International Workshop on Belief Revision
Trento, Italy
May 30 - June 1, 1998

Bayesian Cognitive Diagnosis in Believable Multiagent Systems

Fiorella de Rosis and
Emanuele Covino
Departement of Informatics
University of Bari
{derosis,covino}@di.uniba.it

Rino Falcone and
Cristano Castelfranchi
Institute of Psychology
National Research Council, Rome
{cris,falcone}@pscs2.irmkant.rm.cnr.it

Abstract

This paper considers how beliefs can be ascribed in socially intelligent multiagent systems by observing the agents' behaviour, to build an image of their mental state. When agents are 'believable', the actions they perform are a function of their mental pattern and their personality traits. These traits involve different behaviours, and therefore different orderings of beliefs to revise after each interaction step, to keep the mental state image consistent. We examine how a probabilistic ordering of beliefs can be assigned in the hypothesis of a 'closed' multiagent world, and how belief revision can be treated in a bayesian way. We show, in particular, that gradual updating of beliefs enables getting hints for a suspect of deceipt and suggestions on when and how to deceive.

1 Introduction

A rational agent takes decisions about how to behave in a given circumstance by reasoning on its own mental state, that is on first order beliefs and goals. When it is programmed so as to show a 'socially intelligent' behaviour, it will consider, in its reasoning, also the mental state of the agents with which it interacts (second and higher order beliefs). Second order beliefs may be acquired through 'self introduction' of agents or may be inferred by observing their behaviour. In absence of any information (for instance, as its entry in the world), the observing agent may start from a default and possibly incomplete model of the other agents: this model will have to be revised as far as the interaction proceeds, through a reasoning process that has been named *belief ascription* or *cognitive diagnosis* [14]. The following problems characterize this belief revision process:

- the behaviour of the observed agent may be 'ambiguos', and subject to different interpretations. In this case, the most likely hypothesis has to be selected, by leaving aside the alternatives for future revisions;
- this behaviour may be 'contradictory', for different reasons: because the observed agent may change of mind, because it may lie or just because a previous interpretation was incorrect.

In both cases, a more general knowledge of the observed agent (like the stereotype to which it belongs) may help in selecting the interpretation and in revising it; on the other side, observing the agent behaviour may help in defining that stereotype. To clarify our idea, let us consider the scenario that Nebel emploies to introduce his revision algorithm in [10]:

> "Assume that a suspect tells you that he went to the beach for swimming, and assume that you have observed that the sun was shining. Further, you firmly believe that going to the beach for swimming when the sun is shining implies a sun tan. If you discover that the suspect is not tanned, there is an inconsistency to resolve".

Nebel resolves the inconsistency by assigning a low priority, in the belief ordering, to the suspect's assertion that he 'went to the beach', and by concluding that he was a liar. However, in more realistic multiagent interactions, this (reasonable) explanation should be contrasted with some alternatives: (i) the 'suspect' might be someone who does not tan easily (in that case, the rule should be retracted); (ii) the sun might be strong for someone coming from Germany, but not enough for someone from southern Italy (in this case, the 'shining sun' belief shold be retracted), ... and so on. Every hypothesis about the context (let say, about the stereotype to which the observed agent may be assigned) involves a different ordering of beliefs, and therefore implies a different selection of the beliefs to retract. A reasonable belief revision should keep all these alternatives open, even if it will select one of them to take an immediate decision: knowing, in a future step of interaction, that the observed agent was coming fron Bari will increase the likelihood of the second explanation. At the same time, coming to know that he is not a liar will increase the likelihood that he is someone who resist to sun, ... and so on.
In this paper, we examine the stereotype-based belief ascription and revision in multiagent systems. We consider particular types of stereotypes, that we call 'personalities', which play a relavant role in believable agent systems [12]. We show some examples from an agent factory (that we called GOLEM), which was designed to test multiagent cooperation theories: in this factory, two personality-rich rational agents interact in a blocks world to try to achieve their own goals by exploiting, when possible, the other agent's help or by helping it, if required [2]. We finally approach two topics which are of interest for multiagent interaction: the simulation of *deceiving* and of *suspecting deception* and the *open world* issue.

2 The problem

Our rational agents have a mental state which is the conjunction of belief atoms; belief atoms take, as usual, an agent name as a first term and apply to a sentence in a language $\mathcal{L}$. In what follows, without further notice, low-case latin (resp. greek) letters denote costants (variables) in $\mathcal{L}$. In particular:

- A and B denote agents' names;
- α denotes a domain action variable;
- a, a^1, a^2, ..., denote domain action costants;
- ϕ^k denote variables over sentences in $\mathcal{L}$: the variable α occurs in these sentences;
- f^h denote ground sentences in $\mathcal{L}$: actions a^1, a^2, ..., a^j occur in these sentences.

Example:
f^1: $Cnd\text{-}A(a^1)$, for 'A is able to make the action a^1'
f^2: $Cnd\text{-}B(a^2)$, for 'B is able to make the action a^2'
ϕ^1: $\neg Cnd\text{-}A(\alpha)$, for '$A$ is not able to make the action α'
ϕ^2: $Cnd\text{-}B(\alpha)$, for 'B is able to make the action α'.

Let A and B be two *rational agents* living in a *world* W, with A observing the behaviour of B. In what follows, we formally define the *cognitive diagnosis problem*, for A, about B. This problem deals with ascribing to B a set of beliefs which are consistent with A's previous beliefs about B and with its behaviour, and in revising them as far as interaction between A and B proceeds. As we formalize the point of view of A, in all the notations that follow belief formulae should be intended as having a 'B_A' as a prefix, and a 'A believes that' should be premised to their interpretation.

Definition 1 Given $B_B f^h$, a *basic belief* of agent B about f^h, we call *mental state* of B the conjunction of B's basic beliefs

$$MS(B) = \bigwedge_{h=1}^{m} B_B f^h$$

Example: $MS(B) = B_B Cnd\text{-}A(a^1) \wedge B_B \neg Cnd\text{-}A(a^2) \wedge B_B Cnd\text{-}B(a^1) \wedge B_B Cnd\text{-}B(a^2)$ stands for "B believes that A is able to make a^1 only, whereas he is able to make a^2, as well".

Definition 2 Given $B_B\phi^k$, a *generic belief* of agent B about ϕ^k, we call *possible mental pattern* of B a conjunction of consistent B's generic beliefs

$$MP^r(B, \Phi) = \bigwedge_{k=1}^{n} B_B\phi^k, \text{ with } r = 1 \ldots w$$

Example: $MP^1(B, \Phi) = B_B Cnd\text{-}A(\alpha) \wedge B_B \neg Cnd\text{-}B(\alpha)$.

Notice that a possible mental pattern of B differs from a description of B's mental state in that it sets up a condition on the combination of values of a subset of B's beliefs, and that it contains variable names in belief terms.

Definition 3 Given $Perf^i(B,\alpha)$ a *generic performative* that B may make, concerning α, where $Perf^i$ is an element of a finite set of performative types, which includes communicative acts or domain transformations, $(i = 1 \ldots p)$, we call *behaviour schema* of B a relationship between a possible mental pattern of B and a generic performative that B may make when its mental state matches that mental pattern:

$$BS^{ri}(B,\Phi) = [MP^r(B,\Phi) \overset{?}{\Longrightarrow} Perf^i(B,\alpha)], \text{ where } \alpha \text{ occurs in } \Phi.$$

Example: $Perf^1(B,\alpha) = \textit{Request-to-do}(B,\alpha)$
$BS^{11}(B,\Phi) = [B_B \textit{Cnd-A}(\alpha) \wedge B_B \neg \textit{Cnd-B}(\alpha) \overset{?}{\Longrightarrow} \textit{Request-to-do}(B,\alpha)]$
which should be interpreted as: " if B believes that A is able to make the domain action α whereas it is not, it might request A to make it". Notice that the $\overset{?}{\Longrightarrow}$ symbol denotes that this relationship may be uncertain.

Definition 4 We call *world model* a finite set of behaviours schemas that represents a theory about the performatives that B might 'reasonably' make in correspondence with some given mental patterns. The meaning of 'reasonable' depends on the particular application domain, and on agent's A vision of the world.

$$WM(B) = \{BS^{11}(B,\Phi), BS^{12}(B,\Phi), \ldots, BS^{ri}(B,\Phi), \ldots, BS^{wp}(B,\Phi)\}.$$

Definition 5 Given the *observation*, from A, of a *specific performative* $Perf^j(B,a)$ made by B and concerning a, we call *cognitive diagnosis*, for A, the problem of updating B's mental state from $MS(B)$ to $MS^R(B)$ so that

$$WM(B) \cup MS^R(B) \models Perf^j(B,a).$$

This requires selecting the behaviour schema that B likely applied in deciding to make the performative, and enables revising B's mental state so as to insure consistency in this state.

3 Believable agents

"Personality" can be defined as a typical behaviour of a given agent (or a given class of agents) in a given situation (or a given class of situations). As we deal with cognitive agents, we call *personality trait* $T^s(B)$ an aspect of an agent's B mental state which influences its behaviour in a 'sufficiently permanent' way, by setting how combinations of values for its beliefs affect its actions. For instance, we defined, in GOLEM, a theory to set up how multiagent cooperation is influenced by the agents' individual and mutually believed abilities [2]; in this theory, we defined personality traits affecting the delegation behaviour:

- a *lazy* agent always delegate tasks if there is another agent which is able to take care of them; it acts by itself only when there is no alternative;
- an *hanger-on* never acts by itself: it either delegates actions to others or 'waits' until some change in the external situation occurs;
- a *delegating-if-needed* asks for help only if it is not able to do the task by itself;
- a *never-delegating* considers that tasks should only be achieved if it can perform them.

A 'believable' agent [12] holds one or more 'personality traits' in addition to a set of beliefs. Cognitive diagnosis in a believable multiagent world consists in including personality factors in the abduction process, at least in two ways:

a. *revising B's mental state*, given a performative made by B and a hypothesis about B's personality traits;

b. *revising B's personality traits*, given a performative made by B and a hypothesis about B's mental state.

An example of relationship among behaviour schemas and personality traits is shown in **Figure 1**, where behaviour schemas encode the informal description of personality traits that we made in the previous paraghaph.

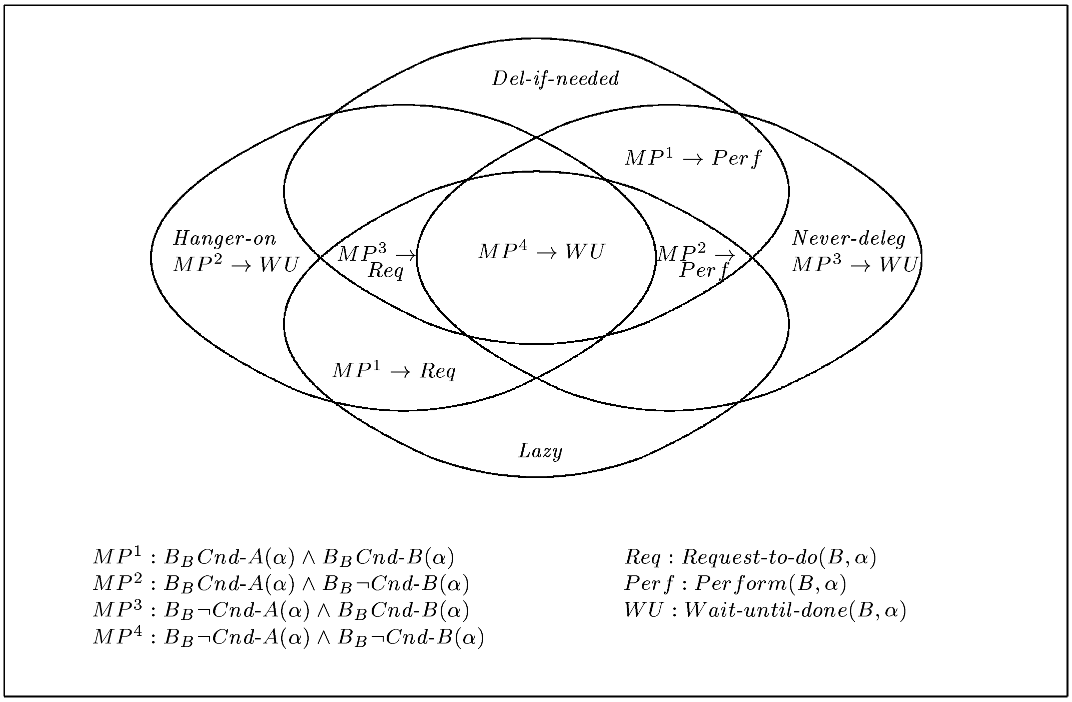

Figure 1: personality-based behaviour schemas in GOLEM

For instance: $MP^1 \rightarrow Req$ stands for $B_B Cnd\text{-}A(\alpha)\ \wedge Lazy(B) \stackrel{?}{\Longrightarrow} Request\text{-}to\text{-}do(B,\alpha)$ which corresponds to the description "a lazy agent always delegates tasks if there is another agent which is able to take care of them". This relationship is such that:

- there are several mental patterns $MP^r(B,\Phi)$ that may lead an agent to make a given performative $Perf^i(B,\alpha)$, each of them corresponding to a distinct personality trait. For example: in Figure 1, both $MP^1(B,\Phi)$ and $MP^2(B,\Phi)$ lead to $Perform(B,\alpha)$ respectively in agents with the personality trait $T^1(B) = Delegating\text{-}if\text{-}needed(B)$ and in $T^2(B) = Lazy(B)$;

- a pattern $MP^r(B,\Phi)$ may lead agents with different personality traits to make different performatives. Again, in Figure 1, MP^1 leads agents with the traits *Lazy* (or *Hanger-on*) and *Delegating-if-needed* (or *Never-delegating*) to make, respectively, the performatives: $Request\text{-}to\text{-}do(B,\alpha)$ and $Perform(B,\alpha)$.

Personality traits thus define a clustering of behaviour schemas in MW, each cluster corresponding to a specific personality trait; this clustering is such that:

- several schemas are defined in correspondence of a given personality trait; for instance, for $T^2(B) = Lazy(B)$:
  $MP^1(B,\Phi) \stackrel{?}{\Longrightarrow} Request\text{-}to\text{-}do(B,\alpha)$
  $MP^2(B,\Phi) \stackrel{?}{\Longrightarrow} Perform(B,\alpha)$
  $MP^3(B,\Phi) \stackrel{?}{\Longrightarrow} Request\text{-}to\text{-}do(B,\alpha)$
  $MP^4(B,\Phi) \stackrel{?}{\Longrightarrow} Wait\text{-}until\text{-}done(B,\alpha)$

- a schema can be associated to several personality traits; for instance,
  $MP^4(B,\Phi) \stackrel{?}{\Longrightarrow} Wait\text{-}until\text{-}done(B,\alpha)$ is associated with all the traits.

4 An example

Let us imagine that B makes the performative: $Request\text{-}to\text{-}do(B,a^1)$. Let us suppose that A doesn't know anything about B's personality, whereas he is certain that she is able to make the domain action a^1, for instance because he saw her doing that action. He therefore 'knows' that B's mental state matches the pattern MP^1 and that she applied the behaviour schema $MP^1 \rightarrow Req$. He asks to himself whether he can infer something about B's personality, from such a communicative act. He may exclude that B is a *never-delegating* or a *delegating-if-needed* (as she is able to make the action). She is, therefore, a *lazy* or an *hanger-on*. A will select one of these hypotheses according to its plausibility and, if the plausibility is the same, according to some different criteria; for instance, his own personality (A may be *trustful* or *suspicious*). He will update his image of B's mental state and will adapt his helping behaviour to this ascription of beliefs. After future B's actions, this image will be revised consistently: knowing something about B's personality traits will now enable A to prune out the research space about B's models. However, as shown in Figure 1, purely logical reasoning enables a partially efficient pruning. In particular, if B makes several performative of the type *Request-to-do* on different turns, A will never have a sufficient evidence to select between the two alternative hypotheses of B being a *lazy* or an *hanger-on*, as the behaviour schema $MP^1 \rightarrow Req$ is common between the two personality traits. In addition, if A is not certain about B's mental state (for instance, about her abilities), he will have a limited possibility of revising his second-order beliefs about B: for instance, only when she performs an action, he will know something more about her abilities.

5 Probabilistic abduction and belief revision

In what follows, we redefine the previous problem in probabilistic terms.

Definition 6 The *r-th believable-mental pattern* of B, denoted with $B\text{-}MP^r(B, \Phi)$, is a consistent combination of beliefs and personality traits.

Example: in our world of cooperating agents, an example of believable mental pattern is $B\text{-}MP^r(B, \Phi) = Lazy(B) \wedge B_B Cnd\text{-}A(\alpha) \wedge B_B Cnd\text{-}B(\alpha)$.

Definition 7 The *general knowledge* of A about the world MW is formalized in probabilistic terms into the table $\mathcal{P}(Perf^i(B, \alpha)|B\text{-}MP^r(B, \Phi))$, where each row is the probabilistic translation of some of the behaviour schemas BS^{ri}, with the specification of the personality traits to which these schemas apply.

Example: by going on with the previous example, we might have
$\mathcal{P}(Request\text{-}to\text{-}do(B, \alpha)|B\text{-}MP^1(B, \Phi)) = .8$
$\mathcal{P}(Perform(B, \alpha)|B\text{-}MP^1(B, \Phi)) = .2$
$\mathcal{P}(Wait\text{-}Until(B, \alpha)|B\text{-}MP^1(B, \Phi)) = 0.$

Definition 9 Given (i) a specific performative (that is, a communicative or domain action) made by B in a particular phase of interaction, denoted with $Perf^i(B, a)$; (ii) a probabilistic knowledge, by A, of B's mental state, which assigns a probability value to the individual components of B's mental state (beliefs and personality traits), denoted with $MS(B)$;
we define *probabilistic cognitive diagnosis* the problem of revising A's knowledge of B's mental state regarding her personality traits and her beliefs that concern the performed action a, by computing the probability values:

$$\begin{array}{ll} \mathcal{P}(T^s(B)|Perf^i(B, a)) & \text{with } j = 1, \ldots, q \\ \mathcal{P}(B_B f^h|Perf^i(B, a)) & \text{with } h = 1, \ldots, m \end{array}$$

We introduce the following simplifying assumptions:

HY1 *B's conformism:*
the set of behaviour schemas in MW is closed, and the relationship among elements of MW and personality traits $\mathrm{T}^S(B)$ is fixed;

HY2 *A's omniscience:*

A knows all behaviour schemas in MW with their relationships with the personality traits;

HY3 *independence of attitudes:*

the probability of a given personality trait $T^s(B)$ for B is not influenced by her beliefs; B's beliefs are conditionally independent among themselfs, as well:

$\mathcal{P}(T^s(B)|B_B\phi^1, B_B\phi^2, \ldots, B_B\phi^n) = P(T^s(B))$

$\mathcal{P}(B_B\phi^k|B_B\phi^1, \ldots, B_B\phi^{k-1}, \ldots, B_B\phi^{k+1}, \ldots) = \mathcal{P}(B_B\phi^k), \forall k = 1, \ldots, n$

Under these hypotheses, A's knowledge of B's mental state can be revised by computing the probability that B's mental state matches $B\text{-}MP^r(B,\Phi)$ after observing the evidence $Perf^i(B,a)$:

$$(1) \qquad \mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]}|Perf^i(B,a)) = \mathcal{P}(Perf^i(B,a)|B\text{-}MP^r(B,\Phi)^{[\alpha/a]}) \times \mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]})/\mathcal{P}(Perf^i(B,a))$$

In this formula:

a. $\mathcal{P}(Perf^i(B,a)|B\text{-}MP^r(B,\Phi)^{[\alpha/a]})$ is taken from the table $\mathcal{P}(Perf^i(B,\alpha)|B\text{-}MP^r(B,\Phi))$, which formalizes A's general knowledge about the world MW; we assume that the probability that an agent with a given believable pattern $B\text{-}MP^r(B,\Phi)$ makes the performative $Perf^i(B,\alpha)$ is independent of the value assigned to the variable α. For example, the definitions of *lazy*, *hanger-on*, *delegating-if-needed*, *never-delegating* agent in Section 3 may be translated into a conditional probability table that establishes which is the probability of making a *Request-to-do* (or a *Perform* or a *Wait-until-done*) conditional on a combination of the agent's beliefs and independently of the specific domain action mentioned in the schema.

If, in particular, a combination of beliefs is 'binding', for an agent with a given personality trait, to perform a particular action in a specific context, the mentioned table will be binary; this is the case if, for instance, a *lazy* B that is not able to do the domain-action α and that believes that A is able to do it will *certainly* perform the action *Request-to-do*(B,α). If, on the contrary, an agent with a given personality trait may make, in a given circumstance, different performatives, each with a given probability value, that table is not binary (like in the example for BS^{1i} at the beginning of this Section).

b. the probability of $B\text{-}MP^r(B,\Phi)^{[\alpha/a]}$ is computed from the probabilities of individual components of B's mental state (under assumption HY3) after applying the unification $[\alpha/a]$:

$$(2) \qquad \mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]}) = \prod_{hs} \mathcal{P}(B_B f^h) \times \mathcal{P}(T^s(B))$$

In absence of any information, and by virtue of B's conformism and A's omniscience assumptions, A will assign a uniform probability distribution to these values.

c. the probability distribution of actions is the following:

(3) $\mathcal{P}(Perf^i(B,a)) =$
$\sum_r \mathcal{P}(Perf^i(B,a)|B\text{-}MP^r(B,\Phi)^{[\alpha/a]}) \times \mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]})$

The posterior probabilities of the individual components of A's model of B' attitudes and personality traits are therefore a function of his general knowledge MW and of his (prior) knowledge about these attitudes and personality traits, *when he observes the performative* $Perf^i(B,a)$; they may be computed from the table of $\mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]} \wedge Perf^i(B,a))$ by summing up over the $B\text{-}MP^r(B,\Phi)^{[\alpha/a]}$ that include $B_B f^h$ or $T^s(B)$, respectively.

(4) $$\mathcal{P}(B_B f^h|Perf^i(B,a)) = \sum_r \mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]}|Perf^j(B,a))$$

over all r such that $B_B f^h \in B\text{-}MP^r(B,\Phi)^{[\alpha/a]}$.

(5) $$\mathcal{P}(T^S(B)|Perf^i(B,a)) = \sum_r \mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]}|Perf^j(B,a))$$

over all r such that $T^s(B) \in B\text{-}MP^r(B,\Phi)^{[\alpha/a]}$.
These probabilities are the new values that A will employ to update in (1) the $\mathcal{P}(B\text{-}MP^r(B,\Phi)^{[\alpha/a]}|Perf^i(B,a^2))$ after B performs the next action a^2 ; if a^2 is different from a, beliefs regarding a will not be revised at the next step. A's knowledge of B's attitudes and personality evolves during interaction as a function of his initial knowledge of these items, of actions performed by B and of the distribution of probability values in the table $\mathcal{P}(Perf^i(B,\alpha)|B\text{-}MP^r(B,\Phi))$.

6 Suspecting and preventing suspect

Personality-based cognitive diagnosis is a kind of stereotypical belief ascription and revision. Like in stereotypical modeling [13], a believable agent inherits the behaviour schemas that are typical of several personality traits. In the example of cooperating agents that we are considering in this paper, B's overall behaviour will result from a consistent integration of a particular *delegation*, a particular *helping* and a particular *reaction* trait. She will therefore conform, in distinct phases of her interaction with A, to the *behaviour schemas* which are attached to these traits. The main difference between stereotypical and personality-based modeling is the *social nature* of the latter. In deciding how to behave, personality-based agents take into consideration the characteristics of agents with which they interact. We already said that a *lazy* (*delegating*) agent does not perform an action that he believes other agents might do; likewise, a *supplier* (*helping*) agent responds positively only to requests of help coming from agents that he presumes would not be able to make the requested action. This mechanism of social relationship among agents results in attaching *values* to personality traits. As we do not represent 'exchange personalities'

(do-ut-des), delegating always implies an advantage, and helping a cost. It is therefore convenient to be helped, even by profiting, if needed, of erroneous or incomplete beliefs of other agents. In this sense, being considered a *delegating-if-needed* is preferable to being considered a *lazy*, because it gives more chance of receiving help. The same applies to the issue of being considered as 'unable to perform the requested action' (in a different context, and with a different scale of values, this might be considered as a negative finding: for instance, in a highly competitive world). Gradual updating of his image of B's mental state results, for A, into attaching different probabilities to attitudes with different 'social values'. This may produce, into A, the onset of a *suspect* towards negative B's characteristics.
In the example in **Figure 2**, we suppose that B is a *Lazy* who is able to make actions a^1 and a^2. We also assume that A initially has no knowledge about B's mental state. He therefore equidistributes probabilities among the four delegating attitudes that B might have (p =.25 for each of them, in particular for B being *Lazy*). A does not know B's abilities either; therefore $\mathcal{P}(B_A B_B Cnd\text{-}B(\alpha)) = .5$, for $\alpha = a^1$ and $\alpha = a^2$. Finally, we may suppose that A does not know either whether B is informed about A's abilities: $\mathcal{P}(B_A B_B Cnd\text{-}A(\alpha)) = .5$, for $\alpha = a^1$ and $\alpha = a^2$. This initial mental state of A is shown in the first block of data, in the first column of **Figure 2**.

A	**B**	
second-order beliefs	**first-order beliefs**	**third-order beliefs**
initial state		
$B_A B_B Cnd\text{-}B(a^1)$ (.5)	$B_B Cnd\text{-}B(a^1)$ (1)	$B_B B_A B_B Cnd\text{-}B(a^1)$ (.5)
$B_A B_B Cnd\text{-}A(a^1)$ (.5)	$B_B Cnd\text{-}A(a^1)$ (1)	$B_B B_A B_B Cnd\text{-}A(a^1)$ (.5)
$B_A B_B Cnd\text{-}B(a^2)$ (.5)	$B_B Cnd\text{-}B(a^2)$ (1)	$B_B B_A B_B Cnd\text{-}B(a^2)$ (.5)
$B_A B_B Cnd\text{-}A(a^2)$ (.5)	$B_B Cnd\text{-}A(a^2)$ (1)	$B_B B_A B_B Cnd\text{-}A(a^2)$ (.5)
$B_A Lazy(B)$ (.25)	$Lazy(B)$ (1)	$B_B B_A Lazy(B)$ (.25)
	$Request\text{-}to\text{-}do(B, a^1)$	
	(passive deception)	
$B_A B_B Cnd\text{-}B(a^1)$ (.4)		$B_B B_A B_B Cnd\text{-}B(a^1)$ (.4)
$B_A B_B Cnd\text{-}A(a^1)$ (1)		$B_B B_A B_B Cnd\text{-}A(a^1)$ (1)
$B_A B_B Cnd\text{-}B(a^2)$ (.5)		$B_B B_A B_B Cnd\text{-}B(a^2)$ (.5)
$B_A B_B Cnd\text{-}A(a^2)$ (.5)		$B_B B_A B_B Cnd\text{-}A(a^2)$ (.5)
$B_A Lazy(B)$ (.4)		$B_B B_A Lazy(B)$ (.4)
	$Request\text{-}to\text{-}do(B, a^2)$	
$B_A B_B Cnd\text{-}B(a^1)$ (.4)		$B_B B_A B_B Cnd\text{-}B(a^1)$ (.4)
$B_A B_B Cnd\text{-}A(a^1)$ (1)		$B_B B_A B_B Cnd\text{-}A(a^1)$ (1)
$B_A B_B Cnd\text{-}B(a^2)$ (.4)		$B_B B_A B_B Cnd\text{-}B(a^2)$ (.4)
$B_A B_B Cnd\text{-}A(a^2)$ (1)		$B_B B_A B_B Cnd\text{-}A(a^2)$ (1)
$B_A Lazy(B)$ (.44)		$B_B B_A Lazy(B)$ (.44)

Figure 2: an example of suspecting (by A) and of preventing suspect through deception (by B).

Let us suppose that B request A to perform the domain-action a^1; A will start 'suspecting' that she is a *lazy* or an *hanger-on* (p=.4: second block of data, again in the first column of the figure). This suspect will increase after a second *Request-to-do*, concerning a different action a^2 (p=.44: third block of data). Belief revision is nonmonotonic, though: if, for instance, after these two requests, B performs herself some domain action a^3 rather than asking help, the probability attached to the *hanger-on* trait will go to zero. We can therefore assume that the arising of a suspect, in A, about negative aspects of B's mental state can be modelled by looking at the variation of probabilities attached to A's second order beliefs that concerns these B's aspects.

Cognitive diagnosis can be applied to several levels of belief nesting. Let us consider, for instance, the reasoning that B performs concurrently with A's reasoning. B may reproduce, in her third order beliefs $B_B B_A B_B f^h$, the abductive reasoning that A is performing on his image of B's mental state, (uncertainty values attached to B's third order beliefs are shown in the third column of the figure). By doing so, B will *anticipate* the consequences of her next action on A's mental state. In this way (to go on with our example), B can try to get advantage of the hypotheses that she imagines A will make about her mental state, to increase the chance of getting help. She adapt her *deceiving strategy* to what she presume A is thinking about her (for an analysis of the possible forms of deception, see [3]). For example, she can make a 'passive deception' at the beginning, when she believes that A does not yet suspect that she is a *lazy*: she will just ask A to help her without revealing her personality and her abilities, by leaving A in his ignorance. She can shift to an 'active deception' as soon as she imagines that this suspect becomes too high (in our example, at the second *Request-to-do*). For instance she can lie about her abilities, by saying that she would not be able to make the requested action ($Inform\text{-}not\text{-}Cnd(B, a^2)$. In this way, if A believes her, the probability values attached to the third blick of data, in columns 1 and 3, will change as follows:

$\mathcal{P}(B_A B_B Cnd\text{-}B(a^2)) = (0)$ and alike for B's third order beliefs, while the probabilities of personality traits will remain unchanged. The possibility, for B, of getting help will be the same as in first turn.

Treatment of suspect is a natural extension of *keyhole* plan recognition [1], in which information available to the recognizing agent is gleaned from incomplete observation rather than being conveyed intensionally by a user or an agent. It is, in our opinion, an inherently uncertain kind of reasoning, which can be treated profitably in probabilistic terms. In another paper, we consider several examples of deception and suspect in the medical domain, by showing that they need attaching, to beliefs, *degrees of strength* of several types: uncertainty is one of them, and probably the most important, but it is not the only one. Other measures of strength, more similar to Galliers' "hardness of revision" [6], are needed as well, to formalize this topic entirely. For instance, one might assign more informational value to A's beliefs about B's personality traits because they support a number of related beliefs, and one might imagine that A prefers revising his beliefs about B's abilities rather than those about her personality. In this case, reordering of A's beliefs about B after every observation will require combining appropriately the two types of measures.

7 Towards an 'open world'

The A's omniscience and B's conformism assumptions that we introduced in Section 5 are particularly hard: the multiagent world is open in several senses, and agents should be able to cope with a situation in which unknown behaviours come up. In the formalization of the problem that we propose in this paper, this means that the *world model* MW should be opened, in several terms:

a. the list of actions $Perf^i(B, \alpha)$ that agents might perform should be open;

b. the list of beliefs $B_B\phi^h$ affecting the choice of an action in the behaviour schema should be open;

c. the list of personality traits $T^s(B)$ should be open;

d. the list of behaviour schemas BS^{ri} should be open

Reasoning agents should then be aware of the fact that their theory may be incomplete; they should be able, as well, to revise it when some other agent behaves unexpectedly. 'Unexpected behaviour' means, in this case, performing unknown actions or behaving inconsistently with the reasoning agent's theory. We are working, at present, on two types of theory revision: refinement of knowledge about personality-based behaviour schemas and updating of expectations about performable actions. In the first case, we assume that A's knowledge of behaviour schemas is incomplete, as he is aware only of a subset of the attitudes which do affect an agent's behaviour; updating of behaviour schemas then results into a refinement of the personality traits description. In the second case, it is A's knowledge of performable actions which is incomplete: in face of a new action, A has to activate a procedure which aims at 'building' a new behaviour schema, by filling its slots with B's attitudes that he came to know so far.

8 Related works

Though we formulated the problem by considering interaction between two agents A and B, our method can be extended to the case of several interacting agents. In this case, A will hold a unique model of the world MW, that he will apply to revise his image of the mental states of several agents B^1, B^2, ..., B^s, ..., after observing action $Perf^i(B^s, f^h)$ by agent B^s.
Research on belief ascription and revision in multiagent systems has focused on plan recognition rather than on cognitive diagnosis. Plans may be represented as recipes for achieving certain states of the world or as complex mental attitudes; they concern, in both case, the domain actions that have to be performed to reach a given domain goal state. Interest on reasoning about the motivations behind a 'message'

in multiagent systems has been limited, probably due to the fact that agents are not programmed to cooperate with *other* agents (to request help or to provide help) according to an analysis of *their* beliefs, goals and capabilities: consequently, the interest towards keeping and updating a model of other agents' attitudes is limited. The main contributions to cognitive diagnosis are due to those who work on conversational systems. In this case, abduction on the interlocutor is made at several levels: for instance, in [5] the domain level, the problem-solving level, the belief level and the discourse level.

In both domains, logics is employed in most of the belief ascription and revision methods which have been proposed so far (for instance, [9], [11]). In fewer cases, the problem has been formulated in probabilistic terms or by Dempster and Shaefer theory [4], [8], [1] while Rich employed an ad hoc theory to treat uncertainty in stereotype-based user modeling in Grundy [13]. The research with which this paper has more similarities is [7], which propose a bayesian updating of beliefs about agents, based on their observed behaviour. These authors maintain an influence diagram that represents belief models at several nestings, and update it incrementally by a recursive, dynamic programming procedure. Their application domain is coordination in air defense and pursuit, and their method does not yet include the possibility of building or triggering new hypothetical models.

9 Acknowledgements

We wish to acknowledge Antonella Miceli for doing the simulation programs (in Java) that we employed in this research, in the scope of her dissertation in Computer Science.

References

[1] D. W. Albrecht, I. Zuckerman, A. N. Nicholson, Bayesian models for keyhole plan recognition in an adventure game, in press on *User Modeling and User Adapted Interaction.*

[2] C. Castelfranchi, F. de Rosis, R. Falcone and S. Pizzutilo, Personality traits and social attitudes in multiagent cooperation, *Applied Artificial Intelligence*, Special Issue on *Socially Intelliget Agents*, in press.

[3] C. Castelfranchi, F. de Rosis and F. Grasso Deception and suspect in medical interaction: towards a simulation of believable dialogues, in *Human-Computer Conversation*, M. Lee and Y. Wilks (Eds), Kluwer Publ. Co. in press.

[4] E. Charniack and R. Goldman, A probabilistic method of plan recognition.

[5] J. Chu. Carrol and s. Carberry, A plan-based model for response generation in collaborative task-oriented dialogues, in *Proceedings of AAAI94.*

[6] J. R. Galliers, Autonomous belief revision and communication, in *Belief Revision*, P. Gardenfors (Ed), Cambridge University Press, 1992.

[7] P. J. Gmytrasiewicz, Bayesian update of recursive agent models, in press on *User Modeling and User-Adapted Interaction.*

[8] M. J. Huber and E. H. Durfee, An initial assessment of plan-recognition-based coordination for multi-agent teams, *ICMAS 96.*

[9] B. Mahleiro, N. R. Jennings and E. Oliveira, Belief revision in multi-agent systems, *ECAI 94.*

[10] B. Nebel, Syntax-based approaches to belief revision, in in *Belief Revision*, P. Gardenfors (Ed), Cambridge University Press, 1992.

[11] A unified view of plans as recipes, *Technical Report of AIIA, 1997.*

[12] W. S. Reilly and J. Bates, Natural negotiation for believable agents, School of Computer *CMU-CS-95-16, 1995.*

[13] E. Rich: Stereotypes and user modeling, in *User Models in Dialog Systems*, A. Kobsa and W. Wahlster (eds), Springer Verlag, 1989.

[14] J. Self, Model-based cognitive diagnosis, in *User Modeling and User Adapted Interaction*, 3,1,1993.