agentes bdi (in portuguese). Technical Report 9, PPGCC/PUCRS, 2000.

On the Conservativity and Stability of Ontology-Revision Operators Based on Reinterpretation

Özgür L. Özçep and Carola Eschenbach

Department for Informatics (AB WSV), MIN Faculty, University of Hamburg
Vogt-Kölln-Str. 30, D-22527 Hamburg, Germany
{oezcep,eschenbach}@informatik.uni-hamburg.de

Abstract. This article deals with the problem of integrating possibly conflicting information into an ontology. We define and analyze a family of ontology-revision operators that resolve conflicts by reinterpreting concept symbols occurring in the triggering information. The analysis of the iterated application of the operators focusses on issues of conservativity and stability of the ontology extension.

1 Introduction

Communication between natural or artificial agents relies on using shared terms with shared meanings. This precondition, however, cannot always be established in advance. While human users of natural language have flexible means to handle situations where different uses of the same term become obvious, such mechanisms of reinterpretations are not well studied for artificial agents. In this article, we are concerned with the specific case of heterogeneity between terminologies, where different agents use the same term with different meanings (conf. [1]) and where this ambiguity is discovered while the sender agent gives information that conflicts with the information the receiver holds. The approach aims at handling the communication between agents that hold kindred ontologies where conflicts are the exception rather than the rule. Therefore no preprocessing stage of aligning the terminologies in advance is assumed.

The specification of the terminology used in communication is based on an ontology the agent holds. For agents whose ontologies are consistent and well-tried the treatment of observed heterogeneities should not lead to the loss of (parts of) its former ontology. Therefore, we are faced with the problem of establishing a semantic mapping between the receiver's (internal) ontology and the sender's terminology during the exchange of information using the terms rather than the exchange of information about the terminologies. In this article we will focus on a lifting process where the incoming information is handled as a sequence of facts and the ontology of the sender is not communicated.

We outline the theoretical basis on which to generate semantic mappings as the product of applying a consistency resolving change operator to an ontology,

represent semantic mappings as description logical formulas in the object language and use them like other logical formulas as premises for inferences needed to calculate the outcome of the change operators.

The ontology-revision operators defined and analyzed in this article are motivated by ideas from the area of belief revision. Along with a treatment of iterated revision (iterated application of an revision operator), we will discuss stability aspects for the operators. There are no considerations about the semantics of our approach in this article. They are left for future work.

A concrete application of the analyzed ontology-revision operators could be to embed them into an information processing system IPS. More concretely imagine a software agent that holds an ontology O_R. The IPS formulates a query (e.g., 'List all cheap books on thermodynamics') and sends it as a request to another agent (the sender) that offers services concerning the request, e.g., services that are needed for online book stores. The sender processes the request, generates a response by using its own ontology O_S, and sends the response as a sequence of information. The IPS processes the sequence by applying the revision operator (incrementally) and, in doing so, resolves conflicts that possibly occur due to the difference between O_R and O_S, thereby, e.g., discovering that the concept *cheap* has different meanings in O_S and O_R.

2 Ontology-Revision Operators: Definitions

Following M. Grove's idea of so called sphere-based belief revision outlined in [2], Wassermann/Fermé ([3]) constructed operators for expanding, revising and contracting a set of concept descriptions by a concept description. As ontologies deal with concepts, [4] adapted these ideas in order to define ontology-revision operators that get as input an ontology O and a sentence α, also called the trigger information, and that have as output a new ontology. Two different types of operators $\odot_1$ and $\odot_2$ were defined in a local and global variant respectively. In this article only the global variants will be dealt with.

For the definition of the operators some preliminary notation is necessary. Throughout this article an ontology will be understood as a finite set of sentences over a description logical (DL) language.[1] The DL constructions and their semantics used in this article—which amounts to the DL $\mathcal{ALUCN}$— are listed in Tab. 1. The question for which specific DLs our framework is suited has to be worked out by the analysis of the operators. An interpretation $\mathcal{I}$ (or $\mathcal{M}$) is a pair consisting of the nonempty domain $\Delta^{\mathcal{I}} = \mathsf{dom}(\mathcal{I})$ and a function assigning to every constant a an element $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$, to every concept symbol K a set $K^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and to every role symbol R a relation $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$.

Concept descriptions will be denoted by C and indexed or primed variants. Concept symbols (or atomic concept descriptions) will be denoted by K, S, T and indexed or primed variants. Constants will be denoted by $a, b, c \ldots$ and indexed variants. Role symbols are denoted by R and indexed variants. An ontology will

[1] For the definitions and the syntax of DLs see [5].

Name	Syntax	Semantics
Top	$\top$	$\Delta^{\mathcal{I}}$
Bottom	$\bot$	$\emptyset$
Intersection	$C_1 \sqcap C_2$	$C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$
Union	$C_1 \sqcup C_2$	$C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$
Negation	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
Value restriction	$\forall RC$	$\{x \in \Delta^{\mathcal{I}} \mid \forall y(x,y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}$
Limited exist. quant.	$\exists R\top$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y(x,y) \in R^{\mathcal{I}}\}$
Unqual. number restriction	$\leq nR$	$\{x \in \Delta^{\mathcal{I}} \mid \lvert\{y \in \Delta^{\mathcal{I}} \mid (x,y) \in R^{\mathcal{I}}\}\rvert \leq n\}$

Table 1. DL syntax and semantics for $\mathcal{ALUCN}$

be denoted by O and indexed or primed variants. As usual the sentences of an ontology are divided into the A-box (world description) and the T-box (terminological knowledge). In this article we focus on A-boxes consisting of sentences of the form $C(a)$ or $R(a,b)$ and on T-boxes consisting of general inclusion axioms (GCI), i.e., sentences of the form $C_1 \sqsubseteq C_2$. Their semantics are respectively given by: $\mathcal{I} \models C(a)$ iff $a^{\mathcal{I}} \in C^{\mathcal{I}}$; $\mathcal{I} \models R(a,b)$ iff $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$; $\mathcal{I} \models C_1 \sqsubseteq C_2$ iff $C_1^{\mathcal{I}} \subseteq C_2^{\mathcal{I}}$. Additionally we use inequalities $a \neq b$ (with the semantic $\mathcal{I} \models a \neq b$ iff $a^{\mathcal{I}} \neq b^{\mathcal{I}}$), which are not contained in the ontologies but are used to formulate unique name assumptions. $\mathsf{Mod}(O)$ is the set of models of O, i.e., the set of interpretations for which all sentences of O are true. Sentences of the form $K(a)$ (for K a concept symbol) will be called positive literals, sentences of the form $\neg K(a)$ will be named negative literals and the union of these sets of sentences will be simply named literals.

An *ontology over a language* $\mathcal{L}$ is a set of sentences in which all non-logical symbols, i.e., the concept symbols, constants and role symbols, are among those in $\mathcal{L}$. $\mathcal{L}(O)$ describes exactly the non-logical symbols occurring in O. Writing $\alpha \in \mathcal{L}$ for a sentence α means that all non-logical symbols of α are in $\mathcal{L}$. $O_{[K_1/K_2]}$ is the outcome of uniformly substituting K_1 by K_2 in O. Sentences that occur as the second arguments of the operators are denoted by α or β and indexed variants and are called trigger information or just trigger.

The (global) operators of [4] are defined with reference to the most specific concept assigned by an ontology to a constant. C is a most specific concept (msc) for a in the ontology O iff $O \models C(a)$ and for all C' such that $O \models C'(a)$ it is true that $O \models C \sqsubseteq C'$. The existence of a most specific concept depends on the ontology O and the underlying description logic.[2] We assume that there is some systematic way (e.g. an ordering over concept descriptions) to pick out for every constant a a unique most specific concept in an ontology O. This unique most specific concept will be denoted by $\mathsf{msc}_O(a)$ and we will talk about *the* most specific concept of a constant in an ontology (or regarding an ontology).

[2] [6] describes a family of description logics for which the most specific concept exists and an algorithm for determining the most specific concept.

The main ideas underlying the definitions below are the following: If the trigger α is compatible with O, then it is just added to O to receive the new ontology $O \cup \{\alpha\}$. If α is incompatible with O, something different is done to guarantee the consistency of the resulting ontology. If, e.g. , α has the form $K(a)$, the incompatibility with the original ontology is assumed to be caused by the fact that the sender uses K to denote a more general concept than in O. This relation between the different terminological uses of K by the holder of O and the sender is expressed by a subsumption, e.g. $K \sqsubseteq K'$ in (1). Here a new symbol K' is introduced in order to disambiguate the different uses of K. In the case of the weak operators (Def. 3) there are no other assumptions concerning the relation of K and K'. For the strong operators (Def. 1, 2) additional bounds are added that try to capture the idea of minimal difference between K and K'.

The inconsistency resolving mechanism used in the definitions below, which we call *reinterpretation*, does not, in a proper sense, 'change' the original ontology as it is usually the case for classical belief revision mechanisms; but it conserves the old ontology and extends it by T-box axioms capturing the relation between the terminology of the holder of O and the sender.

Definition 1. *Let O be an ontology over a DL-Language $\mathcal{L}$, $\alpha = K(a)$ a sentence in $\mathcal{L}$ with K a concept symbol and let a be a constant for which $\mathsf{msc}_O(a)$ exists. Let K' be a new concept symbol not occurring in $O \cup \{\alpha\}$. Then the* ***global operators of type 1 and 2 (for positive literals)*** *are defined by ([4], p. 87):*

$$O \odot_1 K(a) = \begin{cases} O \cup \{K(a)\} & \text{if } O \cup \{K(a)\} \text{ is consistent,} \\ O \cup \{K \sqsubseteq K', K' \sqsubseteq K \sqcup \mathsf{msc}_O(a), & \\ \qquad K'(a)\} & \text{else} \end{cases} \tag{1}$$

$$O \odot_2 K(a) = \begin{cases} O \cup \{K(a)\} & \text{if } O \cup \{K(a)\} \text{ is consistent,} \\ O_{[K/K']} \cup \{K' \sqsubseteq K, K \sqsubseteq K' \sqcup \mathsf{msc}_{O_{[K/K']}}(a), & \\ \qquad K(a)\} & \text{else} \end{cases} \tag{2}$$

The operators $\odot_1$ and $\odot_2$ are similar as one can obtain $\odot_2$ by changing the roles of K' and K in the definition of $\odot_1$. The case that $O \cup \{\alpha\}$ is consistent is handled by both operators in the same way by adding α to O.

The difference of $\odot_1$ and $\odot_2$ comes into play in the inconsistency case. The operators $\odot_1$ and $\odot_2$ differ regarding which concept (the concept represented in O vs. the concept represented in α) is denoted by the new symbol K'. The type-1 operator $\odot_1$ substitutes the occurrence of K in the trigger information by a new symbol K' while O is not changed. We will also say that K in α is *reinterpreted*. The type-2 operator $\odot_2$ substitutes the occurrences of K in the ontology O by a new symbol K', while preserving α. We will also say that K in O is reinterpreted. The difference between $\odot_1$ and $\odot_2$ can also be described by saying that $\odot_1$ preserves the terminology of the ontology O while $\odot_2$ adapts to the terminology of the sender of α.

As a consequence, O is not changed by applying $\odot_1$ and α is put in a reinterpreted form into the resulting ontology. The operator $\odot_1$ fulfills the condition of monotonicity (see below) but only a weak form of the success postulate mentioned in the classic belief revision postulates of AGM ([7], p. 513). The operator

$\odot_2$, on the other hand, fulfills the success postulate, i.e., $\alpha \in O \odot_2 \alpha$, but not monotonicity.

Both operators declare upper and lower bounds in which the old symbol K and the new symbol K' occur. In case of $\odot_1$ the bounds are given for K' depending on K and in case of $\odot_2$ the bounds are given for K depending on K'.

Example 1. There is an asymmetry concerning how positive and negative literals are dealt with in revising with a positive literal $K(a)$. Consider, e.g., the ontology $O = \{\neg K(a), K(b), \neg K(c)\}$ and the trigger $\alpha = K(a)$. Then $\mathsf{msc}_O(a) = \mathsf{msc}_O(c) = \neg K$ and we have on the one hand $O \odot_2 K(a) \models K(b)$ (i.e. $K(b)$ is preserved along the revision) but on the other hand $O \odot_2 K(a) \not\models \neg K(c)$ (i.e. $\neg K(c)$ is not preserved). As $K(b)$ has the same prefix as the trigger (no negation), it is preserved: The K of the sender denotes a wider concept, so all individuals (like the one denoted by b) that instantiated K of the original O also do it after the change. For those individuals (like the one denoted by c) that did not instantiate K in O, the situation is different. Only if they are provably different (in O) from the individual referred to in the trigger (here a) they stay outside of K. In this example c and a are not sufficiently different, as $O \not\models \neg\mathsf{msc}_O(a)(c)$. (See also Prop. 2, (12), (13), (16), (17).)

A limitation of the definitions for $\odot_1$ and $\odot_2$ is the fact that they deal only with positive literals. In order to widen the applicability of the operators, we extend the definitions of the operators to deal also with negative literals.[3] Proceeding in this way (from the literals to more complex trigger information) we can check whether it is justifiable also to investigate operators that can handle a wider class of trigger information.

Definition 2. *Let O be an ontology over a DL-Language $\mathcal{L}$, K a concept symbol, and let a be a constant in $\mathcal{L}$ for which $\mathsf{msc}_O(a)$ exists. Let K' be a new concept symbol not occurring in $O \cup \{\neg K(a)\}$. Then the* ***global operators of type 1 and 2 (for literals)*** *are defined according to Def. 1 for the positive cases and for the negative cases by*

$$O \odot_1 \neg K(a) = \begin{cases} O \cup \{\neg K(a)\} & \text{if } O \cup \{\neg K(a)\} \text{ is consistent,} \\ O \cup \{K' \sqsubseteq K, K \sqsubseteq K' \sqcup \mathsf{msc}_O(a), \neg K'(a)\} & \text{else} \end{cases} \tag{3}$$

$$O \odot_2 \neg K(a) = \begin{cases} O \cup \{\neg K(a)\} & \text{if } O \cup \{\neg K(a)\} \text{ is consistent,} \\ O_{[K/K']} \cup \{K \sqsubseteq K', K' \sqsubseteq K \sqcup \mathsf{msc}_{O_{[K/K']}}(a), \neg K(a)\} & \text{else} \end{cases} \tag{4}$$

The use of mscs in the definitions results from the construction in [4], in which the global operators (defined above) originate as generalizations of the local operators—using the msc as a common bound for all local operators. By

[3] The extension of the definitions to other types of trigger information is more complex since it needs to handle more than one candidate for reinterpretation. To this end a choice function (or a preference relation) could to be defined that decides which concept in the trigger has to be reinterpreted.

weakening the specification, yielding definitions for the operators $\otimes_1$ and $\otimes_2$, we get rid of the reference to most specific concepts. The analysis of $\otimes_1$ and $\otimes_2$ aims at preparing the analysis of the stronger operators $\odot_1$ and $\odot_2$ and will show that these weak operators also have some undesirable properties.

Definition 3. *Let O be an ontology over a DL-Language $\mathcal{L}$, K a concept symbol and let a be a constant in $\mathcal{L}$. Let K' be a new concept symbol not occurring in $O \cup \{K(a)\}$. Then the* ***weak global operators of type 1 and 2 (for literals)*** *are defined by*

$$O \otimes_1 K(a) = \begin{cases} O \cup \{K(a)\} & \text{if } O \cup \{K(a)\} \text{ is consistent,} \\ O \cup \{K \sqsubseteq K', K'(a)\}, & \text{else} \end{cases} \tag{5}$$

$$O \otimes_1 \neg K(a) = \begin{cases} O \cup \{\neg K(a)\} & \text{if } O \cup \{\neg K(a)\} \text{ is consistent,} \\ O \cup \{K' \sqsubseteq K, \neg K'(a)\} & \text{else} \end{cases} \tag{6}$$

$$O \otimes_2 K(a) = \begin{cases} O \cup \{K(a)\} & \text{if } O \cup \{K(a)\} \text{ is consistent,} \\ O_{[K/K']} \cup \{K' \sqsubseteq K, K(a)\} & \text{else} \end{cases} \tag{7}$$

$$O \otimes_2 \neg K(a) = \begin{cases} O \cup \{\neg K(a)\} & \text{if } O \cup \{\neg K(a)\} \text{ is consistent,} \\ O_{[K/K']} \cup \{K \sqsubseteq K', \neg K(a)\} & \text{else} \end{cases} \tag{8}$$

The operators $\odot_i$ and $\otimes_i$ for $(i \in \{1, 2\})$ can be considered as special cases of the operators $\oplus_i^{\mathsf{sel}}$ defined (for positive literals) according to (9) and (10):

$$O \oplus_1^{\mathsf{sel}} K(a) = O \otimes_1 K(a) \cup \{K \sqsubseteq K' \sqcup \mathsf{sel}(\{C \mid O \models C(a)\})\} \tag{9}$$

$$O \oplus_2^{\mathsf{sel}} K(a) = O \otimes_2 K(a) \cup \{K' \sqsubseteq K \sqcup \big(\mathsf{sel}(\{C \mid O \models C(a)\})\big)_{[K/K']}\} \tag{10}$$

The operator $\oplus_i^{\mathsf{sel}}$ has a selection function sel as a parameter that, in order to warrant consistency, selects one concept $M = \mathsf{sel}(\{C \mid O \models C(a)\})$ from the set of concepts C instantiated by a in O. If sel is such that $M = \top$, one gets the operator $\otimes_2$. If sel is such that $M = \mathsf{msc}_O(a)$, one gets the operator $\odot_2$.[4]

In this article we will focus on the operators $\odot_i$ and $\otimes_i$ for $(i \in \{1, 2\})$ thereby avoiding the additional complexity due to the selection function sel.

One of the main questions of this article is how the operators behave in case a finite sequence of literals $A = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ or an infinite sequence of trigger information is to be integrated into an ontology. To formulate this question, we use some additional notation: Let $\circ \in \{\odot_1, \odot_2, \otimes_1, \otimes_2\}$ be an operator, $A = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ a finite sequence of literals. Then $O \circ A =_{\text{def}} (\ldots(O \circ \alpha_1) \circ \alpha_2) \ldots) \circ \alpha_n$ describes the outcome of iterated applications of the operator $\circ$ to the resulting ontologies and the trigger information of the sequence A. In case the sequence A is known and has length n we will use $O^{\circ(n)}$ instead of $O \circ A$ and even shorter $O^{(n)}$ if it is clear from the context which operator is meant (or if it is not relevant for which operator repeated application is considered). If $A = (\alpha_1, \ldots, \alpha_i, \ldots)$, then let $O^{\circ(i)} = (\ldots(O \circ \alpha_1) \circ \ldots \circ \alpha_i)$. If A is a sequence of length n, then A^i (for $i \leq n$) is the prefix of A of length i. The set of elements

[4] The function sel has a role similar to the role of the selection functions in partial meet revision and its special cases maxi-choice and full meet revision ([7]).

occurring in a sequence A is denoted by $\tilde{A}$. The symbol '$\circ_1$' will be used as metavariable for type-1 operators, i.e., '$\circ_1$' stands for $\odot_1$ or $\otimes_1$, and '$\circ_2$' will be used as metavariable for type-2 operators, i.e., '$\circ_2$' stands for $\odot_2$ or $\otimes_2$.

3 Monotonicity and Non-Monotonicity

The following observation directly results from the definitions of the operators:

Observation 1. *Let O be an ontology over $\mathcal{L}$, $\alpha \in \mathcal{L}$ and A a sequence of literals. Let $\circ_1$ be a type-1 operator and $\circ_2$ be a type-2 operator. Then:*

1. $O \subseteq O \circ_1 \alpha$ *(monotonicity of $\circ_1$)*
2. *For all $n \in \mathbb{N}$: $O \subseteq O^{\circ_1(n)}$ (monotonicity of iterated $\circ_1$)*
3. $O \otimes_i \alpha \subseteq O \odot_i \alpha$, *for $i \in \{1, 2\}$ ($\odot_i$ is at least as strong as $\otimes_i$)*
4. $\alpha \in O \circ_2 \alpha$ *(success for $\circ_2$)*
5. $O \circ \alpha = O \cup \{\alpha\}$ *iff* $O \cup \{\alpha\}$ *is consistent. (vacuity)*
6. $O \subseteq O \circ A = O \cup \tilde{A}$ *iff* $O \cup \tilde{A}$ *is consistent.*
7. $O \circ \alpha$ *is consistent.*[5] *(consistency)*
8. *If* $\mathsf{Mod}(O_1) = \mathsf{Mod}(O_2)$, *then* $\mathsf{Mod}(O_1 \circ \alpha) = \mathsf{Mod}(O_2 \circ \alpha)$ *(extensionality in left argument)*
9. *If $O \cup \{\alpha\}$ is inconsistent and K' is the new symbol introduced in $O \circ_2 \alpha$ resp. $O \circ_1 \alpha$, then: $(O \circ_2 \alpha)_{[K/L,K'/K]} = (O \circ_1 \alpha)_{[K'/L]}$, for $L \neq K' \notin \mathcal{L}(O \cup \{\alpha\})$ and $\alpha = K(a)$ or $\alpha = \neg K(a)$.*

Assertions 1.4,1.5, 1.7 and 1.8 of the observation are four adapted variants from six of the AGM postulates.[6] The other two postulates deal with the revision of belief sets/propositions with complex information (conjunction) which we cannot (yet) simulate in our setting as we defined the operators only for literals.

In the case of inconsistency, one can say a little bit more about the behavior of type-1 operators: The integration of α into O results in a conservative extension. According to logical terminology a theory O' in a language $\mathcal{L}'$ is called a *conservative extension of the theory O* in a language $\mathcal{L} \subseteq \mathcal{L}'$ iff for all sentences α in $\mathcal{L}$: $O \models \alpha$ iff $O' \models \alpha$.[7] The following proposition states conservativity:

Proposition 1. *Let O be an ontology over a language $\mathcal{L}$, and $\alpha \in \mathcal{L}$ be a literal. Then: If $O \cup \{\alpha\}$ is inconsistent, then $O \odot_1 \alpha$ and $O \otimes_1 \alpha$ are conservative extensions of O.*

Proof. See Appendix.

In the consistency case one cannot guarantee $O \circ_1 \alpha$ to be a conservative extension, only the property of monotonicity holds. As a consequence it is not the case that for all n: $O^{\circ_1(n)}$ is a conservative extension of O. Additionally the following observations can be made:

[5] This can be proved as a corollary to Prop. 1.
[6] Compare the (re-)formulation of the postulates in [8].
[7] [9], p. 208 and [10], p. 625.

Observation 2. *For an ontology O over $\mathcal{L}$ and literals $\alpha, \alpha_i, \alpha_j \in \mathcal{L}$:*

1. *The outcome of applying $\circ_1$ to a sequence A of literals depends on the order of the elements in A. In case of $O \models \neg(\alpha_i \wedge \alpha_j)$ for $\alpha_i, \alpha_j \in \tilde{A}$ and $i < j$ it is possible that α_i wins/survives when resolving the conflict in step j.*
2. *There is a subset $\tilde{A}'$, such that: $O \cup \tilde{A}' \subseteq O^{\circ_1(n)}$ and $O^{\circ_1(n)}$ is a conservative extension of $O \cup \tilde{A}'$.*
3. *The monotonicity of $\circ_1$ preserves conflicts: If $O \cup \{\alpha\}$ is inconsistent, then $O^{(n)} \cup \{\alpha\}$ is also inconsistent. Thus, if $O \cup \{\alpha\}$ is inconsistent, repeated occurrences of α in A never result in $O^{(n)} \models \alpha$ for any $n \in \mathbb{N}$.*

The operators of type 2 are not monotone: $O \not\subseteq O \circ_2 \alpha$ if $O \cup \{\alpha\}$ is inconsistent. Therefore the analysis of the type-2 operators is more complicate. But in combination with the fact that the success postulate is fulfilled, stability (in an intuitive sense defined below) is provable (at least for the weak type-2 operators). From now on we will concentrate on type-2 operators.

4 Detailed Analysis

4.1 Restricted Conservativity

The following proposition states restricted conservativity properties for the operators $\odot_2$ and $\otimes_2$. More precisely, (14) and (18) state conservativity for all sentences β that do not contain one of the concept symbols K, K' (directly) involved in the reinterpretation. Assertions (11) and (15) express conservativity for those sentences that are literals and in which the reinterpreted symbol K occurs with the same prefix (negation vs. no negation symbol) as in the trigger. Similarly (12) and (16) express conservativity (only in case of the strong operator $\odot_2$) for those sentences that are literals and in which the reinterpreted symbol K occurs with a different (complementary) prefix as in the trigger. Assertions (13) and (17) express the fact that $\otimes_2$ does not preserve literals in which the reinterpreted symbol K occurs with a different prefix than the prefix of the occurrence of K in the trigger.

Proposition 2. *Let a and c be constants, K be a concept symbol, O be an ontology such that $\mathsf{msc}_O(a)$ exists. Let $\mathcal{L} = \mathcal{L}(O \cup \{K(a), K(c)\})$. Then for all formula $\beta \in \mathcal{L} \setminus \{K, K'\}$:*

– *If $O \models \neg K(a)$, then:*

$$O \circ_2 K(a) \models K(c) \;\; \textit{iff} \;\; O \cup \{a \neq c\} \models K(c) \tag{11}$$

$$O \odot_2 K(a) \models \neg K(c) \;\; \textit{iff} \;\; O \models \neg K(c) \textit{ and } O \models \neg \mathsf{msc}_O(a)(c) \tag{12}$$

$$O \otimes_2 K(a) \not\models \neg K(c) \tag{13}$$

$$O \circ_2 K(a) \models \beta \;\; \textit{iff} \;\; O \models \beta \tag{14}$$

– *If* $O \models K(a)$, *then:*

$$O \circ_2 \neg K(a) \models \neg K(c) \quad \textit{iff} \quad O \cup \{a \neq c\} \models \neg K(c) \tag{15}$$

$$O \odot_2 \neg K(a) \models K(c) \quad \textit{iff} \quad O \models K(c) \textit{ and } O \models \neg \mathsf{msc}_{\mathsf{O}}(a)(c) \tag{16}$$

$$O \otimes_2 \neg K(a) \not\models K(c) \tag{17}$$

$$O \circ_2 \neg K(a) \models \beta \quad \textit{iff} \quad O \models \beta \tag{18}$$

Proof. See Appendix.

The operators $\odot_2$ and $\otimes_2$ nearly fulfill the same restricted conservativity assertions. The crucial difference is expressed by (12) and (13) (for positive literals) and (16) and (17) (for negative literals). Because of this we can infer more about $\otimes_2$ than is expressed in Prop. 2. (See Sect. 4.2 on stability.)

The conservativity properties expressed in Prop. 2 are called 'restricted' because of two reasons: 1) Conservativity holds only for a subset of the sentences (the set of literals) and 2) the 'if'-directions of two of the proposed assertions ((11), (15)) hold only with additional assumptions concerning the uniqueness of constants. These additional assumptions will be called *'local unique name assumptions'* and will be abbreviated by 'UNA'. They express for some (not all) constants occurring in the ontology and the trigger information the condition that they denote different entities.

The local unique name assumptions have a crucial role in the question of stability which we deal with in the next subsection.

4.2 Stability

The main setting we consider is that of an agent holding some ontology O and receiving a sequence A of trigger information (all being literals) and integrating them into its ontology by using an operator of type 2. If the trigger stems from the same source ontology and this ontology is consistent, also A is consistent.[8] We focus on cases for which A contains only a finite set of different literals and for which some literals can occur infinitely often in A. As the operators of type 2 fulfill success the question arises whether there is a step during integrating A from which on the ontology does not change anymore. Formally: Is there some $i \in \mathbb{N}$ such that $O^{\circ_2(i+m)} = O^{\circ_2(i)}$ for all $m \in \mathbb{N}$?

For the weak operator $\otimes_2$ stability holds under some local unique name assumptions. Stability in general does not hold without a (local) UNA. This can be demonstrated by a simple example.

Example 2. Consider the ontology $O = \{R(c,a), R(c,b), (\leq 1R)(c)\}$. It says that c is in R-relation to a and b and that there is at most one individual to which c is R-related. Thus $O \models a = b$. If A is the infinite sequence $(K(a), \neg K(b), K(a), \neg K(b), \ldots)$ (having finite different literals), then stability cannot occur. In

[8] This seems to be a plausible assumption to be found in the discussion of the belief revision community concerning the interpretation of iterated revision. See, e.g., [11].

other words: If according to the ontology of the receiver one object is denoted by different constants a, b but according to the ontology of the sender a, b denote different objects, then this mismatch cannot be solved by an operator of type 2.

The stability question for $\odot_2$ is a bit more complex because of the additional bound containing the most specific concept. The problematic fact in case of $\odot_2$ is that information integrated in one step i may disappear in a later step $i + m$ and perhaps be replaced by its negation in another (or the same) step. This is demonstrated by the following example.

Example 3. Let the ontology O and the sequence A be given by

$$O = \{\neg K(a_1), L(a_1), L(a_2)\}$$
$$A = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (\neg K(a_2), K(a_1), \neg L(a_1), \neg L(a_2))$$

Applying the definition of $\odot_2$ results in

$$O \odot_2 A \equiv \{\neg K'(a_1), L'(a_1), L'(a_2), \neg K'(a_2), K(a_1), K' \sqsubseteq K,$$
$$K \sqsubseteq K' \sqcup L', \neg L(a_1), L \sqsubseteq L', L' \sqcap (\neg K' \sqcup \neg K)) \sqsubseteq L, \neg L(a_2)\}$$

Consequently $O \odot_2 A \models \neg\alpha_1$, i.e., the trigger α_1 from the first step is abandoned in a later step and its negation follows from $O \odot_2 A$. Thus success for α_1 is not warranted.

This example does not show that stability cannot hold for $\odot_2$, but it shows that we cannot prove it by proving $O^{(i)} \not\models \neg\alpha_1$.

The main result of this article is the stability of $\otimes_2$ and can be proved as a corollary to the following theorem. The operator una used in the theorem explicates the unique name assumption implicitly contained in A:

$$\mathsf{una}(A) = \{a \neq b \mid K(a), \neg K(b) \in \tilde{A}, \text{ for a concept symbol } K\}$$

Defining $\mathsf{una}(A)$ in this way, also expresses the assumption that the set of literals in A is consistent.

Theorem 1. *Let O be a consistent ontology over $\mathcal{L}$. Then for all finite sequences A of literals in $\mathcal{L}$: If $(O \otimes_2 A) \cup \mathsf{una}(A)$ is consistent, then $(O \otimes_2 A) \cup \mathsf{una}(A) \cup \tilde{A}$ is consistent as well.*

Sketch of Proof. We need some additional notation. For a sequence $A = (\alpha_i)_{i \in I}$ and a concept symbol K let

$$A_K = \{\alpha_j \mid j \in I, \alpha_j = K(a_j) \text{ for some constant } a_j\}$$

be the set of literals contained in A in which K occurs positively. Accordingly

$$A_{\neg K} = \{\alpha_j \mid j \in I, \alpha_j = \neg K(a_j) \text{ for some constant } a_j\}$$

is the set of literals contained in A in which K occurs negatively. Let $A_{(K)} = A_K \cup A_{\neg K}$. With $O_K = \{\beta \in O \mid \beta \text{ contains } K\}$ we describe that part of the

ontology O that syntactically contains K. Let $\mathcal{K} = \{K_i \mid i \in I\}$ be the set of all concept symbols in $\mathcal{L}$ for some $I \subseteq \mathbb{N}$.

The main ideas in the proof outlined in the following are first to separate the ontologies in different parts according to the concept symbols and second to check the following two facts: 1) If a conflict resolution for a literal $\alpha_i = K(a)$ is done in step i, then a conflict resolution for a literal α_j (integrated in step $j > i$) containing the same concept symbol K can only occur if α_j has the form $\neg K(b)$. (Accordingly if $\alpha_i = \neg K(a)$, then α_j must have the form $K(b)$.) 2) There can be at most two conflict resolutions with respect to the same concept symbol.

The proof is done by induction on the length of A. In fact the assertion proved by induction is stronger than the one formulated in the theorem, and it contains the assertion of the theorem as (the last) conjunctive part. The assertion is:

For all finite sequences A of length n: There are two disjoint sets of concept symbols $\mathcal{K'}_n$ and $\mathcal{K''}_n$ that are disjoint from $\mathcal{K}$ and have the form $\mathcal{K}'_n = \{K'_i \mid i \in I'_n\}$ and $\mathcal{K}''_n = \{K''_i \mid i \in I''_n\}$ for $I''_n \subseteq I'_n \subseteq I$. And there is a substitution σ_n defined by $K_i\sigma_n = K'_i$ if $i \in I'_n$ and $K_i\sigma_n = K_i$ else, such that the following five assertions holds:

1. $O\sigma_n \subseteq O^{(n)}$
   This expresses the fact that the (original) ontology O in some way is preserved along the integration. It can be found in the resulting ontology $O^{(n)}$ by applying the substitution σ_n which maps the concept symbols of the old ontology onto the corresponding (primed new) symbols of the new ontology and thereby acts like a semantic mapping.
2. All concept symbols contained in $O^{(n)}$ are contained in $\mathcal{K} \cup \mathcal{K}'_n \cup \mathcal{K}''_n$.
3. $O^{(n)}$ can be represented by

$$O^{(n)} = O\sigma_n \cup \overbrace{\bigcup_{i \in I \setminus I'_n} A_{(K_i)}}^{\text{no revision}} \cup \overbrace{\bigcup_{i \in I'_n \setminus I''_n} \left(O^{(n)}_{K_i} \cup (A_{(K_i)} \setminus O^{(n)}_{K_i})\sigma_n\right)}^{\text{simple revision}} \cup \underbrace{\bigcup_{i \in I''_n} \left(O^{(n)}_{K_i} \cup O^{(n)}_{K''_i} \cup \left(A_{(K_i)} \setminus O^{(n)}_{K_i} \setminus O^{(n)}_{K''_i[K''_i/K_i]}\right)\sigma_n\right)}_{\text{twofold revision}}$$

   As the comments under and over the cambered brackets suggest, there can be maximally two conflict resolutions with respect to the same concept symbol.
4. For all $i \in I$:
   (a) If $i \notin I'_n$, then $A_{(K_i)} \subseteq O^{(n)}$.
   (b) If $i \in I'_n \setminus I''_n$, then $(A_{(K_i)} \setminus O^{(n)}_{K_i})\sigma_n \subseteq O^{(n)}$ and there is exactly one T-box axiom in $O^{(n)}$ of the form $K'_i \sqsubseteq K_i$ (case (A)) or $K_i \sqsubseteq K'_i$ (case (B)).
   (A) In this case additionally
   $O^{(n)}_{K_i} \subseteq \{K'_i \sqsubseteq K_i\} \cup A_{(K_i)}$ and if $O^{(n)} \models \neg K_i(a_j)$, then $K_i(a_j) \notin \tilde{A}$.
   (B) In this case additionally
   $O^{(n)}_{K_i} \subseteq \{K_i \sqsubseteq K'_i\} \cup A_{(K_i)}$ and if $O^{(n)} \models K_i(a_j)$, then $\neg K_i(a_j) \notin \tilde{A}$.

(c) If $i \in I''_n$, then there is $K''_i \in \mathcal{K}''_n$ such that $(A_{(K_i)} \backslash O^{(n)} \backslash O^{(n)}_{K''_i[K''_i/K_i]})\sigma_n \subseteq O^{(n)}$ and there is exactly one T-box axiom of the form $K''_i \sqsubseteq K_i$ (case (A)) or of the form $K_i \sqsubseteq K''_i$ (case (B)).

(A) In this case additionally
- $K''_i \sqsubseteq K'_i \in O^{(n)}$ and $O^{(n)}_{K_i} \subseteq \{K''_i \sqsubseteq K_i\} \cup A_{(K_i)}$
- $O^{(n)}_{K''_i} \subseteq \{K''_i \sqsubseteq K'_i, K''_i \sqsubseteq K'_i\} \cup (A_{(K_i)})_{[K_i/K''_i]}$
- $O^{(n)} \models (A_{\neg K_i})_{[K_i/K''_i]} \cup (A_{K_i} \cap O^{(n)}_{K_i}) \cup (A_{K_i} \setminus O^{(n)}_{K_i})\sigma_n$
- If $O^{(n)} \models \neg K_i(a_j)$, then $K_i(a_j) \notin \tilde{A}$ and
- if $O^{(n)} \models K_i(a_j)$, then $\neg K_i(a_j) \notin \tilde{A}$.

(B) In this case additionally
- $K'_i \sqsubseteq K''_i \in O^{(n)}$ and $O^{(n)}_{K_i} \subseteq \{K_i \sqsubseteq K''_i\} \cup A_{(K_i)}$
- $O^{(n)}_{K''_i} \subseteq \{K_i \sqsubseteq K''_i, K'_i \sqsubseteq K''_i\} \cup (A_{(K_i)})_{[K_i/K''_i]}$
- $O^{(n)} \models (A_{K_i})_{[K_i/K''_i]} \cup (A_{\neg K_i} \cap O^{(n)}_{K_i}) \cup (A_{\neg K_i} \setminus O^{(n)}_{K_i})\sigma_n$
- If $O^{(n)} \models \neg K_i(a_j)$, then $K_i(a_j) \notin \tilde{A}$ and
- if $O^{(n)} \models K_i(a_j)$, then $\neg K_i(a_j) \notin \tilde{A}$.

5. If $(O \otimes_2 A) \cup \mathsf{una}(A)$ is consistent, then also $(O \otimes_2 A) \cup \mathsf{una}(A) \cup \tilde{A}$ is consistent.

The proof of the 5th assertion relies on the assertions before and is done by a model construction which completes the proof. Let $\mathcal{M}$ be a model of $(O \otimes_2 A) \cup \mathsf{una}(A)$. We construct a model $\mathcal{M}'$ of $(O \otimes_2 A) \cup \mathsf{una}(A) \cup \tilde{A}$ as follows:

- $\mathsf{dom}(\mathcal{M}') = \mathsf{dom}(\mathcal{M}) = D$; $\mathcal{M}'(a) = \mathcal{M}(a)$ for all constants a;
- $\mathcal{M}'(R) = \mathcal{M}(R)$ for all role symbols R; $\mathcal{M}'(K'_i) = \mathcal{M}(K'_i)$ for all $i \in I'_n$
- $\mathcal{M}'(K_i) = \begin{cases} \mathcal{M}(K_i) & \text{if } i \notin I'_n \\ \mathcal{M}(K'_i) \setminus \{\mathcal{M}(a_i) \mid \neg K_i(a_j) \in \tilde{A}\} & \text{if } i \in I'_n \setminus I''_n \text{ and } K_i \sqsubseteq K'_i \in O^{(n)} \\ \mathcal{M}(K'_i) \cup \{\mathcal{M}(a_i) \mid K_i(a_j) \in \tilde{A}\} & \text{if } i \in I'_n \setminus I''_n \text{ and } K'_i \sqsubseteq K_i \in O^{(n)} \\ D \setminus \{\mathcal{M}(a_j) \mid \neg K_i(a_j) \in \tilde{A}\} & \text{if } i \in I''_n \end{cases}$
- $\mathcal{M}'(K''_i) = \begin{cases} \mathcal{M}(K'_i) \setminus \{\mathcal{M}(a_i) \mid \neg K_i(a_j) \in \tilde{A}\} & \text{if } K''_i \sqsubseteq K'_i \in O^{(n)} \\ \mathcal{M}(K'_i) \cup \{\mathcal{M}(a_i) \mid K_i(a_j) \in \tilde{A}\} & \text{if } K'_i \sqsubseteq K''_i \in O^{(n)} \end{cases}$

The theorem does not state success to be fulfilled with respect to a sequence A, i.e., it is not generally the case that $\tilde{A} \subseteq O \otimes_2 A$, but it states that a weakening of success is true in the sense that $O \otimes_2 A$ is at least compatible with $\tilde{A}$. But note that the interpretation of concept symbols that are subject to two revisions (K_i with $i \in I''_n$) solely depends on A and is completely independent of the original ontology. Therefore, further investigations on the behavior of the more complex operators of type 2 are called for.

As a corollary of the theorem the stability of $\otimes_2$ results.

Corollary 1. *Let O be a consistent ontology and A an infinite sequence of literals containing a finite amount of different literals. Then if for all $j \in \mathbb{N}$ $O^{(j)} \cup \mathsf{una}(A)$ is consistent, there is a step $i \in \mathbb{N}$ such that*

$$O^{\otimes_2(i+m)} = O^{\otimes_2(i)} \text{ for all } m \in \mathbb{N}.$$

5 Related Work

Among the approaches that deal with belief-revision techniques to solve problems from the field of semantic integration, [12] and especially [13] are most closely connected to our approach.

The idea of reinterpreting concepts is similar to the idea of weakening A-box axioms in [13] adapted from [12]. The authors of [13] describe operators for revising a consistent DL knowledge base KB by a another knowledge base KB' that contains at least one A-box axiom involved in the inconsistency. In the refined version of the operator, sentences of KB that are in conflict with those in KB' are replaced by some weakened versions. The leading idea behind the weakening strategy is to consider the cases that lead to the conflict as exceptions.

The main differences between [12], [13] and our approach are that our conflict resolution is done by weakening a concept rather than by weakening sentences of the knowledge base. We focus on literals as triggering information whereby the construction of [12], [13] handles knowledge bases consisting of more complex sentences. We consider iterated applications on a sequence of literals while [12], [13] consider the revision with a set of sentences. Finally our conflict resolution involves a language extension that makes it possible to preserve the old ontology (knowledge base) and declare relations between the old and the new concepts.

6 Conclusion

The analysis of the type-2 operators yields restricted conservativity results and a stability theorem (for the weak version $\otimes_2$). The property of (restricted) conservativity in the inconsistency case is a form of informational conservativity as mentioned in the discussion of rationality postulates[9] for revision operators; this property offers the possibility to use the operators in those areas of information processing that include *refinement* as a main operation.[10]

The property of being stable makes the behavior of the (weak) operators of type 2 predictable. Coming back to the intended application scenario of an information processing system IPS with the embedded operator $\otimes_2$, this means that if we want a predictable behavior of the IPS, we should at least demand two conditions to be fulfilled in the scenario: 1) There should be only finitely many different literals in the sequence A of triggering literals. 2) The sequence A should be consistent. Scenarios in which both conditions are likely to be fulfilled

[9] [14], p. 52–61.

[10] See [10] for a discussion of refinement.

are those in which A stems from a single sender whose knowledge base (ontology) is consistent. Scenarios in which A consists of trigger information from different senders consistency of A is likely not to be fulfilled. For those scenarios type-1 operators could be more appropriate than type-2 operators.

Theorem 1 only asserts compatibility of $O \otimes_2 A$ and $\tilde{A}$ but not success for the whole sequence A (in the sense that $\tilde{A} \subseteq O \otimes_2 A$). This weakness could be compensated by equipping an IPS with an additional memory in which all literals of A are stored and put into $O \otimes_2 A$ after the last literal of A was received.

Example 2 demonstrated the importance of (local) unique name assumptions without which stability is not warranted, and in fact the theorem presupposes the unique name assumption $\mathsf{una}(A)$. So a correctly working IPS would have to check the violation of the UNA and report it. (But this is not handled yet).

Appendix: Proofs

Proof of Prop. 1. Let $\circ_1 \in \{\odot_1, \otimes_1\}$ and $K' \notin \mathcal{L}$.

If β is a sentence in $\mathcal{L}$ and $O \models \beta$, then also $O \circ_1 \alpha \models \beta$, because $O \subseteq O \circ_1 \alpha$. Now suppose that $O \not\models \beta$ for $\beta \in \mathcal{L}$. We show the proposition for positive literals $\alpha = K(a)$. We have to show that $O \circ_1 K(a) \not\models \beta$. By assumption, there is a model $\mathcal{M} \models O \cup \neg\beta$ over $\mathcal{L}$. Define $\mathcal{M}'$ for the language $\mathcal{L}' = \mathcal{L} \cup \{K'\}$ as an extension of $\mathcal{M}$ with $\mathsf{dom}(\mathcal{M}) = \mathsf{dom}(\mathcal{M}')$, $\mathcal{M}'(S) = \mathcal{M}(S)$ for all symbols S different from K' and $\mathcal{M}'(K') = \mathcal{M}(K) \cup \{\mathcal{M}(a)\}$. Then $\mathcal{M}' \models O \odot_1 K(a) \cup \{\neg\beta\}$ and $\mathcal{M}' \models O \otimes_1 K(a) \cup \{\neg\beta\}$ because per definition $O \odot_1 K(a) = O \cup \{K \sqsubseteq K', K' \sqsubseteq K \sqcup \mathsf{msc}_O(a), K'(a)\}$ and $O \otimes_1 K(a) = O \cup \{K \sqsubseteq K', K'(a)\}$ and:
$\mathcal{M}' \models O \cup \{\neg\beta\}$, because $K' \notin \mathcal{L}$ and $\mathcal{M}'$ is the same as $\mathcal{M}$ for all symbols in $\mathcal{L}$, and $\mathcal{M} \models O \cup \{\neg\beta\}$;
$\mathcal{M}' \models (K \sqsubseteq K') \wedge (K' \sqsubseteq K \sqcup \mathsf{msc}_O(a)) \wedge (\mathcal{M}' \models K'(a))$ (construction of $\mathcal{M}'$).
The proof for negative literals $\alpha = \neg K(a)$ is done similarly by constructing a new model $\mathcal{M}'$ from a model $\mathcal{M} \models O \cup \{\neg\beta\}$ setting $\mathcal{M}'(K') = \mathcal{M}(K) \setminus \{\mathcal{M}(a)\}$.

Proof of Prop. 2. The proofs for the assertions in which $\circ_2$ is mentioned, i.e. (11), (14), (15), (18), will be done by proving it either for $\odot_2$ or for $\otimes_2$. The proof for the other operator then follows as a corollary using Obs. 1.

In the proofs the substitution $\sigma = [K/L, K'/K]$ will be used. Because of the fact that $O \subseteq (O \odot_2 K(a))_\sigma$ (see Obs. 1.6), the transformations of the models constructed in the proofs will be more readable. We will use the fact that for all formulas F that do not contain L, F has a satisfying model iff F_σ has one.

Proof of (11): First assume $O \cup \{a \neq c\} \models K(c)$. Then also $O \circ_2 K(a) \cup \{a \neq c\} \models K'(c)$ and since $K' \sqsubseteq K \in O \circ_2 K(a)$ also $O \circ_2 K(a) \cup \{a \neq c\} \models K(c)$. Now let $\mathcal{M}$ be a model of $O \circ_2 K(a)$. If $\mathcal{M}(a) \neq \mathcal{M}(c)$, then $\mathcal{M} \models a \neq c$, and $\mathcal{M} \models K(c)$ follows. If, on the other hand, $\mathcal{M}(a) = \mathcal{M}(c)$, then because of $K(a) \in O \circ_2 K(a)$ also $\mathcal{M}(c) \in \mathcal{M}(K)$ results, i.e., $\mathcal{M} \models K(c)$.

Now assume $O \cup \{a \neq c\} \not\models K(c)$. Let $\mathcal{M}$ be a model of $O \cup \{a \neq c, \neg K(c)\}$. Consequently $\mathcal{M}(a) \neq M(c)$ and $\mathcal{M}(c) \notin \mathcal{M}(K)$. We have to show $O \odot_2 K(a) \not\models$

$K(c)$. Applying the substitution σ to both sides of the entailment results in the task to show

$$O \cup \{L(a), K \sqsubseteq L, L \sqsubseteq K \sqcup \mathsf{msc}_{\mathsf{O}}(a)\} \not\models L(c) \tag{19}$$

Construct a new model $\mathcal{M}'$ over $\mathcal{L}'' = \mathcal{L} \cup \{L\}$ from $\mathcal{M}$ as follows: $\mathsf{dom}(\mathcal{M}') = \mathsf{dom}(\mathcal{M})$, $\mathcal{M}'(S) = \mathcal{M}(S)$ for all symbols $S \in \mathcal{L}$ and $\mathcal{M}'(L) = \mathcal{M}(K) \cup \{\mathcal{M}(a)\}$. Then $\mathcal{M}'$ is a model of $O \cup \{\neg K(c)\}$ and additionally a model of $\{L(a), K \sqsubseteq K, L \sqsubseteq K \sqcup \mathsf{msc}_{\mathsf{O}}(a), \neg L(c)\}$ showing (19). Applying Obs. 1.3 results in $O \otimes_2 K(a) \not\models K(c)$.

Proof of (12): First assume $O \models \neg K(c)$ and $O \models \neg \mathsf{msc}_{\mathsf{O}}(a)(c)$. Then $O \odot_2 K(a) \models \neg K'(c)$ and because of $((K \sqcap \neg \mathsf{msc}_{\mathsf{O}}(a)) \sqsubseteq K') \in O \odot_2 K(a)$ also $O \odot_2 K(a) \models (\neg K \sqcup \mathsf{msc}_{\mathsf{O}}(a))(c)$ so that $O \odot_2 K(a) \models \neg K(c)$.

Now we want to show, if $O \not\models \neg K(c)$, then $O \odot_2 K(a) \not\models \neg K(c)$ and if $O \not\models \neg \mathsf{msc}_{\mathsf{O}}(a)(c)$, then $O \odot_2 K(a) \not\models \neg K(c)$.
Assume $O \not\models \neg K(c)$. Let $\mathcal{M}$ be a model of $O \cup \{K(c)\}$ and construct $\mathcal{M}'$ as an extension of $\mathcal{M}$ with $\mathcal{M}'(L) = \mathcal{M}(K) \cup \{\mathcal{M}(c)\}$. Then $\mathcal{M}'(c) \in \mathcal{M}'(L)$ and $\mathcal{M}' \models (O \odot_2 K(a))_\sigma$ and so also $\mathcal{M}' \models (O \odot_2 K(a) \cup \{K(c)\})_\sigma$ resulting in $O \odot_2 K(a) \not\models \neg K(c)$.
Assume $O \not\models \neg \mathsf{msc}_{\mathsf{O}}(a)(c)$. Let $\mathcal{M}$ be a model of $O \cup \{\mathsf{msc}_{\mathsf{O}}(a)(c)\}$. Construct $\mathcal{M}'$ as an extension of $\mathcal{M}$ by setting $\mathcal{M}'(L) = \mathcal{M}(K) \cup \{\mathcal{M}(a), \mathcal{M}(c)\}$. Then as above $\mathcal{M}' \models (O \odot_2 K(a) \cup \{K(c)\})_\sigma$ and $O \odot_2 K(a) \not\models \neg K(c)$ results.

Proof of (13): Let $\mathcal{M} \models O \otimes_2 K(a)$; then the new model $\mathcal{M}'$ defined by $\mathsf{dom}(\mathcal{M}') = \mathsf{dom}(\mathcal{M})$, $\mathcal{M}'(S) = \mathcal{M}(S)$ for all symbols S different from K and $\mathcal{M}'(K) = \mathsf{dom}(\mathcal{M})$ is a model of $O \otimes_2 K(a)$ and of $K(c)$. (Remember that $K' \sqsubseteq K$ and $K(a)$ are the only formula of $O \otimes_2 K(a)$ that involve K.)

Proof of (14): As $K, K' \notin \beta$ we have $\beta\sigma = \beta$. First assume $O \models \beta$. We have to show $O \circ_2 K(a) \models \beta$. Applying σ this reduces to showing $(O \circ_2 K(a))_\sigma \models \beta$. But this is the case because of $O \subseteq (O \circ_2 K(a))_\sigma$ and the monotonicity of $\models$.

Now assume $O \circ_2 K(a) \models \beta$ for $\odot_2$ in place of $\circ_2$, i.e., applying σ again suppose that the following entailment holds:

$$O \cup \{L(a), K \sqsubseteq L, L \sqsubseteq K \sqcup \mathsf{msc}_{\mathsf{O}}(a)\} \models \beta \tag{20}$$

Let $\mathcal{M}$ be a model over $\mathcal{L}(O \cup \{\beta\})$ of O. Extend $\mathcal{M}$ to $\mathcal{M}'$ by setting $\mathcal{M}'(L) = \mathcal{M}(K) \cup \{\mathcal{M}(a)\}$. Then $\mathcal{M}' \models O \cup \{L(a), K \sqsubseteq L, L \sqsubseteq K \sqcup \mathsf{msc}_{\mathsf{O}}(a)\}$ and hence $\mathcal{M}' \models \beta$. As $\mathcal{M}$ is the reduct of $\mathcal{M}'$ to $\mathcal{L}(O \cup \{\beta\})$ also $\mathcal{M} \models \beta$. We have shown the assertion that if $O \odot_2 K(a) \models \beta$, then $O \models \beta$. The assertion for $\otimes_2$ in place of $\odot_2$ follows with Obs. 1.3.

The proofs of (15), (16) and (18) are similar. For (15) and (18) one constructs $\mathcal{M}'$ from a model $\mathcal{M} \models O \cup \{a \neq c, K(c)\}$ by setting $\mathcal{M}'(L) = \mathcal{M}(K) \setminus \{\mathcal{M}(a)\}$. For the proof of (16) one constructs the extension $\mathcal{M}'_1$ of $\mathcal{M}_1 \models O \cup \{K(c)\}$ by setting $\mathcal{M}'_1(L) = \mathcal{M}_1(K) \setminus \{\mathcal{M}_1(a)\}$. And one constructs the extension $\mathcal{M}'_2$ of $\mathcal{M}_2 \models O \cup \{\mathsf{msc}_{\mathsf{O}}(a)(c)\}$ by setting $\mathcal{M}'_2(L) = \mathcal{M}_2(K) \setminus \{\mathcal{M}_2(a), \mathcal{M}_2(c)\}$.

Proof of (17): Let $\mathcal{M} \models O \otimes_2 \neg K(a)$; then the new model $\mathcal{M}'$ defined by $\mathsf{dom}(\mathcal{M}') = \mathsf{dom}(\mathcal{M})$, $\mathcal{M}'(S) = \mathcal{M}(S)$ for all symbols S different from K and $\mathcal{M}'(K) = \emptyset$ is a model of $O \otimes_2 K(a)$ and of $\neg K(c)$.

Acknowledgments. We want to thank the anonymous reviewers whose valuable comments we tried to integrate in the last version of the article.

References

1. Noy, N.F.: Semantic integration: A survey of ontology-based approaches. SIGMOD Record **33**(4) (2004) 65–70
2. Grove, A.: Two modellings for theory change. Journal of Philosophical Logic **17** (1988) 157–170
3. Wassermann, R., Fermé, E.: A note on prototype revision (1999) *"Spinning Ideas". (Electronic Essays dedicated to Peter Gärdenfors on his 50th Birthday).* http://www.lucs.lu.se/spinning/.
4. Özçep, Ö.L.: Ontology revision through concept contraction. In Artemov, S., Parikh, R., eds.: Proceedings of the Workshop on Rationality and Knowledge, 18th European Summerschool in Logic, Language, and Information, Universidad de Malaga, 7–11 August. (2006) 79–90
5. Baader, F.: Description logic terminology. In Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook. Cambridge University Press (2002) 495 – 505
6. Küsters, R., Molitor, R.: Computing most specific concepts in description logics with existential restrictions. LTCS-Report 00-05, LuFG Theoretical Computer Science, RWTH Aachen, Germany (2000)
7. Alchourrón, C., Gaerdenfors, P., Makinson, D.: On the logic of theory change: partial meet contraction and revision functions. Journal of Symbolic Logic **50** (1985) 510–530
8. Katsuno, H., Mendelzon, A.: On the difference between updating a knowledge base and revising it. In Allen, J.F., Fikes, R., Sandewall, E., eds.: KR'91: Principles of Knowledge Representation and Reasoning. Morgan Kaufmann, San Mateo, California (1991) 387–394
9. Monk, D.J.: Mathematical Logic. Springer (1976)
10. Antoniou, G., Kehagias, A.: A note on the refinement of ontologies. International Journal of Intelligent Systems **15**(7) (2000) 623–632
11. Delgrande, J.P., Dubois, D., Lang, J.: Iterated revision as prioritized merging. In Doherty, P., Mylopoulos, J., Welty, C.A., eds.: KR, AAAI Press (2006) 210–220
12. Meyer, T., Lee, K., Booth, R.: Knowledge integration for description logics. In Veloso, M.M., Kambhampati, S., eds.: AAAI, AAAI Press / The MIT Press (2005) 645–650
13. Qi, G., Liu, W., Bell, D.: A revision-based approach to handling inconsistency in description logics. In: Proceedings of the 11th International workshop on Non-Monotonic Reasoning (NMR06). (2006)
14. Gärdenfors, P., Rott, H.: Belief revision. In Gabbay, D., Hoger, C., Robinson, J., eds.: Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 4. Oxford University Press (1995) 35–132

Dynamic *T-Box*-Handling in Agent-Agent-Communication

Moritz Goeb, Peter Reiss, Bernhard Schiemann, and Ulf Schreiber

Artificial Intelligence Division, Department of Computer Science,
University of Erlangen–Nuremberg
{simogoeb,siulschr}@stud.informatik.uni-erlangen.de
{reiss,schieman}@informatik.uni-erlangen.de

Abstract. FIPA ACL speech acts are the core of ontology based communication between two agents in a multi agent system (MAS). We formulate the content of an ACL Message in W3Cs *OWL-DL* as ontological encoding of the content field and use the tableau reasoner RACER for validation. If validated, the contents can be added directly into receiver agents Knowledge Base (KB), respectively its *T-Box*/*A-Box*. What if the sender agent has extended his *T-Box* before so that terminology used in the message content is unknown to the receiver? What if the message content is referring to concepts that have been extended with additional restrictions in the receiver *T-Box*?
We draw up a solution based on finding semantic 'agreement' between the *T-Box* of the sender and the *T-Box* of the receiver.

Key words: *OWL-DL*, FIPA ACL, content language, agent, T-Box, merging

1 Introduction

In recent years, software agents (abbr. agents) have become a well studied and frequently applied implementation technique for distributed systems. From a users perspective, agents enable innovative capabilities like proactivity, reactivity, social ability and autonomy. For research work in AI, such agents build systems that are considered as multi agent systems (MAS).

One characteristic of agents in a MAS is the capability to communicate with each other, which enables social abilities. A specific language protocol was defined for communication, the FIPA Agent Communication Language (FIPA ACL). A FIPA ACL conformant message contains a certain set of parameters: The sender, the intended receiver and a conversation ID to identify the message uniquely are given. A performative parameter is set to specify the type of communicative act. The definitions for performatives are influenced by the *Conversation Acts Theory* [15], which classifies natural language utterances with respect to their role in a discourse. Besides these parameters that contain metainformation for the communication protocol, every message has a content field, which contains the actual information and can be seen as the payload of the message. The format of the content is not specified, although there are several proposals for content languages. The FIPA SL Content Language Specification [6] proposes a very expressive language, which unfortunately comes at the price of making the

interpretation of an incoming message very complex and even allows the formulation of undecidable message contents. One main criticism of FIPA ACL is the fact that the semantics of a message is somewhat spread between the performative and the content. E.g., the performative *request*– which is intended to request the receiver to perform some action – might contain a complete ACL message in its content field that should be forwarded to another agent under certain circumstances. Also, the definitions of some performatives require a certain minimum expressivity of the content language used. In [13], we motivate our approach to use *OWL-DL* [9] as content language in FIPA messages. *OWL-DL* is designated to describe the concepts in the emerging Semantic Web. It has clearly defined and decidable semantics, while its expressivity is sufficient for a lot of tasks. Two agents that talk to each other using this language will be able to use all the tools that are already available or will be built to deal with *OWL-DL*. Since the *OWL-DL*-encoded content holds the semantics, a small subset of FIPA ACL performatives is sufficient for communication. Currently, we support the performatives *inform* and *query-ref*, the algorithm using this performatives is described in section 2.

In this paper, we focus on agent-agent-communication with respect to its social ability in terms of cooperativity and autonomy. For communication, formal ontologies are used to provide the concepts, that are required to interpret the "content words" of the messages.

There is more than one definition of the term "ontology". We take a *OWL-DL* document as a formal ontology and follow [2] to use it setting up a Knowledge Base (KB). In addition, we use the definition of a knowledge based agent given by [12], according to which an agent uses its KB containing theorems to reason about the application domain. Combining these two approaches, each agent of an MAS holds a KB based on the domain ontology (application ontology) formulated in *OWL-DL*. Unlike other knowledge representation languages, *OWL-DL* builds on the RDF format.

2 Setting: *OWL-DL* and agent-agent-comunication

In [13], we describe a method for using *OWL-DL* as a FIPA content language. We introduce a separate *OWL-DL* KB for each agent. Using *OWL-DL* not only for message contents but also for the representation of the knowledge of the agents avoids unnecessary translations between different knowledge representation languages when message contents are transferred into the KB.

Here, an agent's KB consists of separate representations for the *T-Box*, which holds the domain specification, and the *A-Box*, where assertions about the current state of the world are stored. The DIG interface [3] is employed to access an *OWL-DL* reasoner like RACER [11], in order to do *A-Box*-*T-Box*-reasoning for inference and consistency checks.

While there are no restrictions on the use of *OWL-DL* in the *T-Box* part, the *A-Box* representation is simplified by disallowing class membership assertions in anonymous concepts like restrictions. Instead, membership in restrictions can only be asserted indirectly by referencing a named concept that is defined as a subset of the restriction inside the *T-Box*. This restriction of *A-Box* expressivity might seem undesirable for some, but it greatly simplifies the task of understanding the meaning of an *A-Box*.

2.1 Communication model

In general, the communication model in [13] enables agents to use *OWL-DL* as a content language for exchanging information about the current state of the world. More specifically, the performatives *query-ref* and *inform* (see [5] for the definition of FIPA performatives) are used to exchange *A-Box* assertions. According to [5], the contents of *inform* messages are propositions. Applied to *OWL-DL* as a content language this means that the *A-Box* assertions in the content field of a message are taken as sentences stating that all those assertions are true.

query-ref messages are supposed to contain referential expressions that specify which propositions should be contained in the content of an *inform* message sent in reply to the *query-ref* messages. Thus, an *OWL-DL* document contained in the content field of such a *query-ref* message contains temporary concepts, each of which represents a referential expression. These concepts define necessary and sufficient criteria for concept membership. The actual *A-Box* assertions referenced by these concepts are those assertions that describe one or more individuals that can be identified to be a member of those concepts.

In the following we will focus on a pair of two agents of a MAS that are trying to transfer some of the knowledge from the *A-Box* of the first agents KB to the KB of the other agent. As we will see, this sometimes makes it necessary to change the *T-Boxes* of one or even both of the agents. Knowledge transfer is done by performing an *inform* speech act, as described in [13]. The agents will be identified as the sender and the receiver, based on which part they act in the execution of that initial *inform* speech act, even if subsequent messages might be sent in the opposite direction.

2.2 Dynamic *OWL-DL T-Boxes*

The *T-Box* contains concept and role definitions, e.g. $\mathsf{Animal} \sqsubseteq \mathsf{LivingMatter}$[1] or $(\exists \mathsf{born-by}.\top) \sqsubseteq \mathsf{Animal}$ and $\top \sqsubseteq (\forall \mathsf{born-by}.\mathsf{Animal})$[2]. In contrast, the *A-Box* contains only membership definitions like e.g. $\mathsf{DOGCLAIRE123} : \mathsf{Animal}$[3].

$$\mathsf{Animal} \sqsubseteq \neg(\exists \mathsf{use}.\mathsf{Photosynthesis} \sqcup \mathsf{Plant} \sqcup \mathsf{Fungus}) \quad (1)$$

$$\mathsf{Plant} \sqsubseteq \neg\mathsf{Fungus} \quad (2)$$

$$\mathsf{Animal} \sqsubseteq \mathsf{LivingMatter} \quad (3)$$

Fig. 1. A *T-Box* defining animals

A dynamic *OWL-DL T-Box* is a *T-Box* where it is allowed to add (expand) or remove/replace (contract/revise/update) expressions. The *T-Box* expressions in figure 1

[1] This expression defines that the concept Animal is a subconcept of the concept LivingMatter

[2] The two expressions in conjunction define that the domain of the role born − by has the concept Animal as domain and range, which means that Animal could only be born − by Animal.

[3] DOGCLAIRE123 is an instance of Animal

define the concept of an animal as a biologist would define it nowadays. Let expression 3 be seen as a stable (not changing, invariant) part of the *T-Box*. Figure 2 shows

$$\mathsf{Animal} \sqsubseteq \forall \mathsf{eat}.(\mathsf{Animal} \sqcup \mathsf{Plant}) \tag{4}$$
$$\mathsf{Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{5}$$

Fig. 2. *T-Box* expression added to make the definition of animals more precise

one of the possibilities of expressions that could be added to the *T-Box* from Figure 1. As stated in subsection 2.1, there are multiple (two) agents with separate KBs and therefore separate *T-Boxes*. Parallel to the *T-Box* presented in Figure 2, a second dynamic *T-Box* (belonging to the second agent) including the expressions from Figure 1 and the expressions shown in Figure 3 exists. These two *T-Boxes* are not in conflict with each

$$\mathsf{Animal} \sqsubseteq \exists \mathsf{need}.\mathsf{Oxygen} \tag{6}$$

Fig. 3. A second *T-Box* expression added to precise the definition of animals

other so that they can easily be merged by extending one with the other. This simple

$$\mathsf{Animal} \sqsubseteq \neg(\exists \mathsf{use}.\mathsf{Photosynthesis} \sqcup \mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{7}$$
$$\mathsf{Plant} \sqsubseteq \neg \mathsf{Fungus} \tag{8}$$
$$\mathsf{Animal} \sqsubseteq \mathsf{LivingMatter} \tag{9}$$
$$\mathsf{Animal} \sqsubseteq \forall \mathsf{eat}.(\mathsf{Animal} \sqcup \mathsf{Plant}) \tag{10}$$
$$\mathsf{Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{11}$$
$$\mathsf{Animal} \sqsubseteq \exists \mathsf{need}.\mathsf{Oxygen} \tag{12}$$

Fig. 4. The merged *T-Box* defining animals

example (Figure 4) shows that two *T-Boxes* that have been concurrently extended could easily be merged if there is no conflict.

In [13] it was assumed that the *T-Boxes* of the agents involved are identical. Now we increase the flexibility in that we divide the *T-Box* into a stable part (expression 3, stable in both agents, see subsection 2.3) and a dynamic part, that is changed in both KBs in parallel.

In *OWL-DL* the Open World Assumption (OWA) holds. Thus, it holds in our setting for both agents. This situation gives room for independent expansions of the *T-Boxes* of the communicating agents. This is further supported by the RDF base of *OWL-DL*, in which all names are URIs, so that collision free new names can be introduced without central coordination. But even if we assume that the KBs of both agents have to be internally consistent all the time and that all their concepts have to be satisfiable. So, concepts/roles are changed/expanded in a monotonic way. The same applies to *T-Boxes* in which an existing expression has been replaced with a different one in at least one of the *T-Boxes*. In Figure 5 an example of an independently changed *T-Box* is given. Here,

$$\mathsf{Animal} \sqsubseteq \neg\exists \mathsf{use}.\mathsf{Photosynthesis} \sqcap \neg(\mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{13}$$
$$\mathsf{Plant} \sqsubseteq \neg\mathsf{Fungus} \tag{14}$$
$$\mathsf{Animal} \sqsubseteq \mathsf{LivingMatter} \tag{15}$$
$$\mathsf{Animal} \sqsubseteq \forall \mathsf{eat}.\mathsf{Fungus} \tag{16}$$
$$\mathsf{Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{17}$$
$$\mathsf{Animal} \sqsubseteq \exists \mathsf{need}.\mathsf{Oxygen} \tag{18}$$

Fig. 5. A *T-Box* conflicting with figure 4

expression 16 conflicts with expression 10 in Figure 4. This example will be used to demonstrate the process of merging.

To explain the merge algorithm, the example stated in 2.2 is described in detail: Both agents are aware of the introduced concepts Animal, Plants and Fungus being disjoint. Additionally they know about the role eat. We ignore here the expressions about the need for Oxygen, the Photosynthesis and the relation to LivingMatter (subconcept) to simplify matters. For this example we reduce the two KBs to the ones shown in figure 6 and 7.

$$\mathsf{Animal} \sqsubseteq \neg(\mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{19}$$
$$\mathsf{Plant} \sqsubseteq \neg\mathsf{Fungus} \tag{20}$$
$$\mathsf{Animal} \sqsubseteq \forall \mathsf{eat}.\mathsf{Fungus} \tag{21}$$
$$\mathsf{Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{22}$$

Fig. 6. Senders *T-Box*

$$\mathsf{Animal} \sqsubseteq \neg(\mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{23}$$
$$\mathsf{Plant} \sqsubseteq \neg\mathsf{Fungus} \tag{24}$$
$$\mathsf{Animal} \sqsubseteq \forall \mathsf{eat}.(\mathsf{Animal} \sqcup \mathsf{Plant}) \tag{25}$$
$$\mathsf{Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{26}$$

Fig. 7. Receivers *T-Box*

2.3 Stable part of the *T-Boxes*: Common Grounding

The already mentioned expression 3 in example 5 represents a stable, invariant part of all *T-Boxes* of all agents (sender and receiver). We divide a *T-Box* in three different parts:

1. A stable (not changing) part of the *T-Box*
2. A dynamic part, where concepts (included in agents KB) change over time
3. A completely free part, where concepts are defined, that are not shared with other agents.

A stable part (1) can be used for concepts and roles that are defined at an imported *OWL-DL* ontology. To ensure stability, we have to restrict: (1) New concepts/roles (dynamic part) are not equivalent to imported, invariant ones. A concept of the dynamic part is not a superconcept of an imported one. Since equality between concepts ($\equiv$) can be seen as a pair of subconcept relations, like $\mathsf{A} \sqsubseteq \mathsf{B}$ and $\mathsf{B} \sqsubseteq \mathsf{A}$, this restriction disallows equality relations. (2) It is not allowed to change expressions that define these imported concepts/roles.

So, agents may use these concepts and roles as a "Common Ground" (shared knowledge). Therefore, the concepts defined as Common Ground should be limited to a set of abstract, general definitions (reference/fundamental ontologies). In addition to this invariant part of the *T-Box* there also exist shared, but changeable concepts and roles (part 2). The merge process works on this part of the *T-Box*. Finally there might by a third part of the *T-Box* (part 3). The definitions here are "private"and are not shared to others. For this paper we simplify this part: we leave this last part empty.

3 Merging process

We follow [8] and [4] when we define the 'merging of ontologies' as the creation of an unified ontology from two already existing ontologies with an overlapping part. As already stated above, agents exchange knowledge by sending *A-Box* expressions (knowledge about individuals). If the corresponding *T-Box* expressions of these individuals are known, the receiving agent tries to integrate these individuals into its *A-Box*. If there are *A-Box* consistency conflicts, an '*A-Box*-updater' should reconstruct consistency by e.g. contracting *A-Box* expressions. In [14] an algorithm for revising the receivers *A-Box* with the content of incoming *inform* messages is introduced. Since this algorithm only affects the *A-Box* of the receiver and does not contract any knowledge from the *T-Box*,

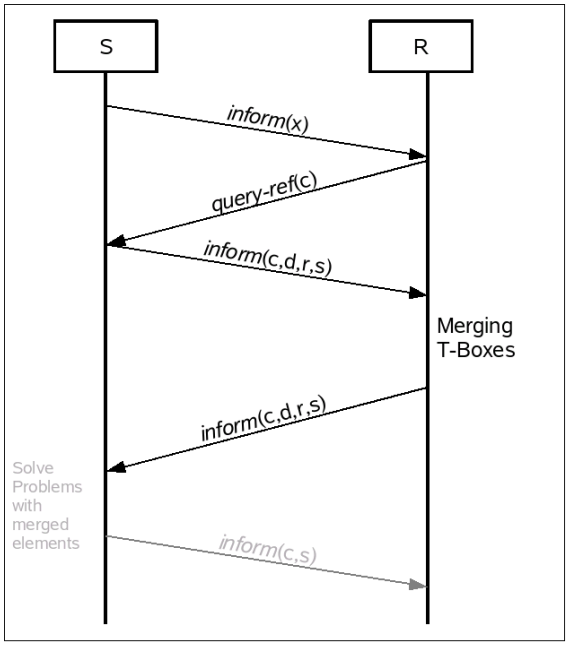


Fig. 8. Communication between sender and receiver agent about needed definitions

it would be unable find a solution if the message content is inconsistent with the *T-Box* of the receiver. In this case the algorithm described here can be applied to find a shared understanding of the names mentioned in the *inform* message.

In figure 3 the receiver tries to insert an *A-Box* expression (individual; here x), but due to additional needed *T-Box* expressions, it demands the sender using a *query-ref* (here c is a concept). The answer (*inform*) contains the *T-Box* expressions c, d, s and r (c, d concepts; s, r roles). In general the answer consists of all *T-Box* expressions needed by the concept/role definition (co/ro for short). Now, the receiver merges these *T-Box* expressions with its *T-Box*. Afterwards, feedback consisting of merged concept and role expressions (c, d, r, s) is sent to the sender using an *inform*. Another communication act is needed if the feedback itself interferes with other parts of the *T-Box* or *A-Box* of the sender.

In our example (Figure 6) the sender informs the receiver about an individual from its KB, which is an Animal **only** eating Fungus. This individual can not be integrated by the receiver, since its KB only knows about Animals, that eat Plants or Animals. The receivers *T-Box* additionally restricts that Fungus is a disjoint concept from Plants and Animals.

Now, we throw a glance at the *T-Box* expressions the original sender must add to its *inform* (with the co/ro):

- equivalent concepts/roles
- superconcepts/-roles
- disjoint concepts/roles

– roles and concepts used in the definition of the co/ro

The last line of this list results in a recursion that only stops at $\top$, $\bot$ or at datatype values. The stable Common Ground described in subsection 2.3 can be used as an additional termination condition in order to reduce runtime complexity.

3.1 Concept and role names

A first step to make the incoming *inform* message content processable is to separate the received part of the sender *T-Box* from the receiver *T-Box* by temporarily introducing two new names for each of the concepts and roles that occur in both *T-Boxes*. This can be done without risking name conflicts by using the hierarchical structure of URIs together with the unique names of the different agents as prefixes[4]. Additionally the receiver agent adds version numbers (timestamps) to these names (and the names of the sender respectively) to distinguish them from already generated ones (by repeated communication and merging). The mapping between the old names and the pair of new names is stored and then applied to the *T-Box* and *A-Box* of the receiver (with the receiver prefix) as well as to the known part of the senders *T-Box* and the message content (with the sender prefix) in a search and replace operation. After these new names are introduced the two *T-Boxes* can be safely joined into one (receiver side). So far the pairs of new names in the receiver *T-Box* are not related at all. A relation between the new names is established by adding the old name of such a pair as a shared superconcept (or superproperty) of both of them. In the following a set of two new names related to the matching original name will be referred to as a concept/role triangle. Another possibilities relating old and new names are described in subsection 5.1.

In our example this results in the *T-Box* shown in Figure 9. Expression 35 inserts the old name as superconcept of senders animal (S_Animal) and expression 42 the superrole to the receivers role eat (r_eat). The expressions between 35 and 42 do this for all concepts that appeared in both initial *T-Boxes*.

3.2 Recovering shared concept definitions

After the introduction of these new subconcepts any new knowledge an agent might have previously acquired about a common concept is effectively lost. This is caused by the fact that only the concept membership assertions currently found in the *A-Box* can go through the replacement process, while subsequent messages will only reference either the old name or the new name of the communication partner. None of these names would allow that agent to apply his previously acquired knowledge (initially about the original concept) to individuals found in those messages.

In order to minimize this loss our algorithm now searches for unnecessary concept (and role) triangles. These are identified by additionally asserting that both new names are equivalent to the old name and then checking KB consistency and concept satisfiability.

[4] A name of a concept/role encoded in *OWL-DL* is an URI consisting of scheme, host, path and fragment. The name of the agent which owns the KB is used in the host name of this URI.

$$\mathsf{S\_Animal} \sqsubseteq \neg(\mathsf{S\_Plant} \sqcup \mathsf{S\_Fungus}) \quad (27)$$
$$\mathsf{S\_Plant} \sqsubseteq \neg\mathsf{S\_Fungus} \quad (28)$$
$$\mathsf{S\_Animal} \sqsubseteq \forall \mathsf{s\_eat}.\mathsf{S\_Fungus} \quad (29)$$
$$\mathsf{S\_Animal} \sqsubseteq \geqq 1\ \mathsf{s\_eat} \quad (30)$$
$$\mathsf{R\_Animal} \sqsubseteq \neg(\mathsf{R\_Plant} \sqcup \mathsf{R\_Fungus}) \quad (31)$$
$$\mathsf{R\_Plant} \sqsubseteq \neg\mathsf{R\_Fungus} \quad (32)$$
$$\mathsf{R\_Animal} \sqsubseteq \forall \mathsf{r\_eat}.(\mathsf{R\_Animal} \sqcup \mathsf{R\_Plant}) \quad (33)$$
$$\mathsf{R\_Animal} \sqsubseteq \geqq 1\ \mathsf{r\_eat} \quad (34)$$
$$\mathsf{S\_Animal} \sqsubseteq \mathsf{Animal} \quad (35)$$
$$\mathsf{R\_Animal} \sqsubseteq \mathsf{Animal} \quad (36)$$
$$\mathsf{S\_Fungus} \sqsubseteq \mathsf{Fungus} \quad (37)$$
$$\mathsf{R\_Fungus} \sqsubseteq \mathsf{Fungus} \quad (38)$$
$$\mathsf{S\_Plant} \sqsubseteq \mathsf{Plant} \quad (39)$$
$$\mathsf{R\_Plant} \sqsubseteq \mathsf{Plant} \quad (40)$$
$$\mathsf{s\_eat} \sqsubseteq \mathsf{eat} \quad (41)$$
$$\mathsf{r\_eat} \sqsubseteq \mathsf{eat} \quad (42)$$

Fig. 9. Name-expanded *T-Box*

These checks are done by querying an *OWL-DL* reasoner. In preparation for these queries the temporary *T-Box* and *A-Box* are both translated into DIG 1.1 syntax and then told into a new reasoner instance. The queries run on this new reasoner instance are:

consistency The first step is a check for consistency. DIG 1.1 does not define a query type for consistency checks. In RACER 1.9, however, consistency of a KB can be determined with the (proprietary) query `<asks><consistentKB/></asks>`.

satisfiability As already mentioned in 2.2, besides the consistency of the KB, each of the defined concepts must be satisfiable. This is checked with queries of the following pattern:
`<asks><satisfiable>[Concept]<satisfiable/></asks>`
(DIG 1.1, understood by RACER).

Based on these checks, our algorithm determines which new concepts and roles can be merged back into their original names and which have to remain separate. Those new names that are allowed to be equivalent with their original names are then replaced with the original name, based on the stored mapping between the original URIs and the newliy introduced ones. Obviously it is not possible to merge all new concepts/roles back into their original counterpart, because in that case the initial inconsistency that triggered the execution of this algorithm would not have occurred.

Since conflicts can also be the result of one concept being made more specific in one *T-Box* and a different, but related concept being made more specific in the other *T-Box*, interdependencies between different pairs of new names can occur that lead to

one pair only being consistently mergable if certain other pairs are not merged. That is why different combinations of merged and separated concepts/roles have to be examined. Because ideally only few of the new names would remain, we ignore those combinations of merged concepts/roles that are contained in a different combination. Finding maximal subsets is a well studied step of belief revision algorithms e.g. [10, page 170ff].

Searching for these combination is computationally expensive since it requires many consistency and satisfiability checks. Effort is reduced by two rather straight forward optimizations. One is to reuse reasoner instances if after a successful test with a given set of equality assertions a test with additional equality assertions is needed. In this case the existing reasoner instance can be extended with the additional assertions and the next consistency/satisfiability check can be run without reloading the whole KB.

The other one is to skip combinations that are contained in a previously found solution, which can be explained with the following little example: Let us consider the case that one possible solution is to merge concepts A, B and C while D and E will be kept separate. Now if the next combination to be checked is to merge A and B while C, D and E are kept separate, this second combination will be consistent as well and therefore does not need to be checked.

After the maximal subsets of mergable name pair combinations have been found, a choice has to be made. Here we are following the naming introduced in the AGM theory (the Partial Meet hierarchy, ranging from Full Meet to Maxichoice, as described in [1, 7]), even though the elements of the sets between which the choice is made are not propositional sentences but mergable concepts and roles. This is not the only difference, because even the worst case (Full Meet) would not result in a total loss of knowledge, but merely in the creation of more subconcepts and roles than absolutely necessary.

Since subsequent computationally expensive steps are applied to the remaining unmerged name pairs later in our algorithm (see 3.3), it is favorable to choose a single combination of mergable triples. This relates our algorithm to Maxichoice contraction. Maxichoice contraction is often criticized for keeping too much knowledge that might be wrong, but in our case the *A-Box* content of the receiver KB and the incoming *inform* message is taken as additional input during the search, which implicitly rules out combinations that do not match actual observations stored in the *A-Boxes*.

In our example, combinations of the four pairs of possible equivalence expressions shown in Figure 10 are examined, which, all taken together, would result in an inconsistent KB. If the result of adding a combination of these expressions to the KB from Figure 9 is consistent, the combination represents a potential possible solution. Each of these valid combinations, that is not a subset of another potential possible solution, is considered a maximal possible solution.

Our algorithm applied to this example will generate two maximal possible solutions: one in which the equivalences from {44, 45, 46} are true and one in which {43, 44, 45} are true. There are other solutions, like {43, 44}, but these are not maximal and are therefore not considered. In the following the first solution will be further examined.

$$\mathsf{S\_Animal} \equiv \mathsf{Animal} \equiv \mathsf{R\_Animal} \tag{43}$$
$$\mathsf{S\_Plant} \equiv \mathsf{Plant} \equiv \mathsf{R\_Plant} \tag{44}$$
$$\mathsf{S\_Fungus} \equiv \mathsf{Fungus} \equiv \mathsf{R\_Fungus} \tag{45}$$
$$\mathsf{s\_eat} \equiv \mathsf{eat} \equiv \mathsf{r\_eat} \tag{46}$$

Fig. 10. Combination options

3.3 Finding common parts in the remaining unmerged concepts

Once a combination has been chosen, it has to be applied. Instead of adding the found equivalency expressions to the *T-Box*, the original names are put back into all *T-Box* and *A-Box* expressions where they had previously been replaced with one of the new names. This makes the subconcept and subproperty relations that were introduced between the old names and the now removed new names redundant, so they are removed from the *T-Box*. Since we chose solution 43, Plant, Fungus and eat will be merged. This results

$$\mathsf{Plant} \sqsubseteq \neg\mathsf{Fungus} \tag{47}$$
$$\mathsf{S\_Animal} \sqsubseteq \neg(\mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{48}$$
$$\mathsf{R\_Animal} \sqsubseteq \neg(\mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{49}$$
$$\mathsf{S\_Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{50}$$
$$\mathsf{S\_Animal} \sqsubseteq \forall \mathsf{eat}.\mathsf{Fungus} \tag{51}$$
$$\mathsf{R\_Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{52}$$
$$\mathsf{R\_Animal} \sqsubseteq \forall \mathsf{eat}.(\mathsf{R\_Animal} \sqcup \mathsf{Plant}) \tag{53}$$
$$\mathsf{S\_Animal} \sqsubseteq \mathsf{Animal} \tag{54}$$
$$\mathsf{R\_Animal} \sqsubseteq \mathsf{Animal} \tag{55}$$

Fig. 11. Partially merged *T-Box*

in the *T-Box* shown in figure 11.

After the pairs of new names that can be completely merged have been removed, the remaining concept and role triangles are further examined in order to find expressions specifying one of the two newly introduced subconcepts (or subroles) that can as well be applied to the superconcept (or superrole) without causing an inconsistency. This is done in a similar way as the search for equality combinations.

Here, the expressions describing the subconcepts/subroles of a triangle are focused. The expressions of a subset are temporarily rewritten using the original name (the name of the superconcept/superrole of the triangle) instead of the new name. If the KB ist consistent, the subset is a possible solution. The resulting set of expression combinations is then used as the input for a search for maximal subsets following the same steps

as the search described in subsection 3.2. This results in a set of expressions that can be moved to the superconcept/superrole without making the KB inconsistent.

This does not result in new knowledge, because the moved expressions were either part of one (or both) of the original *T-Boxes*, or are at least closer to the originals than the new versions. Instead, the found expressions are those of the expressions related to original names of concepts or roles, that do not have to be 'pushed down' to the newly created subconcepts (or subroles). In order to avoid unnecessary redundant *T-Box* expressions, the now redundant versions (those relating to the old names) of the expressions from the result set are then removed from the *T-Box*. If, on one side of the triangle, all examined expressions can be rewritten using the original name, the new name can be completely dropped, including the generated 'triangle leg' $\mathsf{generatedSubUri} \sqsubseteq \mathsf{originalUri}$.

Back to our example we now examine the remaining separated (not merged) concept Animal. The expressions that have to be considered for the concept triangle consisting of the superconcept Animal and the two generated subconcepts S_Animal and R_Animal are the expressions 48, 49, 50, 51, 52, and 53 from 11. These have to be searched for possible combinations, which results in only one maximal possibility. This is keeping 51 and 53 in their old form (the expressions that are referencing the subconcept) and replacing 48, 49, 50 and 52 with the corresponding expressions using the original name Animal. This can result in pairs of identical expressions (e.g. the original name versions of 48 and 49) which are of course only inserted once into the *T-Box*.

$$\mathsf{Plant} \sqsubseteq \neg\mathsf{Fungus} \tag{56}$$
$$\mathsf{Animal} \sqsubseteq \neg(\mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{57}$$
$$\mathsf{Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{58}$$
$$\mathsf{S\_Animal} \sqsubseteq \forall \mathsf{eat}.\mathsf{Fungus} \tag{59}$$
$$\mathsf{R\_Animal} \sqsubseteq \forall \mathsf{eat}.(\mathsf{R\_Animal} \sqcup \mathsf{Plant}) \tag{60}$$
$$\mathsf{S\_Animal} \sqsubseteq \mathsf{Animal} \tag{61}$$
$$\mathsf{R\_Animal} \sqsubseteq \mathsf{Animal} \tag{62}$$

Fig. 12. Final merged *T-Box*

As already mentioned before, it is possible that some of the introduced new names can be dropped beause all of the expessions on their side of their triangle can be rewritten using the original name. In our example both subconcepts, S_Animal and R_Animal, are each used in one remaining expression, so both new subconcepts have to remain.

3.4 Informing the communication partner

The final step is to inform the original sender about the changes made. The sender then has to do the appropriate replacements in his own KB. This can result in further inconsistencies, since the receiver was not able to verify the changes made so far against

the *A-Box* of the sender. In this case, the sender has to run the algorithm a second time, so that consistency with his whole KB is checked, not just with the parts he sent to the receiver. Since this second run of our algorithm will not make any existing concept definitions more specific, the result is a pair of *T-Boxes* that are open enough to be consistent with the *A-Boxes* of both agents. Lastly, the resulting final changes have to be sent to the receiver of the original message.

4 The algorithm

initialize $uriMapping$ {the mapping from each original URI to the pair of generated URIs for sender and receiver}
initialize new temporary KB $tempTBox_r$ with new names
initialize new temporary KB $tempTBox_s$ with new names
$tempTBox$:= $tempTBox_r \cup tempTBox_s$
build triangles in $tempTBox$
$tempABox$:= copy of receiver ABox with with new names
$tempMessage$:= copy of message content ABox with new names
initialize new $trianglesets$
for all $trianglesToKeep$ in all subsets of $uriMapping$ **do**
  for all concept or role in $uriMapping$ not in $trianglesToKeep$ **do**
    add $C_s \equiv C \equiv C_r$ or $R_s \equiv R \equiv R_r$ respectively
  end for
  check consistency/satisfiability of $tempTBox$, $tempABox$, $tempMessage$
  if consistency/satisfiability check successful **then**
    add $trianglesToKeep$ to $trianglesets$
  end if
end for
remove non minimal subsets from $trianglesets$
$solution$:= a set selected in the choice operation from $trianglesets$ {implemented: random maxichoice of smallest sets of $trianglesets$}
for all mapping uri_s, uri_r, $uri_{original}$ from $uriMapping$ not in $solution$ **do**
  replace $uri_r \rightarrow uri_{original}$ in $tempTBox$, $tempABox$
  replace $uri_s \rightarrow uri_{original}$ in $tempTBox$, $tempMessage$
  remove resulting redundant $uri_{original} \sqsubseteq uri_{original}$
end for
for all mapping uri_s, uri_r, $uri_{original}$ in $solution$ **do**
  {search additional parts of knowledge about the URI that both agents can agree on, not yet a complete search}
  $expressionSet$:= any expression from $tempTBox$ defining uri_s or uri_r
  remove the generated sentences $uri_s \sqsubseteq uri_{original}$ and $uri_r \sqsubseteq uri_{original}$ from $expressionSet$
  initialize $subExpressionsets$ as a set of subsets of $expressionSet$
  for all $expressionsForNewUri$ all subsets of $expressionSet$ **do**
    for all $expression$ in $expressionSet$ not in $expressionsForNewUri$ **do**
      rewrite $expression$ with uri_s and uri_r replaced with $uri_{original}$
    end for
    check consistency/satisfiability of $tempTBox$, $tempABox$, $tempMessage$
    if consistency/satisfiability check successful **then**
      add $expressionsForNewUri$ to $subExpressionsets$

37: **end if**
38: **end for**
39: remove non minimal subsets from $subExpressionsets$
40: $expressionSolution$:= a set selected in the choice operation from $subExpressionsets$ {implemented: random maxichoice again}
41: **for all** $expression2$ from $expressionSet$ not in $expressionSolution$ **do**
42: rewrite $expression2$ with uri_s and uri_r replaced with $uri_{original}$
43: **end for**
44: **if** $expressionSolution$ contains no expression about uri_r **then**
45: replace uri_r with $uri_{original}$ in $tempTBox$ and $tempABox$
46: check consistency/satisfiability of $tempTBox$, $tempABox$, $tempMessage$
47: remove resulting redundant $uri_{original} \sqsubseteq uri_{original}$
48: **if** consistency/satisfiability check successful **then**
49: undo replace uri_r with $uri_{original}$
50: **end if**
51: **end if**
52: **if** $expressionSolution$ contains no expression about uri_s **then**
53: replace uri_s with $uri_{original}$ in $tempTBox$ and $tempMessage$
54: check consistency/satisfiability of $tempTBox$, $tempABox$, $tempMessage$
55: remove resulting redundant $uri_{original} \sqsubseteq uri_{original}$
56: **if** consistency/satisfiability check successful **then**
57: undo replace uri_s with $uri_{original}$
58: **end if**
59: **end if**
60: **end for**

5 Discussion and conclusion

The algorithm introduced in this paper has one major advantage. It will always return a *T-Box* that is consistent. The *T-Box* containing renamed concepts and roles, is always compatible with the merged *A-Box* and does not lack any previously available implicit knowledge after all occurrences of the original in the *A-Box* have been updated. Both agents can still use their own definitions by using the subconcepts and and subroles created for them. Likewise they are not only able to understand the common part of the concept and role definitions if their communication parter subsequently sends messages referencing the partner's subconcepts and subroles, but can also benefit from additional implications of those names since their definitions are known as well. Furthermore, compatible extensions made to the TBox of one agent can shared without constructing new subconcept and subroles,

5.1 An alternative triangle

Unifying the non-conflicting expressions (shown in section 2.2) is a correct merge solution. We restrict the change possibilities in each agent to monotonic ones. The superconcept is the result of more general versions of the initial concepts. It contains the common expressions as well as the nonconflicting ones. This will change the original

concept, but it will always include all individuals matching one of the two conflicting definitions.

In our example, every animal from both conflicting *T-Boxes* is still an animal afterwards (due to the fact that S_animal and R_animal are subconcepts of it).

An alternative way to build a triangle is to add the union of the newly built subconcepts to the superconcept:

$$\mathsf{SuperConcept} \equiv \mathsf{SubConcept1} \sqcup \mathsf{SubConcept2} \tag{63}$$

Additionally the subconcepts are defined as intersection between the conflicting expressions and the superconcept:

$$\mathsf{SubConcept} \equiv \mathsf{SuperConcept1} \sqcap \mathsf{ConflExp1} \sqcap \mathsf{ConfExp2...} \tag{64}$$

The alternatives to build the triangle differs in that the superconcept is partitioned into only this two subconcepts. So, individuals of the superconcept have to be in one of the subconcepts. In contrast to this, the triangle built in subsection 3.1 allows other subconcepts, e.g. $\mathsf{T\_animal} \sqsubseteq \mathsf{Animal} \sqcap \neg(\mathsf{R\_animal} \sqcup \mathsf{S\_animal})$, but expressions (63) and (64) restrict the subconcept to S_animal and R_animal. As one easily can see, the correctness of the solution is preserved, since none of the original *T-Boxes* would have allowed an instance of animal not matching the definitions that now apply to either R_animal or S_animal. Figure 13 shows this alternative applied to the example from the previous sections. One argument to add expressions like (63) and (64) is that it

$$\mathsf{Plant} \sqsubseteq \neg\mathsf{Fungus} \tag{65}$$
$$\mathsf{Animal} \sqsubseteq \neg(\mathsf{Plant} \sqcup \mathsf{Fungus}) \tag{66}$$
$$\mathsf{Animal} \sqsubseteq \geqq 1\ \mathsf{eat} \tag{67}$$
$$\mathsf{S\_animal} \equiv (\mathsf{Animal} \sqcap \forall \mathsf{eat}.\mathsf{Fungus}) \tag{68}$$
$$\mathsf{R\_animal} \equiv (\mathsf{Animal} \sqcap \forall \mathsf{eat}.(\mathsf{R\_animal} \sqcup \mathsf{Plant})) \tag{69}$$
$$\mathsf{Animal} \equiv \mathsf{S\_Animal} \sqcup \mathsf{R\_Animal} \tag{70}$$

Fig. 13. Integration alternative

contains the semantic of both original *T-Boxes* in terms of '**All** animals I know are like this'(a sentence both agents can say).

A disadvantage of this alternative appears if more than one merge situation occur. If so, the union in the superconcept triggers the creation of 'middle' concepts. Using the original triangle alternative, a second merging process with another more specific concept will generate only **one** new subconcept, since the superconcept is generalized allowing other subclasses. Using the alternative, the union is the conflicting expression and needs to be moved to a newly built subconcept. The other newly built subconcept of the triangle is the one generated for the incoming concept. The superconcept is the

union of these concepts. **Two** new concepts were added. So, in some situations the alternative builds more concepts.

Another disadvantage of adding expressions like (63) and (64) is that they were not contained in the original *T-Boxes*. Once merged, the concept with the original URI will always contain a union. So, the common name can not be generalized to its original state, since the added expression will not be removed by our algorithm. Therefore, recovery as defined in [7] is not possible.

Also, this alternative way of defining triangles can not yet be applied to roles, since in OWL DL 1.0 the union constructor can only be used with concepts.

To make this more obvious, we will compare the alternatives in an example. We consider three Agents A,B and C having disjunct concepts identified by the same URI.

$$original : \mathsf{X} \sqsubseteq \top \tag{71}$$
$$AgentA : \mathsf{X} \sqsubseteq (= 1\ \mathsf{R}) \tag{72}$$
$$AgentB : \mathsf{X} \sqsubseteq (= 2\ \mathsf{R}) \tag{73}$$
$$AgentC : \mathsf{X} \sqsubseteq (= 3\ \mathsf{R}) \tag{74}$$

Fig. 14. Agents view on X

Now agent A and B communicate and merge X. This results in figure 15. X is not

$$\mathsf{X_A} \equiv \mathsf{X} \sqcap (= 1\ \mathsf{R}) \tag{75}$$
$$\mathsf{X_B} \equiv \mathsf{X} \sqcap (= 2\ \mathsf{R}) \tag{76}$$
$$\mathsf{X} \equiv \mathsf{X_A} \sqcup \mathsf{X_B} \tag{77}$$

Fig. 15. Agents A and B merge

the original concept, because of (77). Next step is agent C communicating with agent A. X can not be merged since the union permitts a cardinality of 3. The resulting *T-Box* is shown in figure 16.

The difference to the other way to build the triangles is, that concepts the $\mathsf{X_{AB}}$ is not generated. The concept identified by the original URI, here X, is genralized to its original state. The results can be seen in figure 17

5.2 Further work

Section2.3 introduced the idea to divide the *T-Box* into multiple parts in which different levels of change are allowed to happen. The comments of the reviewers of this paper lead to an ongoing work/discussion in our group on this topic.

$$X_A \equiv X_{AB} \sqcap (= 1\ R) \tag{78}$$
$$X_B \equiv X_{AB} \sqcap (= 2\ R) \tag{79}$$
$$X_{AB} \equiv X \sqcap (X_A \sqcup X_B) \tag{80}$$
$$X_C \equiv X \sqcap (= 3\ R) \tag{81}$$
$$X \equiv X_{AB} \sqcup X_C \tag{82}$$

Fig. 16. Agents A and C merge

$$X_A \sqsubseteq (= 1\ R) \tag{83}$$
$$X_A \sqsubseteq X \tag{84}$$
$$X_B \sqsubseteq (= 2\ R) \tag{85}$$
$$X_B \sqsubseteq X \tag{86}$$
$$X_C \sqsubseteq (= 3\ R) \tag{87}$$
$$X_C \sqsubseteq X \tag{88}$$

Fig. 17. Using other triangle alternative

5.3 Conclusion

The paper presented an algorithm that merges *T-Boxes* in agent-agent-communication. Even though we have to evaluate this approach with different possible merge problems, we think that the handling of the dynamicalised *T-Boxes* adds important flexibility to agents using this algorithm.

5.4 Acknowledgments

The authors want to thank Prof. Görz for reviews and usefull discussions on previous versions of this paper.

Bibliography

[1] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *J. Symb. Log.*, 50(2):510–530, 1985.

[2] F. Baader and W. Nutt. Basic description logics, 2003.

[3] Sean Bechhofer. *The DIG Description Logic Interface: DIG/1.1*. University of Manchester, Oxford Road, Manchester, M13 9PL, 2003.

[4] Marc Ehrig. *Ontology Alignement*. Springer Science+Business Media LLC, 223 Spring Street, New Yoork, NY 10013, 2007.

[5] Foundation for Intelligent Physical Agents FIPA. FIPA Communicative Act Library Specification, 2002.

[6] Foundation for Intelligent Physical Agents FIPA. FIPA SL Content Language Specification, 2002.

[7] Peter Gardenfors, editor. *Belief Revision*. Cambridge University Press, New York, NY, USA, 1992.

[8] Michel Klein. Combining and relating ontologies: an analysis of problems and solutions. In Asuncion Gomez-Perez, Michael Gruninger, Heiner Stuckenschmidt, and Michael Uschold, editors, *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle, USA, august 2001.

[9] Eric Miller et al. Web Ontology Language (OWL), 2004.

[10] Bernhard Nebel. *Reasoning and Revision in Hybrid Representation Systems*, volume 422 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[11] RACER Systems GmbH. The features of racerpro version 1.9, 2005.

[12] S. Russel and P. Norvig. *Artificial Intelligence*, chapter 7. Logic Based Agents. Prentice Hall, 2003.

[13] Bernhard Schiemann and Ulf Schreiber. OWL DL as a FIPA ACL content language. In Nicola Guarino Roberta Ferrario and Laurent Prevot, editors, *Proceedings of the Workshop on Formal Ontology for Communicating Agents (FOCA)*, 18th European Summer School of Language, Logic and Information, pages 73–80, University of Malaga, July 2006.

[14] Ulf Schreiber. ABox–Updates von OWL DL Wissensbasen in JADE basierten Agenten. Diplomarbeit, University of Erlangen-Nuremberg, Haberstr. 2 91058 Erlangen, May 2007.

[15] David R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, Rochester, 1994.