Information acquisition from multi-agent resources

Zhisheng Huang* and Peter van Emde Boas**

*Center for Computer Science in Organization and Management(CCSOM)
University of Amsterdam
**ILLC, Department of Mathematics and Computer Science
University of Amsterdam

Abstract

Rational agents, information systems and knowledge bases all share the property that they may become more effective by combining information from multiple sources. However, as was clearly indicated by the notorious "Judge puzzle" proposed by W. J. Schoenmakers in 1986, combining information from several sources is a dangerous operation. The resulting database may turn out to be inconsistent, or even worse: there are situations where the result is consistent but supports inferences which contradict the beliefs of all contributing agents.

In this paper we investigate the possibilities and limitations of strategies for coping with this problem. Our first attempt tries to characterize those situations where information can be combined without risking the undesirable situation that some derivable proposition contradicts the beliefs of all agents involved. The resulting notion is called *Absolute safety*. It turns out however that for that case only trivial solutions exist. Consequently any non-trivial strategy must use information about the epistemic states of the agents involved.

Subsequently we investigate less restrictive notions of safety. The more interesting ones involve not only propositions about the world but also epistemic information about the knowledge of the agents. This information can be formulated conveniently using the logic of belief dependence which has been designed by the first author, and which has been used previously for designing effectively computable belief revision procedures.

The results characterizing the alternative safety notions generalize for this extended logic. We present a notion of *restricted almost safety* within this framework which describes the safety of combining information under the hypothesis that the contributing agents eventually would have exchanged their information among themselves. For this notion an explicit solution to the Judge puzzle is given.

1 Introduction

The construction of models for multi-agent epistemic systems has become one of the most interesting and popular topics in artificial intelligence and in the theory of knowledge based expert systems. Information systems in the real world are loaded by combining information from many (possibly unrelated) sources. As is generally known merging information may produce inconsistent knowledge bases. However, an even more subtle risk was indicated by W. J. Schoenmakers [0] when he published his *Judge puzzle*. This puzzle describes the situation where an agent, called the *judge*, by combining information from two other agents, called the *witnesses*, consistently obtains a conclusion which contradicts the belief sets of both witnesses:

> *Once upon a time a wise but strictly formal judge questioned two witnesses. They spoke to her on separate occasions. Witness w1 honestly stated his conviction that proposition p was true. Witness w2 honestly stated that he believed that the implication $p \rightarrow q$ was true. Nothing*

else was said or heard. The judge, not noticing any inconsistency accepted both statements and concluded that q had to be true. However, when the two witnesses heard about her conclusion they were shocked because they both were convinced that q was false. But they were too late to prevent the verdict to be executed...

As pointed out by Schoenmakers, in the above story nobody can be blamed for this situation to arise. The witnesses, even though formally required to tell everything they know, are not responsible since neither of them was asked about q and hardly could know at the time of interrogation that the truth of q was at stake. The judge on the other hand had no reason to even consider the possibility that her argument was unsound, since there is not the slightest trace of contradiction in the testimony. She might have asked on, and confronted the witnesses with her conclusion that q was true. For the judge this would have been possible, but, as Schoenmakers indicates, this possibility is lost in the case of a knowledge base being loaded with information from independent sources, since by the time proposition q turns out to be relevant the two informants no longer are accessible. And therefore Schoenmakers concludes:

Intelligent database systems may behave perfectly in splendid isolation, operating on one world without inconsistencies, but even when they are consistent they may produce unacceptable results when operating on the information that is accessible in a community of such systems. Their results will be acceptable, most of the time, but nobody knows when.

Consequently it becomes relevant to look for a characterization of situations where combining information from multi sources is *safe*, which informally means that no conclusion drawn from the combined information is disbelieved by all informants. At the same time our combination operator should support at least the derivation of one proposition not already supported by one of the contributing agents; otherwise the problem of obtaining the right information reduces to the identification of the right source.

However, having formalized this problem, we prove a triviality theorem expressing that a combining operation satisfying the above form of *absolute safety* doesn't exist. Consequently, a more refined approach is required which takes into account both the information contributed by the agents and their complete belief sets. In this context the notions of *safety* and *strong safety* are defined, and some characterizations are obtained. It follows that dangerous situations only arise when every agent disagrees with some other agent about some of the propositions which are actually communicated.

These results once more indicate that in a multi agent environment one should maintain a strict distinction between information accepted on behalf of an other agent, and information which is incorporated in your own belief set. The resulting process of accepting information followed by incorporating it, is one of the main motivations for the introduction of the *logic of belief dependence* [0] by the first author. This logic has been previously applied for effectively choosing between several belief revision strategies [0], and for an earlier solution to the Judge puzzle [0].

Compared to our previous paper [0] we believe that we presently can make a much stronger case for the "contrived" solution to the judge puzzle presented in the final section of that paper. The triviality theorem shows that there is no simple solution for the problem. The characterization of the less restrictive safety notions shows that danger is caused by disagreement between agents and disagreement between agents is a fact of life we can't get around. The case for a two stage process for belief incorporation has been argued elsewhere [0, 0]; it is also supported by psychological research. However, when generalizing the safety notions to the case of our epistemic logic of belief dependence, the characterizations for the propositional case extend, and so do their negative consequences. Therefore, the best we can hope for is a specific belief incorporation strategy for the judge which is approximatively safe.

The proposed notion of *restricted almost safety* characterizes the situation where the conclusion of the judge will not contradicted by all witnesses, provided they will eventually have access to each other's

information. This hypothetical situation can be expressed in terms of sub-beliefs in our logic of belief dependence, leading to an effectively testable condition for deciding whether a specific belief revision operator for the judge is almost safe or not.

2 Combining information from multiple agents; the triviality result

In the sequel I denotes a finite and non-empty set of the agents called *informants* and a the *receiver*, an agent who receives and combines information from the informants I. In this section, we study the case of propositional logic $\mathbf{L_P}$, where information communicated between agents consists of pure propositional formulas without modal operators. The language $\mathbf{L_P}$ is recursively constructed from a primitive proposition set Φ_0 and the Boolean connectives as usual. Moreover, the logical notions of a semantic model, the satisfiability relations $\models$, and the consequence operation Cn, are defined as usual.

The receiver's obtained information, is a mapping ψ from the informants I into the formula set $\mathbf{L_P}$. We use the notation $\{\psi_i\}_{i\in I}$ to denote the set $\{\psi(i) \in \mathbf{L_P} : i \in I\}$. The set $\{\psi_i\}_{i\in I}$ is called the *obtained information set.* Each informant may contribute a finite set of formulas which expresses his share in the information exchange; this finite set clearly can be reduced to a single formula by taking the corresponding conjunction formula.[1] Furthermore, the informants' original belief sets are represented by a mapping Ψ from the informant set I into the powerset of the formula set. We also use the notation $\{\Psi_i\}_{i\in I}$ to denote the set $\{\Psi(i) \in \mathcal{P}(\mathbf{L_P}) : i \in I\}$, which is called an *original information set.* These sets $\{\Psi_i\}_{i\in I}$ are not required to be finite. In this paper, we only consider the case where all informants honestly offer information they actually support. This leads to the following definition:

Definition 2.1 (Potential information set) *An original information set $\{\Psi_i\}_{i\in I}$ is said to be a potential information set of an obtained information set $\{\psi_i\}_{i\in I}$ iff it satisfies the following conditions:*
(i) (Honesty Condition) $\Psi(i) \models \psi(i)$, for all $i \in I$, and
(ii) (Consistency Condition) $\Psi(i)$ is consistent, for all $i \in I$.

In the sequel we shall use the word set for information set when no confusion can arise.

Definition 2.2 (Danger) *Suppose that some original set $\{\Psi_i\}_{i\in I}$ is a potential set of an obtained set $\{\psi_i\}_{i\in I}$. Then the set $\{\psi_i\}_{i\in I}$ is said to be* dangerous *with respect to the set $\{\Psi_i\}_{i\in I}$ iff there exists a $\varphi \in \mathbf{L_P}$ such that*
(i) $\{\psi_i\}_{i\in I} \models \varphi$
(ii) $\Psi(i) \models \neg\varphi$ for all $i \in I$.

Remarks: Condition (i) means that the receiver's obtained information implies some fact φ for which according to Condition (ii) all informants originally believe its negation. The more general notion where some derivable fact φ is disbelieved by some but not necessarily all informants is not interesting for our purposes; a contributed set will be "dangerous" in this more general sense with respect to an original set, unless it represents a proposition which is already compatible with the original belief set of all informants. The latter situation is frequently considered in artificial intelligence, where collected information always represent a partial description of the true world. In our approach we don't require such a true world in the background; we just want to ensure that derivable information is at least compatible with the beliefs of some agent.
In the following, $\{\psi_i\}_{i\in I}$ and $\{\Psi_i\}_{i\in I}$ denote an obtained set and an original set respectively if it cannot cause any ambiguities.

[1] Here we use the fact that the languages considered in this paper are closed under conjunction; the case where we don't assume this closure property is a subject for further research.

Definition 2.3 (Absolute Safety) *A consistent set $\{\psi_i\}_{i\in I}$ is said to be* absolutely safe *iff it is not the case that $\{\psi_i\}_{i\in I}$ is dangerous with respect to any of its potential sets $\{\Psi_i\}_{i\in I}$.*

Definition 2.4 (Triviality) *A set $\{\psi_i\}_{i\in I}$ is trivial iff for any formula φ, such that $\{\psi_i\}_{i\in I} \models \varphi$, there exists an $i \in I$ such that $\psi(i) \models \varphi$.*

Clearly a set is trivial iff some formula $\psi(i)$ is logically equivalent to $\bigwedge\{\psi_i\}_{i\in I}$, which means that in fact one informant has already contributed all available information by himself. This observation easily follows by taking $\varphi = \bigwedge\{\psi_i\}_{i\in I}$

It turns out that absolute safety is a condition which is so strong that it supports only trivial situations:

Theorem 2.1 (Triviality Theorem) *A consistent set $\{\psi_i\}_{i\in I}$ is absolutely safe iff it is trivial.*

The proof for this result is easy. Assuming non-triviality there exists a proposition ϕ such that for no i one has $\{\psi_i\}_{i\in I} \models \varphi$; consequently the potential set $\Psi(i) = \{\psi(i), \neg\varphi\}$, for all $i \in I$ is dangerous with respect to $\{\psi_i\}_{i\in I}$. The converse implication is a direct consequence of the triviality condition.

Consequently, the best one can hope for are safety notions which explicitly relate the obtained set and the potential set. Two possible definitions are:

Definition 2.5 (Safety) *If an obtained $\{\psi_i\}_{i\in I}$ is consistent, and an original set $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is said to be* safe *with respect to the set $\{\Psi_i\}_{i\in I}$ iff the set $\{\psi_i\}_{i\in I}$ is not dangerous with respect to the set $\{\Psi_i\}_{i\in I}$.*

Definition 2.6 (Strong Safety) *If a set $\{\psi_i\}_{i\in I}$ is consistent, and $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is said to be* strongly safe *with respect to the set $\{\Psi_i\}_{i\in I}$ iff*
for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then there exists an $i \in I$ such that $\Psi(i) \models \varphi$.

The connection between these two notions is illustrated by the following:

Propositions 2.1 *If $\{\psi_i\}_{i\in I}$ is a consistent set, and $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is safe with respect to its potential set $\{\Psi_i\}_{i\in I}$ iff*
for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then it is not the case that for all $i \in I$, $\Psi(i) \models \neg\varphi$.

So where the safety condition requires that every derivable formula φ is not disbelieved by all informants, the condition of strong safety requires that at least one of the informants positively supports φ. It follows that strong safety is a stronger notion than safety; see the third example below.

Example 2.1

- *$\langle p, p \to q\rangle$ is neither strongly safe nor safe with respect to the potential set $\langle\{p, \neg q\}, \{\neg p, \neg q\}\rangle$.* (**Judge puzzle**)
- *$\langle p \to q, q \to p\rangle$ is strongly safe and safe with respect to $\langle\{\neg p, q\}, \{\neg p, \neg q\}\rangle$.*
- *$\langle p, p \to q\rangle$ is safe with respect to $\langle\{p, p \vee q\}, \{p \to q, q \to p\}\rangle$, but not strongly safe with respect to $\langle\{p, p \vee q\}, \{p \to q, q \to p\}\rangle$.* (**Distinction between safety and strong safety**)

One can easily give alternative characterizations of these safety notions. Evidently a trivial obtained set is strongly safe with respect to every potential set. The safety notions are moreover trivial for the case of a single informant. The two theorems below relate safety to consistency and to disagreement:

Theorem 2.2 (Safety Theorem) *If an obtained set $\{\psi_i\}_{i\in I}$ is consistent, and $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$, then the set $\{\psi_i\}_{i\in I}$ is safe with respect to the original set $\{\Psi_i\}_{i\in I}$ iff there exists an $i \in I$ such that $\Psi(i) \cup \{\psi_i\}_{i\in I}$ is consistent.*

Lemma 2.1 *If a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$, then for all $i \in I$, $\psi(i) \not\models \bigwedge\{\psi_i\}_{i\in I}$ and $\Psi(i) \models \neg\bigwedge\{\psi_i\}_{i\in I}$.*

PROOF. Suppose that a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$. Then, by definition, there exists a φ such that $\{\psi_i\}_{i\in I} \models \varphi$ and $\Psi(i) \models \neg\varphi$ for all $i \in I$. Therefore, $\bigwedge\{\psi_i\}_{i\in I} \models \varphi$, and consequently $\models \bigwedge\{\psi_i\}_{i\in I} \rightarrow \varphi$ and by contraposition $\models \neg\varphi \rightarrow \neg\bigwedge\{\psi_i\}_{i\in I}$.
However, since $\{\psi_i\}_{i\in I}$ is dangerous $\Psi(i) \models \neg\varphi$ for any $i \in I$ and therefore, $\Psi(i) \models \neg\bigwedge\{\psi_i\}_{i\in I}$ for any $i \in I$.
Finally, it is easy to see that for any $i \in I$, $\psi(i) \not\models \bigwedge\{\psi_i\}_{i\in I}$, because, if $\psi(i) \models \bigwedge\{\psi_i\}_{i\in I}$ for any $i \in I$, then $\{\psi_i\}_{i\in I}$ is trivial, and then, by the triviality theorem, $\{\psi_i\}_{i\in I}$ is absolutely safe, whence $\{\psi_i\}_{i\in I}$ cannot be dangerous with respect to any potential set, and a contradiction follows. □

Theorem 2.3 (Disagreement Theorem) *If a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$, then there exists for every $j \in I$ some formula φ and an $i \in I$ such that $\psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi$.*

PROOF. Suppose that a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$. Then by the above lemma, we have,
(A) $\Psi(j) \models \neg\bigwedge\{\psi_i\}_{i\in I}$ for all $j \in I$.
Now, suppose that the conclusion (B) of the disagreement theorem is false, then we have (C).
(B) $(\forall j \in I)(\exists\varphi)(\exists i \in I)(\psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi)$.
(C) $(\exists j \in I)(\forall\varphi)(\forall i \in I)(\psi(i) \models \varphi \Rightarrow \Psi(j) \models \varphi)$.
However, we know that $\psi(i) \models \psi(i)$ for any $i \in I$. Therefore, by (C), we have,
$(\exists j \in I)(\forall i \in I)(\Psi(j) \models \psi(i))$.
So, we have,
(D) $(\exists j \in I)(\Psi(j) \models \bigwedge\{\psi_i\}_{i\in I})$.
Consequently, from (D) and (A), we conclude that this particular $\Psi(j)$ is inconsistent, contradicting our assumption that $\{\Psi_i\}_{i\in I}$ is a potential set. □

Observe that the conclusion of the disagreement theorem can't be strengthened to a form which expresses definite disagreement: *there exists for every $j \in I$ some formula φ and an $i \in I$ such that $\psi(i) \models \varphi$ and $\Psi(j) \models \neg\varphi$*. This conclusion only can be proven if we assume that the sets $\{\Psi_i\}_{i\in I}$ satisfy the strong condition $\Psi(j) \models (\phi \vee \psi)$ iff $\Psi(j) \models \phi$ or $\Psi(j) \models \psi$, i.e., if we assume that our agents use an intuitionistic interpretation of disjunction.

Corollary 2.1 *If a consistent set $\{\psi_i\}_{i\in I}$ is dangerous with respect to a potential set $\{\Psi_i\}_{i\in I}$, then there exist for every $j \in I$ an $i \in I$ and a formula φ in the consequence set $Cn(\{\psi(i)\})$ such that $\Psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi$; such a formula φ is called a* disagreement formula *for j.*

In the judge puzzle story, the formula $p \rightarrow q$ is a disagreement formula for w_1, since $\Psi(w_1) = \{p, \neg q\} \not\models p \rightarrow q$ and $\Psi(w_2) = \{\neg p, \neg q\} \models p \rightarrow q$. The implication of the disagreement theorem is that in a multi-agent information system in order to guarantee safety, agents must be prohibited to talk about something

if they disagree with someone else about it! Therefore each agent will need full information about the others' propositional attitudes, and this clearly represents an unrealistic assumption. Nonetheless the result implies that we should focus on the cases where disagreement may arise, and look for mechanisms for coping with it. As we indicate below, the logic of belief dependence turns out to be a useful tool in this direction.

3 Logic of belief dependence

The logic of belief dependence was introduced [0] in order to model the situation where agents rely on each other with respect to information. It also provides a tool for modeling a two stage process for information acquisition in a multi-agent system: in the first stage agents include information of other agents in *compartimentalized sub-beliefs* and in the second stages these sub-beliefs are processed and *incorporated* into the agents own beliefs. For further information and motivation we refer to [0].

Our logic contains in the first place the general notions of knowledge and belief; these notions are the equivalents of those in epistemic and doxastic logic. For our purposes the difference between knowledge and belief is almost never important. Therefore we generally use $L_i\varphi$ to represent the fact that agent i knows or believes the formula φ. If we need to emphasize that we are talking about knowledge rather than belief, we will enforce this by adding the required axioms for the operators L_i.

The second important notion used for reasoning about dependent knowledge and beliefs is called the *dependent operator*, or alternatively *rely-on relation*, and it is denoted by $D_{i,j}$. Intuitively, we can give $D_{i,j}\varphi$ a number of different interpretations: "agent i relies on agent j about the formula φ", or, "agent j is the credible advisor of agent i about φ".

The dependent operator resembles the awareness operator introduced by Fagin and Halpern [0] in the sense that it operates on the formulas rather than their meaning. For example it is possible in our system that $D_{i,j}(p \wedge q)$ whereas $D_{i,j}(q \wedge p)$ comes out to be false. Evidently one may prevent such anomalies by axiomatizing the $D_{i,j}$ operator. The most natural condition to enforce is neutrality: $D_{i,j}\varphi \leftrightarrow D_{i,j}\neg\varphi$, which states that dependence is independent of whether some fact is stated in a positive or negative way. A slightly less convincing condition is closure under conjunction; the problem is that closure under conjunction in combination with neutrality enforces that agent i depends on agent j concerning the true and false proposition as soon as i depends on j concerning anything at all. The reader is referred to [0, 0] for further information on reasonable axioms for the dependent operator.

The third element in our logic is the *compartment operator*, or alternatively called the *sub-belief operator*, written $L_{i,j}$. Intuitively, $L_{i,j}\varphi$ can be read "agent i believes φ due to agent j". From the viewpoint of *minds society*, $L_{i,j}\varphi$ can be more intuitively interpreted as "agent i believes or knows φ on the mind frame indexed j.

The resulting language is sufficiently rich for formalizing both stages in the multi-agent information acquisition process mentioned above: compartmentalized information is modeled by sub-beliefs $L_{i,j}\varphi$ for agent i, whereas incorporated information corresponds to general beliefs of agent i, namely, $L_i\varphi$.

Supposed we have a set $\mathbf{A_n}$ of n agents, and a set Φ_0 of primitive propositions, the language $\mathbf{L_D}$ for belief dependence logics is the minimal set of formulas closed by usual syntactic rules.

Definition 3.1 (D-model) *A belief dependence D-model is a tuple $M = (S, \pi, \mathcal{L}, \mathcal{D})$, where S is a set of states, $\pi(s,.)$ is a truth assignment for each state $s \in S$, and $\mathcal{L} : \mathbf{A_n} \rightarrow \mathcal{P}(S \times S)$, which consists of n binary serial accessibility relations on S, and $\mathcal{D} : \mathbf{A_n} \times \mathbf{A_n} \times S \rightarrow \mathcal{P}(\mathbf{L_D})$.*

Remarks: Note that the structure of the D-model is similar to the semantic model in Fagin and Halpern's general awareness logic, which was designed to cope with the problem of logical omniscience [0].

The truth relation $\models$ is defined inductively as follows:

$M, s \models p$, iff $\pi(s,p)$ =true, for p a primitive proposition
$M, s \models \neg\varphi$ iff $M, s \not\models \varphi$
$M, s \models \varphi_1 \wedge \varphi_2$ iff $M, s \models \varphi_1 \wedge M, s \models \varphi_2$,
$M, s \models L_i\varphi$ iff $M, t \models \varphi$ for all t such $(s,t) \in \mathcal{L}(i)$
$M, s \models D_{i,j}\varphi$ iff $\varphi \in \mathcal{D}(i,j,s)$.

For D-models, we define sub-beliefs as $L_{i,j}\varphi \stackrel{\text{def}}{=} D_{i,j}\varphi \wedge L_j\varphi$; this implies that agents in our system are honest because the honesty axiom $L_{i,j}\varphi \rightarrow L_j\varphi$ holds.[2]

A minimal logic system for D-models, called **LD system** consists of the following axioms and rules:

Axioms:

(BA) All instances of propositional tautologies.
(KL) $L_i\varphi \wedge L_i(\varphi \rightarrow \psi) \rightarrow L_i\psi$.
(DL) $\neg L_i\bot$.

Rules of Inference:

(MP) $\vdash \varphi, \vdash \varphi \rightarrow \psi \Rightarrow \vdash \psi$.
(NECL) $\vdash \varphi \Rightarrow \vdash L_i\varphi$.

Definitions:

(Lijdf) $L_{i,j}\varphi \stackrel{\text{def}}{=} D_{i,j}\varphi \wedge L_j\varphi$

Theorem 3.1 *The logic* **LD** *is sound and complete for the class of D-models [0].*

We call a formula set which is consistent (with the logic **LD**) a *belief state*. A *belief set* is a subset of a belief state consisting of formula's specifically relevant to some particular agent.

For the dynamic part of our proposed solution to the Judge puzzle we need a belief maintenance operation which will be invoked during the second stage of the information assimilation. We introduce the notion of the *belief maintenance model*, which is an ordered couple $< \mathbf{K}, \Delta >$ such that $\mathbf{K}$ is a collection of belief sets and $\Delta : \mathbf{K} \times \mathbf{L_D} \rightarrow \mathbf{K}$ is a function assigning a belief set $\Delta(K,\varphi)$ to any belief set $K \in \mathbf{K}$ and each formula φ in $\mathbf{L_D}$. We shall write alternatively $K\Delta\varphi$ to represent $\Delta(K,\varphi)$. Again our proposed framework is liberal with respect to the choice of the belief maintenance operation used. We will return to this issue in section 6.

4 Information acquisition in a belief dependence framework

In this section, we consider the information acquisition problem in our framework of belief dependence logic. The extension the definitions which have appeared in section 2 for the case of belief dependence is easy: one simply replaces the propositional language $\mathbf{L_P}$ by the language $\mathbf{L_D}$, propositional models by D-models, and the relation $\models$ for propositional logic by its counterpart for belief dependence logic. Consequently, whenever we say a formula set K is consistent, we mean that K is consistent with respect to the **LD** system unless stated otherwise.

In the resulting theory the (negative) results from section 2 remain valid, indicating that for a solution of problems like the Judge puzzle the formalization of the relevant information into the language of belief dependence logic by itself will be insufficient in order to remove the observed anomaly.

The translation between the propositional formulation of our problem and its formalization in terms of belief states K in the logic of belief dependence invokes a few auxiliary notations defined below:
$L^{-}_{i,j}(K) \stackrel{\text{def}}{=} \{\varphi \in \mathbf{L_D} : K \models L_{i,j}\varphi\}$, denotes agent i's compartimentalized belief set indexed j.

[2] In the section 5 we will need the more demanding notion of reliable sub-beliefs $L^{r}_{i,j}\varphi$ which is defined as $D_{i,j}\varphi \wedge D_{j,j}\varphi \wedge L_j\varphi$.

$L_i^-(K) \stackrel{\text{def}}{=} \{\varphi \in \mathbf{L_D} : K \models L_i\varphi\}$, denotes agent i's (incorporated) belief set.
$L_{a,I}^+(\{\psi_i\}_{i\in I}) \stackrel{\text{def}}{=} \{L_{a,i}\psi(i) \in \mathbf{L_D} : i \in I\}$, is a formula set expressing the information obtained by the receiver from the informants before the receiver has incorporated (part of) this information.

The notion of a *configuration* represents the generalization of a potential set from section 2:

Definition 4.1 (Configuration) *A configuration C is a four place tuple $\langle a, I, \psi, \Psi\rangle$, where $a \in A_n$ denotes an agent, called receiver, $I \subseteq A_n$ is a finite and non-empty set of informants, $\psi : I \to \mathbf{L_D}$ is a mapping from I into $\mathbf{L_D}$, called the obtained information, and $\Psi : I \to \mathcal{P}(\mathbf{L_D})$ is a mapping from I into the powerset of $\mathbf{L_D}$, called the original information.*

Since the required proofs are insensitive to the precise logical language being used it will not surprising that the main results of section 2 remain valid for the logic of belief dependence:

Theorem 4.1 (Triviality Theorem(Restated)) *A consistent obtained set $\{\psi_i\}_{i\in I}$ is absolutely safe iff it is trivial.*

Theorem 4.2 (Disagreement Theorem(Restated)) *Let $C = \langle a, I, \psi, \Psi\rangle$ be a configuration. Suppose that $\{\psi_i\}_{i\in I}$ is consistent, and $\{\Psi_i\}_{i\in I}$ is a potential set of $\{\psi_i\}_{i\in I}$. If $\{\psi_i\}_{i\in I}$ is dangerous with respect to the set $\{\Psi_i\}_{i\in I}$, then there exists for every agent $j \in I$ a formula φ and agent $i \in I$ such that $\psi(i) \models \varphi$ and $\Psi(j) \not\models \varphi$.*

For a belief state K in belief dependence logic and an agent a, we want to induce a configuration for a from K. In this induced configuration agent a becomes the receiver and the remaining agents become the informants. Both the contributed information and the original information is obtained from the belief set K as indicated below.

Definition 4.2 (Induced Configuration) *Suppose that K be a belief state, and a be an agent $\in A_n$. A configuration $C = \langle a, I, \psi, \Psi\rangle$, called the induced configuration for a from K, is constructed as follows:*
(1) I is the set $\{i \in A_n : \exists\varphi(L_{a,i}\varphi \in K)\}$.
(2) If I is not empty, then for all $i \in I$, $\Psi(i) = L_i^-(K)$, otherwise the induced configuration does not exist.
(3) For all $i \in I$, if $L_{a,i}^-(K)$ is finite, then let $\psi(i)$ be $\wedge L_{a,i}^-(K)$, otherwise the induced configuration does not exist.

For an agent $a \in A_n$ a belief state K is said to be a *DB set for a* iff the induced configuration for a from K exists. Evidently the induced configuration $\langle a, I, \psi, \Psi\rangle$ for a from K is unique whenever it exists. We introduce the notation $C(a, K)$ for the induced configuration for a from K. Moreover, due to the honesty condition contained in definition (Lijdf) the original information set $\Psi(i)$ is a potential set for $\psi(i)$ for each $i \in I$. The concept of the induced configuration makes it possible to translate the safety definitions from section 2 to belief states in belief dependence logic:

Definition 4.3 (Safety for a in K) *For an agent $a \in A_n$ and a DB set K for a, let $C(a, K) =< a, I, \psi, \Psi >$ be the induced configuration for a from K, then $\{\psi_i\}_{i\in I}$ is said to be safe for a in K iff $\{\psi_i\}_{i\in I}$ is safe with respect to $\{\Psi_i\}_{i\in I}$.*

Theorem 4.3 (Safety Theorem(Restated)) *Let a and K be an agent and a DB set respectively. Suppose that the induced configuration for a from K, $C(a, K) = \langle a, I, \psi, \Psi\rangle$, then $\{\psi_i\}_{i\in I}$ is safe for a in K iff there exists an $i \in I$ such that $L_i^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent.*

5 Almost safety

In order to evaluate whether obtained information is safe the receiver a still needs information on the true belief states of his informants; the translation into the belief dependence logic and the introduction of configurations has not changed this necessity. However, if we take into consideration which mechanisms might have produced the sub-beliefs in a multi-agent environment, it turns out that these mechanisms themselves may provide us with additional structure supporting the introduction of alternative and weaker safety notions.

The notion of *almost safety* defined in this section is based on one possible hypothesis concerning the creation of sub-beliefs: the so-called *initial role-knowledge assumption*. This hypothesis states that within a multi agent environment the dependency relations are common knowledge: it is not known who knows what or who believes what, but for each proposition it is known how the agents depend on each other concerning this proposition.

That this information is relevant is shown by the example below. Assume that some agent i believes ϕ and says so to the receiver. Suppose moreover that the receiver has learned previously that agent j believes $\neg\phi$. Finally agent i depends on agent j concerning ϕ. According to the initial role-knowledge assumption it is common knowledge that $D_{i,j}\phi$, so the receiver knows that as well. In this situation the receiver can conclude that something strange is going on: would the two agents i and j have been given the possibility to exchange their information, agent i would have been convinced by j that his belief concerning ϕ was wrong. Moreover, this prediction can be made by the receiver without any further interaction with the informants! It is based on this information that the receiver can disregard the information provided by i substituting it by the opposite information provided by agent j.

The notion of almost safety formalizes safety with respect to the hypothetical scenario which will arise when all informants exchange their information before sharing their knowledge with the receiver. In order to be able to reason about these hypothetical belief states we need one further notion:

Definition 5.1 (Combined Sub-belief)
$L^{-}_{i,I}(K) \stackrel{\text{def}}{=} \{\varphi \in \mathbf{L_D} : (\exists j \in I)(K \models D_{i,j}\varphi \wedge D_{j,j}\varphi \wedge L_j\varphi)\}$.

The notion of almost safety is obtained from the safety notion by allowing for one more propositional attitude for an informant with respect to the consequences of the contributed information (clause (ii) below):

Definition 5.2 (Almost Safety) *For an agent $a \in A_n$ and a DB set K for a, if $C(a,K) = \langle a, I, \psi, \Psi\rangle$ is the induced configuration for a from K, $\{\psi_i\}_{i\in I}$ is said to be* almost safe *for a in K iff for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then, either (i) there exists $i \in I$ such that $L^{-}_i(K) \not\models \neg\varphi$, or (ii) there exists an $i \in I$ such that $L^{-}_{i,I}(K)$ is consistent and $L^{-}_{i,I}(K) \models \varphi$.*

We illustrate this notion by an example of a possible configuration for the judge puzzle:
Consider a DB set $K = \{L_{w1}(p \wedge \neg q),\ L_{w2}((p \to q) \wedge \neg q),\ D_{w1,w1}p,\ D_{w2,w2}(p \to q),\ D_{w2,w1}p, L_{a,w1}p,\ L_{a,w2}(p \to q)\}$ So, $I = \{w1, w2\}$, $\psi(w1) = p$, $\psi(w2) = p \to q$, $\Psi(w1) = \{p \wedge \neg q\}$, $\Psi(w2) = \{(p \to q) \wedge \neg q\}$. The induced configuration for a from K is $\langle a, I, \psi, \Psi\rangle$, with $\{\psi_i\}_{i\in I} = \{p, p \to q\}$. Moreover, from $K \models L_{w1}p \wedge D_{w1,w1}p \wedge D_{w2,w1}p$, we have $L^{-}_{w2,I}(K) = \{p, p \to q\}$ so $L^{-}_{w2,I}(K)$ is consistent. Evidently, for any φ, if $\{\psi_i\}_{i\in I} \models \varphi$, then $L^{-}_{w2,I}(K) \models \varphi$. Therefore, $\{\psi_i\}_{i\in I}$ is almost safe for a in K.

It is a direct consequence of the definition that almost safety is a weaker notion than safety. The notions turn out to be equivalent in the degenerate case that the informants don't rely on each other concerning any proposition. By a straightforward generalization of previous characterizations we obtain:

Theorem 5.1 (Almost Safety Theorem) *Let $a \in A_n$ be an agent, K be a DB set for a, and $C(a,K) = \langle a, I, \psi, \Psi \rangle$ be the induced configuration for a from K; then $\{\psi_i\}_{i\in I}$ is almost safe for a in K iff there exists an $i \in I$ such that either $L_i^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent, or $L_{i,I}^-(K)$ is consistent and $L_{i,I}^-(K) \models \{\psi_i\}_{i\in I}$.*

This theorem establishes that almost safety is a property which, in principle, for a given configuration, can be tested effectively: for configuration $C = \langle a, I, \psi, \Psi \rangle$, we say that *almost-safety test statement* (ASTS) holds in C iff:
$(\exists i \in I)(L_i^-(K) \cup \{\psi_i\}_{i\in I}$ is consistent or $(L_{i,I}^-(K)$ is consistent and $L_{i,I}^-(K) \models \{\psi_i\}_{i\in I}))$.

6 Almost Safety on Belief Maintenance Operation

In this section we consider the dynamic process of belief revision corresponding to the second stage of the two stage information acquisition process mentioned in section 3. Given a configuration where the receiver has obtained sub-beliefs by hearing statements by his informants, the receiver will subsequently revise his own belief by incorporation part of these sub-beliefs into his own belief. Clearly he should do so in a safe way; we now have the tools available for formalizing this requirement.

Let $\mathbf{K}$ be a collection of belief sets. As announced before a belief maintenance operation $\triangle : \mathbf{K} \times \mathbf{L_D} \rightarrow \mathbf{K}$ is a function assigning a belief set $\triangle(X, \varphi)$ to any belief set $X \in \mathbf{K}$ and each formula φ in $\mathbf{L_D}$.

A belief maintenance operation $\triangle$ can be defined in many ways. For our application we are interested in belief maintenance operators with a special form: the rational agent checks whether or not a special formula φ_i belongs to the belief set X when she faces the new information ρ_i'. If so, then the result of $\triangle(X, \rho_i')$ is a new belief set Y_i.[3].

$$\triangle(X,\rho) = \begin{cases} Y_{i_1} & \text{if } \varphi_{i_1} \in X, \rho = \rho'_{i_1} \\ \cdots & \cdots \\ Y_{i_n} & \text{if } \varphi_{i_n} \in X, \rho = \rho'_{i_n} \\ X & \textbf{otherwise} \end{cases}$$

For belief maintenance operation $\triangle$, we can use a set of rules with the following form to simplify the representation:

$$\varphi_i \Rightarrow X \triangle \rho_i' = Y_i$$

Each rule represents some case in the definition of the function. We omit the rule representing the "otherwise" case since it represents the default. Intuitively, each rule of the above form says that if φ_i holds in the belief set X, then the result of the maintenance with the new information ρ_i' is Y_i.

We are considering possible safe belief maintenance operations for the receiver a in some belief state K representing the actual exchange of information. For this situation the relevant belief set X equals $L_a^-(K)$.

As in our previous paper [0] we use traditional update operations like *revision*, *contraction* and *expansion*, to define the belief maintenance operation. It is known that, aside from the trivial but unsafe expansion operator, such belief revision and contraction functions are non-trivial to construct and certainly not unique. Therefore we assume that we have selected some group of revision functions to be used in the sequel. In particular we assume that these operations satisfy the AGM postulates [0]. The selected revision function will be denoted $\dotplus$. We define the corresponding contraction function $\dot{-}$ using the Harper Identity in terms of the revision function.

[3]This form is called a *type 3 belief maintenance operation* in [0]

Having made these choices, we obtain a definition for our belief revision operation $\triangle$ in the following form:[4]:

$$\triangle(L_a^-(K),\rho) = \begin{cases} L_a^-(K)\theta_1\psi_{i_1} & \text{if } \varphi_{i_1} \in L_a^-(K), \rho = \rho'_{i_1} \\ \ldots & \ldots \\ L_a^-(K)\theta_n\psi_{i_n} & \text{if } \varphi_{i_n} \in L_a^-(K), \rho = \rho'_{i_n} \\ K & \textbf{otherwise} \end{cases}$$

where $\theta_i \in \{\dot{+}, \dot{-}, +\}$.

This form of the belief revision operator shows that the receiver only revises his private knowledge on the basis of formulas contained therein. Our goal is to define an AS operation for the receiver a with respect to the obtained set $\{\psi_i\}_{i\in I}$. Also the revision should lead to the incorporation of the obtained set, since we want to determine under which circumstances it is safe to do so. Recall that $\wedge L^+_{a,I}(\{\psi_i\}_{i\in I})$ and $\wedge\{\psi_i\}_{i\in I}$ denote respectively the compartimentalized belief and the incorporated belief which corresponds the obtained set $\{\psi_i\}_{i\in I}$. In the sequel these two important formulas will be denoted by $\mathbf{cpart}(\psi)$ and $\mathbf{incorp}(\psi)$ respectively.

Definition 6.1 (AS Operation) *A belief maintenance operation $\triangle : \mathbf{K} \times L_D \rightarrow \mathbf{K}$ is said to be an almost safety one for agent $a \in A_n$ with respect to $\{\psi_i\}_{i\in I}$, iff for any DB set $K \in \mathbf{K}$ for a such that $L_a^-(K) \models \mathbf{cpart}(\psi)$ and $L_a^-(K) \not\models \mathbf{incorp}(\psi)$, it will be the case that $\triangle(L_a^-(K),\mathbf{cpart}(\psi)) \models \mathbf{incorp}(\psi)$ only when $\{\psi_i\}_{i\in I}$ is almost safe for a in K.*

Remarks: (i) We define almost safety for a belief maintenance operation in terms of the general almost-safety notion.
(ii) We consider only the case where the knowledge state K is a DB set for a since otherwise the induced configuration does not exist, and consequently the concept of almost safety does not make sense.
(iii) $L_a^-(K) \models \mathbf{cpart}(\psi)$ means that the receiver a has full knowledge about his compartimentalized information $\wedge L^+_{a,I}(\{\psi_i\}_{i\in I})$.
(iv) $L_a^-(K) \not\models \mathbf{incorp}(\psi)$ and $\triangle(L_a^-(K),\mathbf{cpart}(\psi)) \models \mathbf{incorp}(\psi)$ together means that we consider only the case where the receiver a really assimilates the obtained information.

In other words, agent a originally does not fully believe the fact $\wedge\{\psi_i\}_{i\in I}$, but by invoking the operation, she fully believes this fact. The format for our belief revision operator therefore further specializes to:

$$\triangle(L_a^-(K),\mathbf{cpart}(\psi)) = \begin{cases} L_a^-(K)\dot{+}\mathbf{incorp}(\psi) & \text{if } \varphi_{i_1} \in L_a^-(K) \text{ or } \ldots \text{ or } \varphi_{i_n} \in L_a^-(K) \\ L_a^-(K) & \textbf{otherwise} \end{cases}$$

For this type of belief revision operator we can characterize almost safety:

Theorem 6.1 (AS Operation Theorem) *Consider a belief maintenance operation $\triangle$ of the form:*

$$\triangle(L_a^-(K),\mathbf{cpart}(\psi)) = \begin{cases} L_a^-(K)\dot{+}\mathbf{incorp}(\psi) & \textit{if } \varphi_{i_1} \in L_a^-(K) \textit{ or } \ldots \textit{ or } \varphi_{i_n} \in L_a^-(K) \\ L_a^-(K) & \textbf{otherwise} \end{cases}$$

Operation $\triangle$ is an AS operation for a with respect to $\{\psi_i\}_{i\in I}$ iff every assumption in the sequence $\varphi_{i_1} \in L_a^-(K)$ or ... or $\varphi_{i_n} \in L_a^-(K)$ entails that the almost-safety test holds in $C(a,K)$.

Evidently, our goal is to define an AS operation for general cases. There remain however complications. For example it is not possible to check that a set of formulas K is consistent by testing whether particular formulas belong to K or not. Therefore we need some further assumptions. We only consider DB sets K for a for which the combined sub-belief sets $L^-_{i,I}(K)$ are consistent. Another additional condition is that

[4] called a *type 5 belief maintenance operation* in [0]

we only consider DB set K for an agent a for which knowledge and belief coincide: $K \models L_a\varphi \rightarrow \varphi$ for any φ. We call such an agent *a a skeptic agent in* K. An operation which is AS under the above two additional assumptions will be called a *restricted AS operation.*

Definition 6.2 (Restricted AS Operation) *A belief maintenance operation* $\triangle : \mathbf{K} \times L_D \rightarrow \mathbf{K}$ *is said to be a restricted almost safe one for agent* $a \in A_n$ *with respect to* $\{\psi_i\}_{i\in I}$, *iff for any DB set* $K \in \mathbf{K}$ *for a such that (i)* $L_a^-(K) \models$**cpart**(ψ), *(ii)* $L_a^-(K) \not\models$**incorp**(ψ), *(iii) agent a is a skeptic agent in* K, *and (iv) any combined sub-belief set from* K *is consistent, it holds that* $\triangle(L_a^-(K),$**cpart**$(\psi)) \models$**incorp**(ψ) *only when psii is almost safe for a in* K.

Theorem 6.2 (Restricted AS Operation Theorem) *Suppose that a belief maintenance operation* $\triangle$ *is a type 5 operation like:*

$$\triangle(L_a^-(K),\mathbf{cpart}(\psi)) = \begin{cases} L_a^-(K)\dot{+}\{\psi_i\}_{i\in I} & \text{if } \varphi_{i_1} \in L_a^-(K) \text{ or } \ldots \text{ or } \varphi_{i_n} \in L_a^-(K) \\ L_a^-(K) & \textbf{otherwise} \end{cases}$$

$\triangle$ *is a restricted AS operation for a with respect to* $\{\psi_i\}_{i\in I}$ *iff for every belief state* K *satisfying the conditions (i), (ii), (iii) and (iv) above, every assumption in the sequence* $\varphi_{i_1} \in L_a^-(K)$ *or ...* *or* $\varphi_{i_n} \in L_a^-(K)$ *entails that the almost-safety test holds in* $C(a, K)$.

After these preparations we are finally ready to define a restricted AS operation for the agent a with respect to the obtained information $\{\psi_i\}_{i\in I}$ where $I = \{i_1, \ldots i_k\}$. The defined operation considers two kinds of typical situations. The first situation is that each informant fully relies both on other informants and on herself about what they say, i.e., $\bigwedge_{j=1}^{k}\bigwedge_{j'=1}^{k} D_{i_j,i_{j'}}\psi(i_{j'})$. In this situation, every informant plays a role of an "expert" on the information she offers as defined in [0]. Note that for each informant i_l, the above condition can be reduced to the condition $\bigwedge_{j=1}^{k} D_{i_l,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j)$.

The second situation is one which already supports a stronger notion of safety, meaning that some informant i_l considers the obtained set consistent with her beliefs, i.e., $\neg L_{w_{i_l}} \neg \bigwedge_{j=1}^{k} \psi(i_j)$. Since according to the honesty condition, each informant i_l already believes what she offers, the above condition can be weakened to the less restrictive condition $\neg L_{w_{i_l}} \neg \bigwedge_{j=1,j\neq l}^{k} \psi(i_j)$. Formally we define our operation as a simplified type 5 operation as follows:

The Definition of Operation $\triangle_{ras1}$ (for Agent a):

(A1) $\bigwedge_{j=1}^{k} D_{i_1,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j) \Rightarrow L_a^-(K)\triangle_{ras1}$**cpart**$(\psi) = L_a^-(K)\dot{+}$**incorp**$(\psi)$.

(A2) $\bigwedge_{j=1}^{k} D_{i_2,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j) \Rightarrow L_a^-(K)\triangle_{ras1}$**cpart**$(\psi) = L_a^-(K)\dot{+}$**incorp**$(\psi)$.

.....

(Ak) $\bigwedge_{j=1}^{k} D_{i_k,i_j}\psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j}\psi(i_j) \Rightarrow L_a^-(K)\triangle_{ras1}$**cpart**$(\psi) = L_a^-(K)\dot{+}$**incorp**$(\psi)$.

(B1) $\neg L_{i_1} \neg \bigwedge_{j=2}^{k} \psi(i_j) \Rightarrow L_a^-(K)\triangle_{ras1}$**cpart**$(\psi) = L_a^-(K)\dot{+}$**incorp**$(\psi)$.

(B2) $\neg L_{i_2} \neg \bigwedge_{j=1,j\neq 2}^{k} \psi(i_j) \Rightarrow L_a^-(K)\triangle_{ras1}$**cpart**$(\psi) = L_a^-(K)\dot{+}$**incorp**$(\psi)$.

.....

(Bk) $\neg L_{i_k} \neg \bigwedge_{j=1}^{k-1} \psi(i_j) \Rightarrow L_a^-(K) \triangle_{ras1} \mathbf{cpart}(\psi) = L_a^-(K) \dot{+} \mathbf{incorp}(\psi)$.

For the above operation, the cases (A1)-(Ak) are representative for the original problem as posed by Schoenmakers, since we need no further information about source agents' beliefs other than the general information about the rely-on relations among agents. The cases (B1)-(Bk) deal with the situation where agent a may have previously collected some information about the source agents' beliefs and the obtained information is already safe. Although these situations are not representative for our problem, handling those situation is necessary for obtaining a more general operation.

Theorem 6.3 *The operation $\triangle_{ras1}$ is a restricted AS operation for agent a with respect to $\{\psi_i\}_{i \in I}$.*

PROOF. Let $A(l) = \bigwedge_{j=1}^{k} D_{i_l,i_j} \psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j} \psi(i_j)$, where $l \in \{1, 2, ..., k\}$;
and $B(l) = \neg L_{i_l} \neg \bigwedge_{j=1, j \neq l}^{k}$, where $l \in \{1, 2, ..., k\}$;

We have to show that ($A(1) \in L_a^-(K)$ or $A(2) \in L_a^-(K)$ or ... or $A(k) \in L_a^-(K)$ or $B(1) \in L_a^-(K)$ or ... or $B(k) \in L_a^-(K)$) implies that (ASTS) holds in $C(a, K)$.

It is sufficient to show that (1) $A(l) \in L_a^-(K) \Rightarrow L_{i_l,I}^-(K) \models \{\psi_i\}_{i \in I}$, and
(2) $B(l) \in L_a^-(K) \Rightarrow L_{i_l}^-(K) \cup \{\psi_i\}_{i \in I}$ is consistent, for $1 \leq l \leq k$.

Case (1)
$A(l) \in L_a^-(K)$
$\Rightarrow \bigwedge_{j=1}^{k} D_{i_l,i_j} \psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j} \psi(i_j) \in L_a^-(K)$ (by definition of $A(l)$)
$\Rightarrow K \models L_a(\bigwedge_{j=1}^{k} D_{i_l,i_j} \psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j} \psi(i_j))$ (by Definition of $L_a^-(K)$)
$\Rightarrow K \models \bigwedge_{j=1}^{k} D_{i_l,i_j} \psi(i_j) \wedge \bigwedge_{j=1}^{k} D_{i_j,i_j} \psi(i_j)$ (since a is a skeptic agent)
$\Rightarrow K \models \bigwedge_{j=1}^{k} (D_{i_l,i_j} \psi(i_j) \wedge D_{i_j,i_j} \psi(i_j) \wedge L_{i_j} \psi(i_j))$ (by the honesty condition)
$\Rightarrow \{\psi_i\}_{i \in I} \subseteq L_{i_l,I}^-(K)$ (by definition of $L_{i_l,I}^-(K)$)
$\Rightarrow L_{i_l,I}^-(K) \models \{\psi_i\}_{i \in I}$.

Case (2)
$B(l) \in L_a^-(K)$
$\Rightarrow \neg L_{i_l} \neg \bigwedge_{j=1, j \neq l}^{k} \psi(i_j) \in L_a^-(K)$ (by definition of $B(l)$)
$\Rightarrow K \models L_a(\neg L_{i_l} \neg \bigwedge_{j=1, j \neq l}^{k} \psi(i_j))$ (by definition of $L_a^-(K)$)
$\Rightarrow K \models \neg L_{i_l} \neg \bigwedge_{j=1, j \neq l}^{k} \psi(i_j)$ (since a is a skeptic agent)
$\Rightarrow K \not\models L_{i_l} \neg \bigwedge_{j=1, j \neq l}^{k} \psi(i_j)$ (since K is consistent)
$\Rightarrow L_{i_l}^-(K) \not\models \neg \bigwedge_{j=1, j \neq l}^{k} \psi(i_j)$ (by definition of $L_{i_l}^-(K)$)
$\Rightarrow L_{i_l}^-(K) \not\vdash \neg \bigwedge_{j=1, j \neq l}^{k} \psi(i_j)$ (by soundness)
$\Rightarrow L_{i_l}^-(K) \cup \{\bigwedge_{j=1, j \neq l}^{k} \psi(i_j)\}$ is consistent. (meta reasoning)
$\Rightarrow L_{i_l}^-(K) \cup \{\psi_i\}_{i \in I}$ is consistent. (by honesty)
□

Using the definition of the operation $\triangle_{ras1}$ and the above theorem, it becomes a straightforward application to construct a restricted AS operation for the judge; just consider the special case where

$I = \{w_1, w_2\}$ and $\{\psi_i\}_{i \in I} = \{\psi(w_1) = p, \psi(w_2) = p \rightarrow q\}$, i.e., the agent w_1 offers information p, and agent w_2 offers information $p \rightarrow q$.

The Definition of Operation $\triangle_{jp}$ (for Agent a):

(A1) $D_{w_1,w_2}(p \rightarrow q) \wedge D_{w_1,w_1} p \wedge D_{w_2,w_2}(p \rightarrow q) \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

(A2) $D_{w_2,w_1} p \wedge D_{w_1,w_1} p \wedge D_{w_2,w_2}(p \rightarrow q) \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

(B1) $\neg L_{w_1} \neg q \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

(B2) $\neg L_{w_2} \neg p \Rightarrow L_a^-(K) \triangle_{jp} L_{a,w_1} p \wedge L_{a,w_2}(p \rightarrow q) = L_a^-(K) \dot{+} p \wedge (p \rightarrow q)$.

There remains the task of presenting this rather intricate solution in some more conceptual way. In order to explain our solution to someone who understands the original puzzle but is not able to grasp the full power of the logic machinery called into action, we can present a new sequel to the Judge puzzle story which leads to an unexpected solution. Assuming that the judge drew his conclusion based on our restricted AS operation, we discover that the unacceptability of the state of affairs as indicated by the original story only is represents a temporary stage in the process of exchanging information and incorporation of beliefs. The continuation of the story (the part which Schoenmakers did not include in his paper) goes as follows:

> *When the judge was told that p was true by the witness w1 and learned that the implication $p \rightarrow q$ was true from witness w2, she had to figure out whether these assertions could be accepted together. Now the judge had good reasons for not asking the witnesses for more information about their knowledge, since she could base her decision already on her knowledge of the rely-on relation. She knew that witness w1 was the only authority concerning the statement p, and that witness w2 was the only authority concerning the conditional $p \rightarrow q$. Moreover, this information was common knowledge among both witnesses and herself. Therefore, she could safely conclude that q was true, and consequently she ordered the verdict to be executed. When they learned about this execution both witnesses w1 and w2 came forward and protested against the verdict, claiming that q was false. The judge patiently informed witness w1 about the witness w2's belief that $p \rightarrow q$ was true. Because the witness w1 accepted that w2 was the authority on the implication $p \rightarrow q$, w1 accepted this assertion, and had to agree with the judge. She also told witness w2 about w1's belief, that p was true, and consequently witness w2 also had to agree with her verdict, since w2 accepted that the w1 was the authority about p. In the end everybody was satisfied.*

7 Conclusions

We have formalized the problem of information acquisition in a multi agent environment. The danger of accepting information from several agents as illustrated in the judge puzzle is an inherent consequence of disagreement among the informants; there exists no absolute safe set of obtained information other than trivial sets, and safe or strongly safe sets are defined only relative the full believe state which in general is unknown to the receiver.

Formalizing this problem in a belief dependence framework does not offer an easy way out; however, by assuming the initial role-knowledge assumption, honesty, skepticism for the judge and a few consistency conditions, and by considering a highly specialized belief maintenance operation a restricted almost safe solution for the judge puzzle has been obtained. This solution has moreover the nice property that it is computable.

Notwithstanding its complexity, our solution has some interesting features: it is based on a general theory supported by psychological evidence, and the tools used for the solution were not developed for the purpose of solving the Judge puzzle. We consider it highly unlikely that there exist "cleaner" solutions to this problem (aside of simply denying it to be a problem).

For designers of intelligent database systems and expert systems our results suggest the following guideline: When combining expertise from different expert sources, ensure that the contributing agents involved recognize each other to be the expert on their respective contributions. If the situation should ever arise that some contributing agent starts complaining about the knowledge stored in the system, the designer, by following our guideline, has ensured that during the subsequent debate she won't be forced to redesign the knowledge base; instead the complaining informants will learn something they didn't know before.

References

[1] Fagin, R. F. and Halpern, J. Y., Belief, Awareness, and Limited Reasoning, in: *Artificial Intelligence,* 34 (1988) 39-76.

[2] Peter Gärdenfors, *Knowledge in Flux–Modeling the Dynamics of Epistemic States*, MIT Press, Cambridge, Mass., 1988.

[3] Zhisheng Huang, Logics for Belief Dependence, in: *Proceedings of the 1990 Workshop on Computer Science Logic*(CSL'90), Springer Lecture Notes in Computer Science, LNCS vol 533 (1991), pp. 274–288. Also available in: Institute for Language, Logic and Information, Preprint LP-90-13, University of Amsterdam, 1990.

[4] Zhisheng Huang and Peter van Emde Boas, Belief Dependence, Revision and Persistence, in P. Dekker and M. Stokhof (eds.), *Proceedings of the Eight Amsterdam Colloquium*, Institute for Language, Logic and Computation, University of Amsterdam, 1992, pp. 271–281; Preprint ITLI LP-91-06.

[5] Zhisheng Huang and Peter van Emde Boas, Schoenmakers Paradox: Its Solution in a Belief Dependence Framework, Institute for Language, Logic and Computation, University of Amsterdam, Preprint ITLI LP-91-05.

[6] Zhisheng Huang, *Logics for Agents with Bounded Rationality,* Ph. D. thesis, University of Amsterdam, 1994.

[7] Nebel, B., *Reasoning and Revision in Hybrid Representation Systems*, Lecture Notes in Computer Science 422, 1990.

[8] Potts, G., John, M., and Kirson, D., Incorporating New Information into Existing World Knowledge, *Cognitive Psychology* 21,(1989), 303-333.

[9] W.J. Schoenmakers, A Problem in Knowledge Acquisition, SIGART Newsletter. 95(1986), 56-57.