Computational Intelligence, in press

Can Computers Deliberately Deceive?

A Simulation Tool and its Application to Turing's Imitation Game

Fiorella de Rosis[1], *Cristiano Castelfranchi*[2], *Valeria Carofiglio* [1]*and Giuseppe Grassano*[1]
[1]Department of Informatics, University of Bari
Via Orabona 4, 70126 Bari, Italy
+39 080 5443282, derosis, carofiglio}@di.uniba.it

[2]Institute of Science and Technology of Cognition, CNR
and University of Siena
castel@www.ip.rm.cnr.it

Abstract

In this paper, we describe how agents can deceive within a probabilistic framework for representing their mental state: in doing so, we challenge the so called 'sincerity assumption' in HCI and MAS. We distinguish 'deception' from its special case of 'lie' and characterise different forms of deception, by identifying several criteria for distinguishing among them. In particular, we propose a model of information impact on the Receiver's mind. Since the message Sender must plan its strategy by considering the Receiver's criteria for believing, we also discuss some of these criteria, like content plausibility, source informativity and information safety. We apply this model to a simplified version of Turing's Imitation Game and describe how we implemented a Simulator of deceptive strategies that we called 'Mouth of Truth'. We conclude the paper by describing an evaluation study that enabled us to verify the validity of our method and to revise it in part.

Keywords: deception modeling, belief networks, belief ascription, uncertain reasoning.

1. Introduction

In Human-Computer interaction (HCI) and multiagent systems (MAS), users and agents are presumed to always say the truth: the 'sincerity assertion' is a baseline assumption in the great majority of dialog systems: see, for instance, (Carberry and Lambert, 1999; Paek and Horvitz, 2000; Zukerman et al, 2000). This hypothesis is usually justified by claiming that HCI and MAS are based on the principle of cooperation and that cooperation would imply sincerity in all performances. We argue that this is not always the case and that, on the contrary, in HCI, computer-supported cooperation and organisation, computer mediated commerce, intelligent database management etc, deception occurs not only occasionally and unintentionally, but also on purpose, exactly as it happens in human communication (Buller et al, 1990; Burgoon et al, 1996; Castelfranchi and Poggi, 1998; Castelfranchi et al, 1999 and 2001; Kashy and DePaulo, 1996; Lelouche and Doublait, 1992). "Intelligent deceiving agents" are already a reality of the agent-based paradigm (Zlotkin and Rosenshein, 1991): these agents may deceive because of 'malicious' motives or because they work for 'malicious' owners; but they may do it for good reasons and in the interest of others, as well. For example:

- *information systems* have to misinform an unauthorized user (be it a human or a software agent) in order to protect confidential information; this is a well known problem for databases where 'multi-level security' requires deliberately wrong answers and cover stories (Demolombe and Jones, 1996);
- *in electronic commerce*, we may delegate agents to bid for us in a self-interested perspective or to sell something for us in some online auction; in both cases, we want them to bargain for us, and bargaining is always a bit deceptive. On the other hand, agents could be of help in empowering the consumers and reduce the handicap due to asymmetric information in the market; to this aim, they must be able to discover various forms of deception and contrast them;
- *our personal assistants* might need to deceive us thereby trying to influence us to do the right thing, to protect our interests against our own short term preferences or biases and so on. This is already evident in some naturally occurring dialogs (especially in the medical domain) in which either the doctor or the patient or both of them frequently do not tell the truth or 'all' the truth: that is, they lie or are 'reticent' (Berry et al 1998; Brody, 1982; Castelfranchi et al, 1999; de Rosis et al, 1999). In the near future, recommendations to the user concerning appropriate behaviours will be provided by software agents in a number of situations; it is therefore reasonable to envisage that these agents will be endowed, on the one hand with the ability to deceive and, on the other, with the ability to discover deception in the user (or in other agents) and to react appropriately (Firozabadi and Jones, 1998).

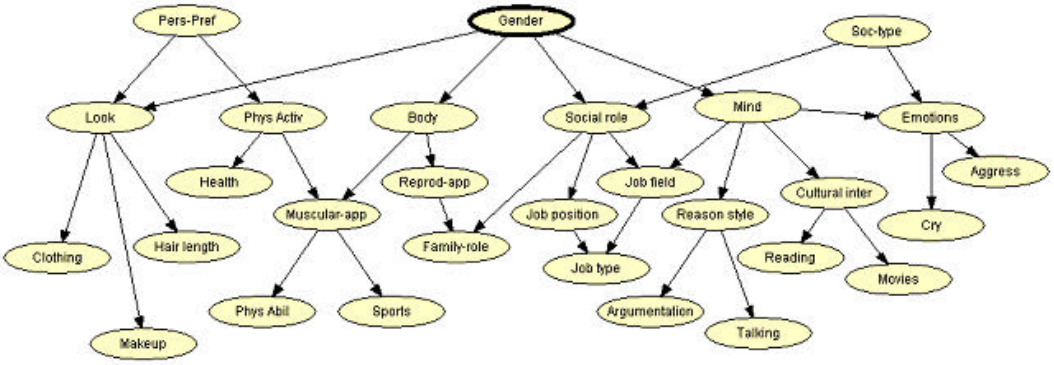
Our notion of deception covers, as we will see, many more cases than deliberately lying. The hypothesis of sincere assertion consists in assuming that all communicated facts are believed by the Sender and that relevant facts in the given context are not hidden. In this paper, we examine *how the decision to deceive can be formalised*. We discuss, in particular, the role of uncertainty in this form of reasoning and propose representing the mental state of the two interlocutors with the formalism of belief networks (Spiegelhalter,1986; Neapolitan,1990; Pearl, 2000).

2. The 'Imitation Game'

Examples in this paper are taken from a scenario that is drawn from Turing's "*Imitation Game*" (Turing, 1950). Three people are involved in Turing's description of this game: a woman (A), a man (B) and an interrogator (C). C

is closed in a room and communicates with A and B by some electronic device; C's goal is to guess their gender: C knows the two participants as 'X' and 'Y' and, at the end of the game, will say: "X is A and Y is B" or the contrary. Both A and B have to answer C's questions. The goal of A is to induce C to incorrectly identify her gender, while the goal of B is to help C to reach the truth. The question Turing raises is the following: "*What will happen when/if a computer will take the place of A? Would the interrogator C give a wrong answer as frequently as in a game involving only humans?*". He claims that this question is equivalent to the very basic one: "*Can machines think?*".

Although Turing does not explicitly mention deception in his paper, the idea that intelligent behavior also involves the ability to deceive is, in our view, a proof of the interest of modeling deception and suspicion in information exchanges. Like several intuitions in that paper, the situation imagined by Turing tends to be frequent in human-computer interaction. In chat lines and games, for instance, users tend to simulate a different gender (from their own). In receiving answers to some questions on the web, we do not know and cannot easily understand whether this answer comes from a person, a data base manager or an intelligent personal assistant. So, a weak form of the Turing Test is already a common experience on the net, although agents or dialogue systems employed for simulating it are not very sophisticated or intelligent. What is amazing is that many uses do not care to understand whether the answer received is produced by a human operator or a machine: there is not much difference between the two alternatives, also considering that both might try to deceive us and that we have to decide whether to trust them or not. For these reasons, we claim that the study of deception is a worthwhile topic for AI and represents a return to its original and deepest challenge.



Legend: variables associated with nodes in the belief network (with their values)

Aggress	*tendency to aggressiveness (high/moderate/low)*
Argumentation	*argumentation style (by appeal to emotion/with logical arguments)*
Body	*body characteristics (female/male traits)*
Clothing	*clothes worn (skirt/trousers/both)*
Cry	*tendency to cry (frequently/unfrequently)*
Cultural inter	*cultural interests (typical of own gender/not clearly defined)*
Emotions	*tendency to show emotions (frequently/unfrequently)*
Family-role	*role within family (cares for children/gets money to family/does both)*
Gender	*gender (man/woman)*
Hair	*hair length (long/medium/short)*
Health	*health condition (good/medium/bad)*
Job field	*job field (industry/agriculture/public administration)*
Jop position	*job position (subordinate/autonomous/leading)*
Job type	*job (clerk/farmer/manager/sales person/free lance)*
Look	*look (typical of women/not clearly defined/typical of men)*
Makeup	*uses cosmetics(yes/no)*
Mind	*mind (typical of own gender/not cleary defined)*
Movies	*preferred movies (war/love stories/comics)*
Muscular-app	*muscular apparatus (well-developed/moderately-developed)*
Pers-Pref	*personal preferences (typical of own gender/ambiguous)*
Phys Abil	*physical abilities (is agile/can lift heavy weights)*
Phys Activ	*physical activity (high/moderate/none)*
Reading	*reading interests (novels/essays/spy stories)*
Reason-style	*privileged reasoning style (logic/intuition)*
Reprod-app	*reproductive apparatus (can bear offspring/can fecundate)*
Social role	*role in the society (typical of own gender/not clearly defined)*
Soc-type	*type of society (conservative/innovative)*
Sports	*sports practiced (yoga/football/aerobics/Aikido)*
Talking	*preferred talking subjects (private/politics)*

Figure 1: the belief network representing knowledge employed in our game

Rather than simulating the Imitation Game (which would be a difficult undertaking indeed) we consider, in this paper, a simplified scenario in which the System (S) is required to say something related to X's gender, and a Receiver (R) has to discover whether X is a man or a woman. The goal of S is to win the game by inducing R to come to a wrong conclusion. S is enabled to deceive R, to achieve this goal; reasoning models are represented as Belief Networks (BNs). **Figure 1** shows an example of BN that represents the knowledge employed in our simulation. The 'gender of X' is one of the roots of this network. The idea is that look, body, social role and mind of individuals ('intermediate nodes') depend on their gender but also on their personal preferences and the type of society in which they live (other 'roots'). Differences in look, body, social role and mind may manifest themselves in several ways, corresponding to the network's 'leaves': hair length, clothing, use of cosmetics, practiced sports, role played in the family and so on. Uncertainty in the relationship between these variables is represented (as usual in BNs) by prior probabilities associated with 'root' nodes and conditional probability tables associated with the other nodes (Spiegelhalter, 1986).

3. Elements of deception planning

Like any form of persuasion, deception requires a 'plan'. Elements of this plan are the decision of 'whether to deceive', the selection of a 'deception object', the form of deception to apply and (in the majority of cases) a 'deception instrument'; mentioning an information source may also help in increasing the chance of success of the deception. Let us define these concepts and discuss how they may be modeled.

3.1 Deciding whether to deceive

Notation 1. Let:

- S and R denote, respectively, the Sender and the Receiver of the message;
- p_i (i= 1,...,n) and g_k (k= 1,...,m) be boolean variables denoting the first one a 'domain-fact' and the second a 'domain goal-state'. The 'eventually' operator $\Diamond$, applied to g_k, should be interpreted as "the domain goal-state g_k will be achieved at some point in the future" (Cohen and Levesque, 1992).

We will denote S's belief and goal concerning p_i with, respectively, $B_S\ p_i$ and $G_S\ p_i$ and S's ignorance of p_i with $\neg BW_S\ p_i$ [1]. Similar notations will be employed for R. Second-order beliefs and goals will be introduced, for S, as combinations of the defined items: $B_S\ B_R\ p_i$, $G_S\ B_R\ p_i$ etc.

Definition 1.

We call *active beliefs* the subset of beliefs which, in a given context or situation, an agent considers, uses or evaluates for a given decision or a given problem-solving. A belief may be active because it was mentioned by R or S or because it was involved recently in R's reasoning. [2]

---

[1] $\neg BW_S\ p_i \Leftrightarrow \neg B_S\ p_i \wedge \neg B_S\ \neg p_i$ and should be read as "S does not believe whether p_i is true or false".

[2] Active beliefs are a relevant concept in dialog planning and argumentation. Chris Reed distinguishes three epistemic states for beliefs: an agent may be *unaware* of a proposition p or *aware but undecided* about p or *not*

We assume that the mental state of S includes two components: (i) a set of *own beliefs* about the domain and *reasoning rules* which establish the way that beliefs are interrelated and (ii) *a image of the Receiver's mental state*, with similar components. In addition, some of the Receiver's beliefs about the domain may influence some Sender's goal:

$$\mathsf{R1}{:}B_S\,(B_R\,p_i \Rrightarrow (\Diamond g_k)); \quad \mathsf{R2}{:}B_S\,(B_R\,(\neg p_i \Rrightarrow (\Diamond g_k)); \quad \mathsf{R3}{:}B_S\,(\neg BW_R\,p_i \Rrightarrow (\Diamond g_k))$$

These are disjunctive rules: only one of them is true in a given context. They should be interpreted as follows: "S believes that the achievement of goal g_k *is influenced by* R's belief about the fact p_i; it might occur when R believes p_i, when R believes $\neg p_i$ or when R ignores whether p_i is true". Notice that the Receiver's belief about p_i may affect achievement of more than one goal and that this belief may be influenced, in turn, by other beliefs which are 'active' when this choice is made. Rules R1, R2 and R3 enable S to make some prediction about the consequences of a particular aspect of the Receiver's state of mind on the achievement of its own domain goal-state.

In our simplified version of the Imitation Game, the System's domain goal g_k is *to win the game*. This will happen if R comes to a wrong conclusion about X's gender (p_i). Let us now assume that S activates the goal $G_S\,\Diamond g_k$, that is, that S *intends to win the game*. If S is a rational agent, this will imply activation of a goal about R's beliefs which varies according to whether rule R1, R2 or R3 applies: $\mathsf{G1}{:}G_S\,B_R\,p_i$, $\mathsf{G2}{:}G_S\,B_R\,\neg p_i$ or $\mathsf{G3}{:}G_S\,\neg BW_R\,p_i$.
In our game, we assume that X is a woman (as we will do in the rest of this paper), the goal of S may be "*R believes that X is a man*" or "*R ignores X's gender*" [3]. Setting one of these goals will activate a strategy enabling the System to achieve the desired situation, which will result in a *'sincere' or a 'deceptive' communication*, depending on what S believes about p_i .

Notation 2. Let us denote:
the *influence goals* that S may have as follows:
($-p_i$): "*leave* R's belief about p_i *invaried*",
($\rightarrow p_i$): "*strengthen* R's belief about p_i",
($\leftarrow p_i$): "*weaken* R's belief about p_i";
and the *'deceptive' and 'sincere' communication types* with *Dec* and *Sinc*.

Influence goals $\rightarrow p_i$ or $\leftarrow p_i$ may require strengthening or weakening the belief of R about p_i differently, depending on her initial degree of belief. Communication types combine with influence goals to produce a message, according to the following factors:

- S's belief about p_i: $B_S p_i$, $B_S\,\neg p_i$ or $\neg BW_S p_i$;
- S's hypothesis about R's belief: $B_S\,B_R\,p_i$, $B_S\,B_R\,\neg p_i$ or $B_S\,\neg BW_R\,p_i$;
- S's goal about R's belief: G1, G2 or G3.

In particular:

a. the influence goal depends on the difference between S's hypothesis and goal about R's belief;
b. the communication type depends on the difference between S's belief and its goal about R's belief.

---

attending to a belief that is included in its model (Reed, 2000). Our notion of 'not believing whether' ($\neg$BW) is similar to Reed's 'awareness with indecision', while our notion of 'activity' his similar to his 'attending to'.

[3] Notice that G1, G2 and G3 may either be 'achievement goals', when S believes that they are not currently true, or 'maintenance goals', when S believes that they are already true (Cohen and Levesque, 1992).

Therefore, in *Dec(-p_i)* and *Sinc(-p_i)* S avoids doing anything that might change R's belief about p_i [4]; if, in some way, it must send a message (for instance, because it has to answer a question), it selects a subject that has a minimum influence on this belief. In *Dec(→p_i), Sinc(→p_i), Dec(←p_i)* and *Sinc(←p_i),* S has to do something (either sincerely or deceptively) to achieve the desired change.

3.2. Deceiving in silence

Definition 2. We call:
passive deception the deception form in which nothing is done by S to achieve its goal about the Receiver's beliefs and *active deception* the form in which S does something to influence R's beliefs.

A passive deception form may be applied only when the influence goal is (-p_i). Here, we want to emphasize that deception does not necessarily require "doing some action aimed at influencing the Receiver's belief". In particular, doing something to leave R's belief unchanged may already be considered as a form of 'active' deception.

Example 1:
Let us suppose that the goal of S is that R ignores X's gender or has an incorrect belief about it. If (as we assumed in Section 2) S believes that "*X is a woman*" and is convinced that R believes that "*X is a man"*or has no idea about X's gender, it may deceive R either passively, by avoiding to talk about the subject, or actively, by delivering some information that does not alter this belief.

3.3 Selecting a deception 'object'

Definition 3.
We call *deception object* (p_i) the fact whose belief by R influences achievement of the goal-state of S through rules R1, R2 or R3, as defined in Section 3.1.

In some cases, several domain facts might influence the goal-state; in these cases, S has to select the deception object that best achieves his or her goal.

3.4 Selecting a deception 'instrument'

Once a deception object p_i and a needed degree of change in this belief have been settled, S may decide to *directly* influence R's belief about this fact by focusing its communicative action on p_i, or to *indirectly* influence this belief by focusing its action on another domain fact p_j which is related to p_i.

Definition 4
We call *deception instrument* (p_j) a domain fact that is related to the deception object p_i in R's mental state.

In indirect deception, S tries to induce R to believe in p_i by some form of inference from p_j. The advantage of this deception form is that it permits *influencing the Receiver in the desired direction* by reducing the risk of being discovered in an attempt to deceive. In our example, leaving the Receiver in a state of doubt as long as possible is the essence of the game. S cannot directly say which is X's gender nor can it pretend to ignore it: therefore it argues indirectly about it. The choice of the deception instrument is influenced by the context in which the communication

[4] Notice that 'doing' something does not necessarily mean 'saying' something. S might try to change U's beliefs without sending messages or speech acts: for instance, it might modify the external world perceived by R. However, as in this paper we deal with dialogue, we will focus on linguistic ways of changing the Receiver's beliefs.

occurs: when S has to respond to a (more or less precise) question, this choice is more limited than when it makes a spontaneous communication; it also depends on what S presumes R is thinking in that particular context (her active beliefs). In the general case, the deception instrument p_j may be one of the following items:

- a *premise* of p_i: p_j = Pre(p_i) that is represented in the BN by *a parent-node of p_i*. In this case, S exploits R's deductive reasoning ability;
- a *consequence* of p_i: p_j = Con(p_i) that is represented by *a child-node of p_i*. In this case, S exploits R's abductive reasoning ability;
- *any other information related to p_i*: p_j = Rel(p_i). A notable case is that of *brother nodes of p_i*: by talking about a brother node, S exploits R's ability to combine abductive and deductive reasoning at the same time;
- if S presumes that R might not believe what it says, it may support its statement p_j by citing an *information source* or, better, the fact that a given source said that p_j. In this case, S may exploit R's ability to reason on the relationship between informativity of sources and credibility of facts to deceive her.

When several deception instruments are available, S has to select the most 'convenient' (and we will see in a while which convenience criteria it may use). If a probabilistic approach is applied, the user's beliefs about the *likelihood* of the deception instrument and the *strength of its link* with the deception object influence this choice. Another influencing factor is the *length of the path* from the deception object to the deception instrument in the reasoning model; if the weights associated with the links are similar, the degree of change produced is a function of the length of the path. Direct deception produces the maximum change; indirect deception on father or children nodes comes next; indirect deception on brothers is less effective, and so on. By selecting the most convenient father node, S may *divert the Receiver's attention from a real premise* p_k of fact p_i *towards a fictitious one,* p_j. This may be convenient when the fictitious premise entails a lower risk of being discovered in a deception attempt, even if it has a lower efficacy. It might also be convenient when S presumes that R already believes in p_i and wants to propose an alternative interpretation of this fact to her, which affects the deception object in the desired direction. To this aim, it will weaken the link $p_k \rightarrow p_i$ while strengthening the link $p_j \rightarrow p_i$, or will reduce the probability of p_k while strengthening the probability of p_j. Similar considerations may be made for the choice of children or brother nodes and information sources. Notice that, as it has been proved that humans tend to use heuristic rules rather than probability theory in their reasoning (Tversky and Kahneman, 1974), only evoking a fictitious premise or consequence (even if this has a weak link with p_i) may achieve the effect of distracting the Receiver's attention from the real one [5].

To illustrate the various types of deception instruments that we described, we consider, in **figure 2**, a subset of the graph describing our game.

Example 2:

In this example, we (temporarily) change the deception object by taking, as p_i, an intermediate node in the subgraph in figure 2: *"the muscular apparatus of X"*(Musc-app). The following are, in this case, candidate deception instruments:

a. *parent-nodes* Pre(p_i): *"Physical activity performed by X"* (Phys-act) and *"Body characteristics of X"*(Body);

b. *child-nodes* Con(p_i): *"Physical abilities of X"* (Phys-abil) and *"Sport practiced by X"*(Sports);

c. *brother-nodes*: *"Health conditions of X"* (Health).

Let us assume that S wants to convince R that X has a *'well developed'* muscular apparatus. To achieve this goal, it might say that *"X performs an intensive physical activity"*, that *"X's body has typical male traits"*, that *"X can lift*

[5] Walton mentions 'distraction' as one of the tactics of deceptive argumentation, by defining it as a situation in which *"the arguer not only argues for the wrong conclusion, but tries to throw the audience off track of the proper line of argumentation leading to the right conclusion"* (Walton, 1999).

heavy weights", that *"X regularly plays football"* or that *"X is in good health"*. It might support the first statement by saying that*"X's best friend says that X practices intensive physical activity"*. These data will all indirectly influence R's belief in the muscular apparatus of X, although in different ways and with different weights.

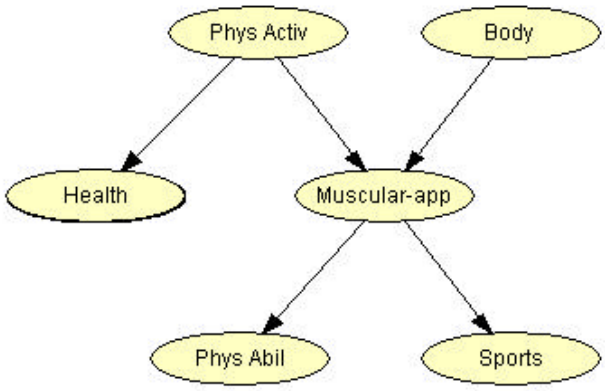


Figure 2: a subset of the network in figure 1

Example 3

Let us now return to our main game and deception object by assuming that S believes that R knows that *"X has a well-developed muscular apparatus"*; she will naturally be inclined to believe that *"X's body has typical male traits"* and therefore that X is a man. To contrast this belief, S may propose to R an alternative interpretation of what she knows: for instance, that the muscular apparatus of X is due to a great physical activity rather than to X's gender. To this aim, it might say that *"X performs an intensive physical activity"* or that "*Physical activity is much more effective in producing a well-developed muscular apparatus than having a male or female body structure" or* that *"X is in good health"* (a consequence of performing physical activity). As for information sources, if R believes that *"X's mother said that X is unable to perform any task involving physical activity at home"*, S may say: *"Well, you know, mothers can be biased sources about these matters. The best friend of X says, on the contrary, that X is always enthusiastic about participating in initiatives entailing a great physical activity*".

3.5 Deceiving while saying the truth

S does not necessarily have to tell a lie when making an attempt to deceive. It may deceive while saying something which it believes, basically in 3 ways:

a. *by planning not to be believed*: for instance, S might exploit R's distrust**;**

b. *by relying on R's self-deception*: S might exploit some incorrect inference that R will derive from its message;

c. *by transmitting an uninfluential truth* that leaves R in uncertainty or error.

A fact may be uninfluencial to both S and R, or only to R; this typically occurs when the System believes that the Receiver's mind is different from its own mind, so that the impact of believing in some particular fact is different for the two agents. For instance, if S believes that p_j is uninfluencial to p_i but presumes that R believes that this link is strong, it might say that p_j is true to induce R to believe in p_i.

Example 4

In Example 2, if S presumes that R believes (erroneously) that practicing Aikido requires a well-developed muscular apparatus, it might say that "*X's preferred sport is Aikido*" (by relying on R's self-deception).

Strategies aimed at diverting R's attention towards a fictitious premise, consequence or information source are particularly suited to deceiving while saying the truth: if several alternatives are available, S might select a fact in which it believes, even if this produces only a slight influence on the deception object. The advantage of falsely-sincere deception is that the risk of being discovered in a deception attempt, either immediately or in the long term, is slight and that the consequences of being discovered are not serious. S may always say*: "When I told you that p_j, I did not know that you would have concluded that p_i!"*. An in depth discussion about 'lying by saying the truth' may be found in (Vincent and Castelfranchi, 1981).

3.6. Evaluating the validity of a deception strategy

When several alternatives are available, S has to *select the deception instrument* that best achieves its goal of influencing R's belief about the deception object in the desired manner. As we anticipated informally in the previous Section, several characteristics of the deception instrument enter in this decision:

a its *impact* on the deception object. This is the main variable, which measures how effective will S's communication be (if accepted by R) in changing her belief in the desired direction and to the desired extent;

b its *plausibility* to R: this variable relates to the likelihood that R will become convinced of what S says, that she will believe it (Castelfranchi and Poggi, 1998). Not all proposed pieces of information are equally acceptable by R, as she already holds a set of beliefs and is able to reason about them;

c its *credibility* as a function of the cited information source: when S mentions a source as the information provider, the degree with which R believes in this fact depends on how informative she considers this source to be;

d its *safety*: this variable is linked to the risk that the deception attempt is discovered and combines two measures: the *risk of being discovered in a deception attempt* and the *seriousness of the consequences* of this event. The risk of being discovered depends, first of all, on how plausible it is to R, that S believes in what it says [6]. If deception is studied in a logical framework, plausibility is directly connected to logical consistency (see, for instance, Lee and Wilks, 1997); probability permits scaling this measure. The second factor affecting the risk of being discovered is the risk that R guesses that her believing what S says is 'convenient' to S, that it produces some advantage to S. This second factor, as well as the seriousness of the consequences of not being believed, requires representing, in S's model of R, some knowledge of what will happen if R comes to believe in the deception instrument: how much advantage R presumes S will make of this belief and how serious is what R herself might do after discovering the deception attempt. Seriousness of consequences may be decreased if S can defend itself if discovered, by justifying what it said.

e its *computational cost*: different strategies may be more or less complex in conception, performance or understanding. The cost of conceiving a strategy, to S, is a function of the computational complexity of the algorithm that selects the inference path. The cost of understanding it, to R, depends on the length of the reasoning chain involved and on the proportion of 'active' beliefs (in R's mind) in this chain.

---

[6] Notice the difference between the *plausibility of a fact to R* and the *plausibility, to R, that S believes in that fact.* The first measure is based on second-order beliefs, while the second one is based on third-order beliefs.

3.7. Selecting a good candidate

S's strategy should be aimed at trying to find the best compromise among these (often conflicting) needs. It must say something that is effective to produce the desired effect on R's belief, something that R finds plausible, that R does not find implausible for S to believe, that does not entail consequences too risky if the deception attempt is discovered and finally, that S may produce and R may understand without too much effort. Finding a compromise among these criteria is difficult in the majority of cases. This is the reason why the ability to deceive is a measure of intelligent behavior. The relative weight of the various factors depends on the context. Let us fix the importance, to S, of convincing R (that is linked to the weight of the domain-goal g_k). The context then includes the presumed *attitude* of the Receiver (how suspicious she is), the *environment* in which the dialog occurs (for example, whether R and/or S are in an emotional state or hurried) and which may be, to S, the *consequences* of being discovered in a deception attempt. If R has no particular reason to be suspicious, she tends to accept rather naturally a plausible fact; she reasons on the plausibility that S really believes it only if she find it implausible. A suspicious R tends, on the contrary, to guess much more about the plausibility of the communications she receives and on the convenience, to S, that she believes in these facts. Therefore, plausibility is very important in communicating with non-suspicious Receivers, while safety is important if R is suspicious. Information sources are a double-edged weapon: on one hand (as we said) they may increase the plausibility of a fact; on the other, they may decrease the level of suspicion of R (*"Why did S cite this source? Is S uncertain of what it says? Is it trying to hide itself behind another source to avoid being blamed of what it said?"*… and so on). Simple (to conceive and perform) strategies are natural when S is hurried or when it is in an emotional state; the same applies to the cost of understanding and the state of R.

A particular consideration merits what may be called the *horizon effect*: a strategy may be good in a given phase of the dialog but may reveal to be bad in the future. A good strategy is therefore also a strategy that leaves good strategies open for the future. First of all, due to the D-separation property of BNs (Pearl, 2000), fixing the truth value of a node makes any further evidence about its descendants uninfluencial, unless these are linked to the deception object by other paths not passing through this node. So, when the deception object is the root of the network (as in our case), peripheral nodes are preferable because they leave the maximum freedom to future choices.

Example 5

Talking about *Personal preferences of X* has a good horizon effect although this is a root of the network, because its descendants (like 'Look' or 'Muscular apparatus') are linked to the deception object also by paths not passing through that node. The same is true for the *Physical activity performed by X*. On the contrary, saying that *The look of X is not clearly defined* prevents using, in the future, arguments based on X's hair, clothing or use of cosmetics.

We focused our study on how to define strategies in which only the first three variables are considered. Therefore, the rest of this paper will not consider safety and cost. Firstly, this is to simplify the problem. But it is also justified by the fact that, in the game we consider, the consequences, for S, of being discovered in a deception attempt with any deception instrument are always the same; R will understand the system's goal and will win. Due to the limited number of nodes in the graph, the various strategies also have very similar complexities.

3.8. Building a image of the Receiver's mental state

To reason about the Receiver beliefs, the System has to make some hypotheses about her mental state, that it will revise as the interaction proceeds and new information is acquired.

Notation 3. We call:

OWN-S the *belief network* that represents the *System's mental state*,

IM-S-R the *belief network* that represents the *System's image of the Receiver's mind*.

Nodes in OWN-S and in IM-S-R represent, respectively, first-order and second-order beliefs. The structure of IM-S-R may be the same as that of OWN-S (*identical model*), a subset of OWN-S (*overlay model*) or may partially overlap it (*shift model*). The two models may also differ in the parameters attached to links among their nodes.

Definition 5. We say that:
IM-S-R is *probabilistically compatible* with OWN-S if its parameters are defined so that the weights of the links among the nodes that the two models share are the same (the conditional probability tables are marginalized, as in (Zukerman et al, 2000)).

The difference between OWN-S and IM-S-R influences the selection of nodes candidate to deception: in identical models, S may select any fact it knows. In overlay models, the selected node should be an item that R is presumed to infer easily. In shift models, S will have to learn how to update its own model according to R's knowledge before selecting a candidate. In building IM-S-R, the facts acknowledged during the dialog are considered: what S said and what R presumably believes. However, knowledge including only these facts would be sparse. To connect them in a network, S has to make some hypothesis about the *line of reasoning* that R followed in formulating or in accepting these facts. This line of reasoning depends on the type of dialog and influences the IM-S-R building algorithm, as we will see later on.

3.9. Synthesis of the constituents of deception strategies

We may now summarise the topics that should be considered in formalising deception strategies: a) how to select a deception object, b) how to build an image of the Receiver's mental state, c) how to decide whether to apply direct or indirect deception and d) how to select, in indirect deception, a 'convenient' deception instrument. If we assume that the deception object is fixed and that indirect deception is applied (as it is reasonable to do in our game), formalisation is focused on steps b. and d. In the rest of the paper, we will discuss how the probability theory may help in solving these issues.

4. A Deception Simulator: 'Mouth of Truth'

Mouth of Truth is a tool that enables the User to learn and test how various forms of deception may be applied in planning utterances that affect a Receiver's belief so as to achieve a given influence goal. The Systems suggests some deception strategies with criteria for evaluating them. The problem considered with the tool may be stated as follows. Given:

- a *domain*;
- a *domain reasoning model* of the System S, represented in a belief network OWN-S;
- a *deception object* p_i corresponding to one of the nodes in OWN-S and one of the following goals about R's belief in p_i: G1: $G_S\, B_R\, p_i$, G2: $G_S\, B_R\, \neg p_i$ or G3: $G_S\, \neg BW_R\, p_i$;
- a set T of *domain-facts* p_k ($k = 1,\ldots, w$) known to R, such that: $\forall k\ B_S\, B_R\, p_k$,[7]
- a set *models* that represent the relationships between *information* provided by *sources* and truth values of the facts they refer, again represented as belief networks:

[7] We will assume, from now on, that reasoning involves multi-valued variables; so, p_j is a statement specifying the value taken by a multivalued variable v_j associated with a node in the belief network.

build a image of the mental state of R (IM-S-R) and find a (not necessarily sincere) 'candidate' fact p_j such that $p_j \neq p_i$ (indirect deception), $p_j \notin T$ (p_j is 'new' to R), G1, G2 or G3 is achieved 'to the best extent' if R believes in p_j. If several candidates exist, analyse advantages and disadvantages of each and suggest the best compromise. Cite, if needed, one or more information sources to increase the plausibility of p_j. Users interact with the System by a Web interface, which guides them in selecting a domain and a deception object and introducing the set T (figure 3).

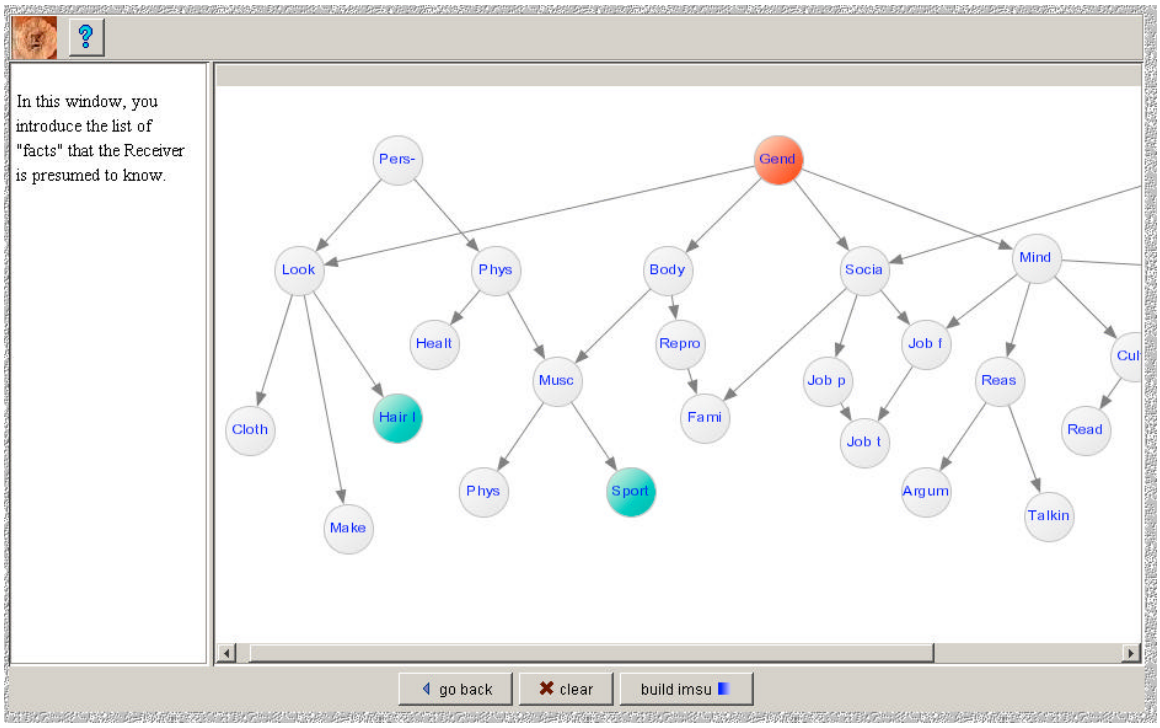


Figure 3: inserting 'facts known to the user' in the interface of Mouth od Truth

The System then proposes a list of facts that might be communicated; these facts correspond to different deception strategies that may be applied in that context and are illustrated with a comment in natural language. Each candidate is associated with a set of evaluation measures that permit assessing its advantages and disadvantages. The User may test how the suggested strategies vary according to the experiment conditions. We now describe the main System modules by showing some examples of results.

4.1. Building a Default Image of the Receiver's Mind

This is done in two steps: in the first one, the structure of the BN is built; in the second, parameters are attached to the BN. As we said in Section 3.8, in building IM-S-R some hypothesis about the *line of reasoning* that R followed in formulating or in accepting the facts she presumably knows has to be made. As we may assume that, in our game, R has no other source of information than S, IM-S-R is a overlay model of OWN-S, which we build by applying a limited notion of 'simulative reasoning' [8]. IM-S-R includes only the beliefs that S may assume to be *active*, in the Receiver's mind, in the present phase of interaction. Active beliefs include, first of all, the set T of facts that R

[8] Simulative reasoning is reasoning of the following form: "R believes p_1, p_2, ..., p_n; if I believed those things, then I would also believe q; therefore, R believes q" (Kaplan and Schubert, 2000).

knows and the deception object p_i. This set is not extended to all the consequences that might be drawn from them, as would happen in simulative reasoning. To minimize assumptions about R's beliefs, only the *minimal chains of reasoning* that are needed to establish a relationship between every fact in T and the node p_i are included in IM-S-R. The idea behind this choice is that, when S communicates a fact p_k to R, R reasons on how p_k would affect her belief in p_i. In following this line of reasoning, she activates the beliefs along a path from p_k to p_i; if several such paths exist, the shortest one is selected. The pruning algorithm is the following:

Algorithm Pr (OWN-S, T, p_i)
    IM-S-R = Copy(OWN-S)
    For All p_k in T, ?[k] = SearchMinimalPaths(T, p_i, OWN-S)
    DeleteNodes(IM-S-R, ?[k]).
Where SearchMinimalPaths(T, p_i, OWN-S) explores, for every p_k in T, the edges in OWN-S
to find a set of minimal paths ?[k] = < p_k =p_1,…,p_m=p_i> connecting p_k with the deception object p_i.
This procedure iterates a bidirectional breadth- first search algorithm |T| times by uniformly expanding,
at every step, the borders of the current p_k and p_i until an intersection between the two borders is found.
Notice that:
1) at the generic step n, all the nodes at distance n from p_k and p_i are explored
2) a vertex's border is the set of nodes that are adjacent to the vertex and have not yet been explored.

The procedure DeleteNodes(IM-S-R, ?[k]) deletes from a copy of OWN-S the nodes ***not*** included in ?[k].

The worst-case time complexity of every step of the SearchMinimalPaths algoritm is the same as that of a breadth-first search algorithm, i.e. O(n + m), where n is the number of nodes and m is the number of edges in OWN-S. As we apply the procedure to all nodes p_k in T, the time complexity of the pruning algorithm is O(|T|*(n + m)), that is (n**2 + n * m) in the worst-case, when |T| = n.

Figure 4 shows an example of the IM-S-R that is built from the OWN-S in Figure 1 after S said that "*X has long hair*" and "*X's preferred sport is football*". Parameters are assigned to IM-S-R so that the two models are probabilistically compatible (see Definition 5): this is a default assumption which is dynamically revised as far as interaction goes on.

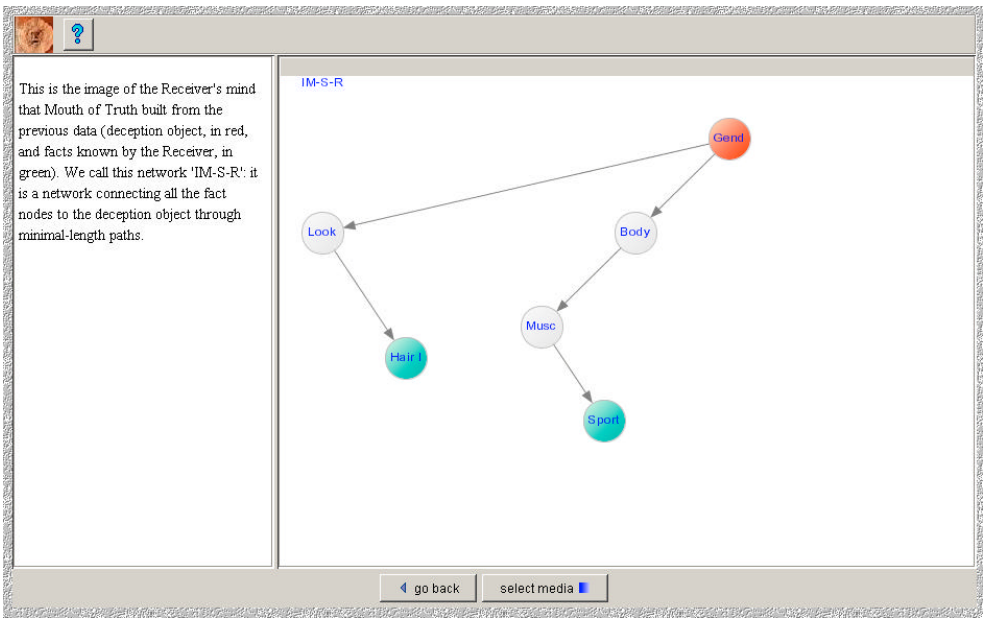


Figure 4: the system's image of the receiver's mind (IM-S-R) built from MoT from data in figure 3

4.2. Defining a communicative goal

S sets a goal-value for the probability of the deception object:

$$\text{goal-value of } P(p_i) = \begin{cases} 0.5 \text{ if } G_S \neg BW_R\, p_i\,, \\ 0 \text{ if } G_S\, B_R \neg p_i \text{ or} \\ 1 \text{ if } G_S\, B_R\, p_i\,. \end{cases}$$

It then propagates, in IM-S-R, the evidence in T and considers the probability that, in its view, R assigns to the deception object p_i, in her present state of knowledge (the 'prior' probability of p_i).

4.3. Listing candidates to indirect deception

In selecting the candidates to indirect deception, *the strategy adopted is to suggest to R an interpretation of the facts which she knows influence the deception object in the desired way*. For this aim, S selects candidates to deception among the 'near relatives' of facts in T. The algorithm includes in a list L of candidates 'parents', 'children' and 'brother' nodes of elements of T whose values may influence p_i, thereby excluding nodes that cannot influence p_i due to the 'D-separation' property of belief networks (Pearl, 2000). With this strategy, the System considers the possibility of applying various forms of deception (on causes, on effects and by diversion) without forgetting the 'horizon effect' (that is, the analysis of the long-term effects of its present move), so as to leave the maximum freedom to the selection of candidates in future moves. Mouth of Truth generates an explanation to justify its choice of candidates, by applying template-based natural language generation (Reiter and Dale, 2000) to knowledge associated with the belief network.

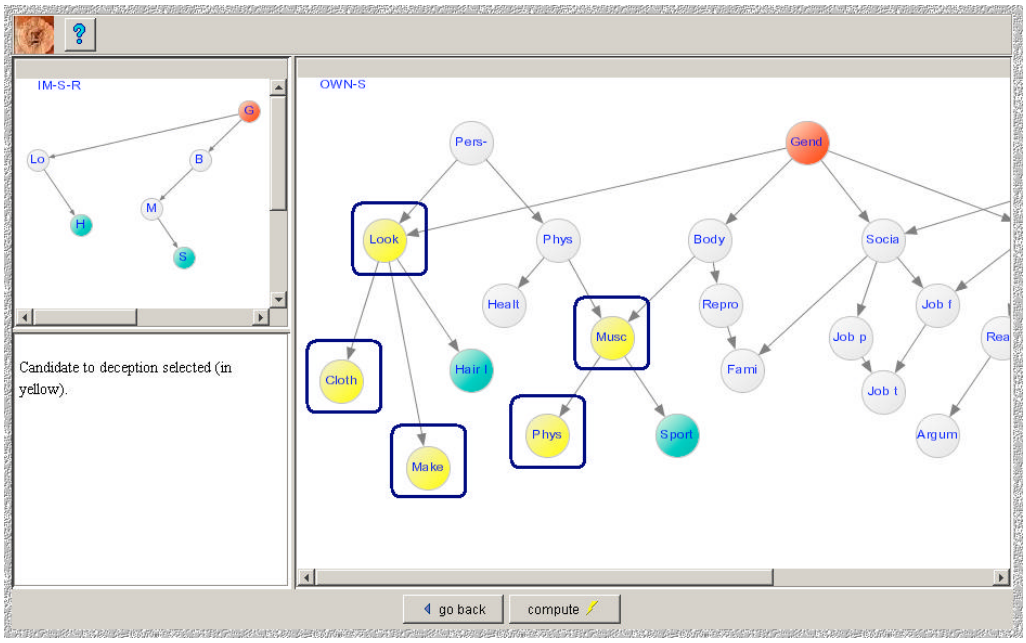


Figure 5: candidates to deception suggested by MoT in the example in figure 4

Example 6:

Let us, again, consider the IM-S-R in Figure 4, where T: {'Hair=long', 'Sports=football'} and p_i = 'Gender'. If S wants R to ignore the gender of X, the goal-value of P('Gender = woman') is set to .5.

The list of candidates to deception includes, in this case, the parents of nodes in T ('Look' and 'Muscular-app') and their brothers (Clothing', 'Makeup' and 'Phys-abil'): see Figure 5. That is, in the order:
"The look of X is typical of women / not clearly defined / typical of men".
"The muscular apparatus of X is well developed / moderately developed"
"X may wear skirts / trousers / both"
"X uses /does not use cosmetics"
"X is agile / can lift heavy weights".

The generated explanation text is the following:
"*We consider a deception strategy that is aimed at inducing R to interpret what she knows in a way that is convenient for your goal. Remember that your goal is that R ignores X's gender.*

a. You said, first of all, that R knows that "The length of X's hair is long". You might then talk to R about the following characteristics of X:

- *'Look' ("X's look is typical of women, not clearly defined or typical of men"). This is a premise of R's belief about the length of X's hair. The choice of this node corresponds to a strategy of 'deception on a cause of a belief'.*
- *'Clothing' ("X may wear skirts / trousers / both skirts and trousers") or 'Makeup' ("X uses cosmetics / does not use cosmetics"). The choice of one of these nodes corresponds to a strategy of 'indirect deception on the cause of a belief' : indirect because it is applied to consequences of this cause rather than directly on the cause itself.*

b. You also said that R knows that "The Sport practiced by X is football". You might then also talk to R about the following characteristics of X:

- *'Muscular apparatus' ("The muscular apparatus of X is well developed / moderately developed"). This is the cause of the belief of R about the sport practiced by X: see previous comment.*
- *'Physical abilities'. The choice of this node corresponds to a strategy of 'indirect deception on the cause of a belief' ('Muscular apparatus'): see previous comment.*"

4.4. Evaluating candidates to deception.

For every candidate in L, impact, plausibility and credibility of the data are evaluated as a function of the informativity of a cited source. Let us describe how we measure these variables.

a. Measuring impact.

Notation 4. Let:

$B_S B_R\ P(p_i)$ denote the prior probability that, in S's view, R attaches to the deception object p_i, in the considered phase of the game;
$G_S B_R\ P(p_i)$ denote the probability that S would like p_i to take, to R;
$B_S B_R\ P(p_i \mid p_j)$ denote the posterior probability that, again in S's view, R will attach to p_i after believing in the deception instrument p_j.

Definition 6. We propose two measures of the impact of p_j on p_i:

- *posterior goal distance:*

$Imp^1\ (p_j$ ¿ $p_i) = 1- |B_S B_R\ P(p_i \mid p_j) - G_S B_R\ P(p_i)|)|$, whose values range in the (0,1) interval

- *probability change*:

Imp^2 (p_j ¿ p_i) = 1- |$B_S B_R$ P(p_i | p_j) - $B_S B_R$ P(p_i)|, whose values range in the (0,1) interval [9]

These measures are useful in different contexts. Imp^1 is appropriate when the influence goal is to change R's belief about p_i ((←p_i) or (→p_i)). Imp^2 is appropriate when the influence goal is to leave R's belief about p_i invaried (-p_i). [10]

Example 7:

Let us, again, consider Example 6. Knowing that X has got long hair and practices football, R suspects that X is a man (prior probability of being a woman = .41). S considers the nodes in the list L of candidates: for instance, X's physical ability (Phys-abil). S might say that "*X is agile*" or that "*X can lift heavy weights*", and evaluates the impact of the two alternatives on the deception object:

Case 1: to protract the game, S wants to leave R in doubt about X's gender: 'Gender of X=unknown", that is $G_S B_R$ P(p_i) = .5 (see Figure 6).

- if p_j = (Phys-abil='is agile'), the probability that "Gender = woman" becomes = .55. Then, Imp^1 (p_j ¿ p_i) = .95 and Imp^2 (p_j ¿ p_i) = .86.
- if p_j = (Phys-abil='can lift heavy weights'), the probability that ("Gender = woman") becomes .39. Then, Imp^1 (p_j ¿ p_i) = .89 and Imp^2 (p_j ¿ p_i) = .98

The first alternative ('X is agile') is preferable: the posterior goal distance Imp^1 (p_j ¿ p_i) which, as we said, is the a appropriate measure to employ in this case, is close to that desired.

Case 2: for some reason, S wants R to believe that X is a man:

- if p_j = (Phys-abil='is agile'), Imp^1 (p_j ¿ p_i) = .47 and Imp^2 (p_j ¿ p_i) = .81
- if p_j = (Phys-abil='can lift heavy weights'), Imp^1 (p_j ¿ p_i) = .74 and Imp^2 (p_j ¿ p_i) = .92

The second alternative ('X can lift heavy weights') is preferable.

The conclusions in the two cases are reasonable: saying that X is agile is a good strategy to strengthen R's doubt about X's gender, while saying that X can lift heavy weights contributes to convincing R that X might be a man.

b. Measuring plausibility

Definition 7. We adopt a notion of '*plausibility as compatibility with previous and active beliefs*'. The following are our measures of plausibility of p_j to R*:*

- *local plausibility*:
  $Plau^1(p_j) = B_S B_R P(p_j)$.
- *global plausibility*:
  $Plau^2(p_j) = 1 - 1/w\ \Sigma_{k=1,...w}$ | $B_S B_R$ P(p_k|p_j)- $B_S B_R$ P(p_k)|, with $p_k \in$ {active beliefs}

Both variables range in the (0,1) interval. $Plau^1$ depends only on the prior belief of R about p_j: the higher this value, the easier it is, for R, to accept S's communication that 'p_j is true'. $Plau^2$ is a function of the overall variation (in absolute value) in the probabilities of *all active beliefs* which is induced, in R, from believing in p_j.

---

[9] The measure of impact in (Zukerman, 2000), considers the same elements, although with a function of different shape (logarithm of the ratio between posterior and prior probability).

[10] Another measure of impact, which is appropriate when one wants to compare situations with different prior probabilities of the deception object, is the following:
Imp^3 (p_j ¿ p_i) = 1- [($B_S B_R$ P(p_i | p_j) - $G_S B_R$ P(p_i)) / ($B_S B_R$ P(p_i) - $G_S B_R$ P(p_i) + .0001)]
The .0001 constant is a 'statistical zero' that avoids divisors with zero value. The closest is Imp^3 to 1, the most effective is p_j in producing the desired change on p_i. Negative values correspond to changes in the wrong direction, positive values to changes in the right direction.

Example 8:

We continue the previous example to measure the plausibility of the two values that the variable "Physical ability of X" might have.

The prior probability that X 'is agile' is =.14; then, the local plausibility $Plau^1$ =.14 while the global plausibility $Plau^2$ = .78.

The prior probability that X 'can lift heavy weights' is = .86; then, $Plau^1$ = .86 while $Plau^2$ = .97.

Both measures indicate that *"X can lift heavy weights"* is more plausible, to R, than *"X is agile"*: remember that the first candidate had a great impact in convincing R that X is a man while (obviously) it had a much lower impact in leaving R in doubt about X's gender.

Figure 6 shows the table displayed by MoT in example 6 (that is, when *R knows that X has got long hair and plays football*). As R suspects that X is a man, the facts that have a great impact on increasing R's uncertainty about X's gender are those that are typical of women (use of cosmetics and being agile) or more ambiguous answers, such as 'wears both trousers and skirts'. This Table also shows that the facts with a good plausibility are those consistent with the previous knowledge of R: 'wearing skirts' and 'use cosmetics' are plausible because they are consistent with 'long hair'; 'ability to lift heavy weights' and 'having a well-developed muscular apparatus' are consistent with the practice of football; finally, wearing trousers is compatible with the two facts known by R.

c. Measuring safety

We measure the risk for S of being discovered in a deception attempt, by estimating how plausible R might consider that S believes in what itself says**.**

Notation 5. Let us denote, with

$B_S\ P^*(p_j \mid p_i,\{T\})$ the probability that S believes in p_j, given what it knows about X's gender (that X is a woman or a man) and given the set of facts that itself declared to believe in previous phases of interaction.

Definition 8. The following is our measures of *safety of* p_j :

$Saf\ (p_j) = B_S\ P^*(p_j \mid p_i,\{T\})$[11]

Again, this variable ranges in the (0,1) interval. Its value is obtained by propagating, in OWN-S, evidence about X's gender (that S knows but R ignores) and about all the facts that S communicated to R: those in which S really believes and those in which S does not believe. This variable enables S to evaluate the risk of contradicting what it said in precedence by talking about p_j and to consider, at the same time, the risk of contradictions in future communications. As a final comment to Figure 6, notice that safety and local plausibility are quite related measures: the few exceptions are 'wearing trousers', 'do not use cosmetics' and 'have a look typical of men' (which are characteristics not true for women, in general) .

d. Introducing an information source

Candidates to deception may be presented directly or may ascribed to some information source; this is especially useful to increase the plausibility of facts that might have a great impact on the deception object but are not sufficiently plausible to R.

Notations 6. Let:

[11] We normalize this measure to eliminate differences due to the different number of states of nodes in the BN.

- IS denote an *information source* for the fact p_j;
- Say (IS, p_j) denote the event: *'IS says that p_j is true'* (and alike, for Say (IS, $\neg p_j$));
- Comp (IS, p_j) denote the *competence* of IS about p_j;
- Sinc (IS, p_j) denote the *sincerity* of IS about p_j;
- Inf (IS, p_j) denote the *informativity* of IS about p_j;
- Cred (p_j, IS) denote the *credibility* of the fact p_j when referred by IS.

Definition 9. The following are probabilistic definitions of the previous concepts:

- Comp (IS, p_j) is a function of the value couple $\{P(B_{IS}\, p_j \mid p_j), P(B_{IS}\, p_j \mid \neg p_j)\}$; that is, of the probabilities that IS believes in p_j when this is true and when this is false.
- Sinc (IS, p_j) is a function of the value couple $\{P(\text{Say (IS, } p_j) \mid B_{IS}\, p_j), (P(\text{Say (IS, } p_j) \mid B_{IS}\, \neg p_j)\}$; that is, of the probability that IS says that p_j is true when it believes it and when it does not believe it.
- Inf (IS, p_j) is a function of the couple of values $\{P(\text{Say (IS, } p_j) \mid p_j), (P(\text{Say (IS, } p_j) \mid \neg p_j)\}$ and is a combination of competence and sincerity [12]. The informativity of an information source about p_j may be low because the source does not recognise when p_j is true (is not much competent) or because, although recognising it, it is not completely sincere about it. One may therefore keep the two concepts (and measures) of competence and sincerity separate or may combine them into a unique, overall evaluation of informativity, as we do in our tool.

The goal is Gender (unknown)

Deception media		Deception object		Impact		Plausibility		Safety
		prior belief	post belief	posterior goal distance	probability change	local plausibility	global plausibility	
Clothing	skirt	0.409387	0.692294	0.807706	0.717093	0.598865	0.780997	0.824166
	trousers	0.409387	0.120302	0.620302	0.710916	0.588065	0.812641	0.227058
	both	0.409387	0.394051	0.894051	0.984665	0.126139	0.827663	0.12461
Makeup	uses cosmetics	0.409387	0.569449	0.930551	0.839938	0.601844	0.864653	0.825696
	does not use cosmetics	0.409387	0.16744	0.66744	0.758053	0.398156	0.807522	0.174304
Look	typical of women	0.409387	0.754094	0.745906	0.655293	0.596203	0.795196	0.866774
	not clearly defined	0.409387	0.17582	0.67582	0.766434	0.497106	0.810598	0.254263
	typical of men	0.409387	0.0146232	0.514623	0.605236	0.310489	0.736051	0.0121891
Phys_abil	is agile	0.409387	0.553171	0.946829	0.856216	0.137442	0.783574	0.191803
	can lift heavy weights	0.409387	0.386476	0.886476	0.977089	0.862558	0.965514	0.808197
Musc_app	well-developed	0.409387	0.383471	0.883471	0.974085	0.953197	0.980401	0.885246
	moderately-developed	0.409387	0.937186	0.562814	0.472201	0.0468027	0.60085	0.114754

Figure 6: measures of impact, plausibility and safety of candidates in figure 5, as computed by MoT

The credibility of a fact referred by an information source: Cred (p_h, IS)= P(p_h | Say (IS, p_h)) and Cred (p_h, IS)= P(p_h | ¬Say (IS, p_h)) may be computed from the informativity of the source and the prior probability of the fact. A fact may have a high credibility because its prior probability is high even when the source is not very informative, or the

[12] P(Say (IS, p_j) | p_j) = P(Say (IS, p_j) | $B_{IS}\, p_j$)*P($B_{IS}\, p_j$ | p_j) + P(Say (IS, $\neg B_{IS}\, p_j$) | p_j)*P($\neg B_{IS}\, p_j$ | p_j), and alike for P(Say (IS, p_j) | $\neg p_j$). This kind of measure is common in medical tests, in which P(Say (IS, p_j) | p_j) is called 'sensitivity' and P(¬Say (IS, p_j) | $\neg p_j$) is called 'specificity': sensitivity and specificity are a combination of the technical characteristics of instruments employed in making the test and of the competence of those that interpret results (Weinstein and Fineberg, 1980)

inverse. In our simulation game, the prior probability that R attaches to a fact communicated by S corresponds to its local plausibility. The less plausible is a fact to R in S's view, the higher should be the informativity of the mentioned source.

Example 9

Let p_j be a fact whose plausibility, to R is equal to .5: for instance: "*X tends to cry frequently*"; this means that, in S's view, the Receiver doesn't know whether X tends to cry frequently.

- Let IS-1 be an *informative* source about p_j, that is a source such that: {P(Say (IS, p_j) | p_j)=.8, (P(Say (IS, p_j) | $\neg p_j$)=.2}: IS-1 is informative because it tends to say that p_j is true when (and only when) it is so. IS-1 might be the best friend of X who knows X and tells the truth (is competent and sincere). Let us now assume that S tells R that IS-1 said that "*X tends to cry frequently*"; the credibility of p_j to R raises to .80. If, on the contrary, S communicates that IS-1 said "*X doesn't cry frequently*", the credibility of p_j is reduced to .20.
- Let now IS-2 be a *'not informative'* source: for instance, a source such that: {P(Say (IS, p_j) | p_j)=.2, (P(Say (IS, p_j) | $\neg p_j$)=.5}. IS-2 is not informative because it tends to conceal that p_j is true when it is so and to behave randomly when p_j is false. Let us now assume that S communicates to R that IS-2 said that "*X tends to cry frequently*"; the credibility of p_j lowers to .29. If, on the contrary, IS-2 said that "*X doesn't cry frequently*", the credibility of p_j raises to .62.
- Finally, let IS-3 be a source that is '*rather informative'* although less so than IS-1, that is such that: {P(Say (IS, p_j) | p_j)=.7, (P(Say (IS, p_j) | $\neg p_j$)=.3}. If S declares that "*X tends to cry frequently*", the credibility of p_j raises to .70 while, if IS-3 said that "*X doesn't cry frequently*", the credibility of p_j is lowered to .30.

Belief networks enable representing how information which is provided by various sources, each with its own informativity and with possibly different views about p_j, may be combined into a unique measure of credibility. In Mouth of Truth, these belief networks are separated from OWN-S and from IM-S-R. They include two types of nodes: $B_S B_R$ p_j and $B_S B_R$ Say(IS-i, p_j); the arcs are oriented from the first type of nodes to the second one, according to the general criterion of orienting arcs in BNs from 'hidden' to 'observable' facts. We use these networks to reason abductively by introducing evidence about the facts referred by one or more information sources and by checking how the value of $B_S B_R$ p_j varies after this evidence has been propagated. We may also simulate situations in which S tries to 'confound' the User's ideas by mixing informative with uninformative sources.

Example 10:

Case 1: convergence of informative sources:

To continue with the previous example, if S declares that both IS-1 and IS-3 said that p_j is true, the credibility of p_j becomes = .90. If they both say that p_j is false, its credibility is lowered to .10.

Case 2: convergence of sources with different levels of informativity:

If S declares that both IS-1 and IS-2 said that p_j is true; the credibility of this fact will be = .62. If both sources said that p_j is false, its credibility is reduced to .29.

Case 3: contradiction between informative sources:

If S declares that IS-1 said that p_j is true while IS-3 said that it is false, the credibility of this fact slightly increases (.63) because IS-1 is more informative than IS-3.

Note that, in the previous example, the effect of a non informative source (IS-2) is that of confounding the Receiver, even when evidence from an informative one (IS-1) is provided at the same time.

4.5. Combining the various measures

As we said in Section 4.1, the purpose of Mouth of Truth is not to try to deceive its Users, but rather to enable them to learn and test how various forms of deception may be applied in a given situation. The final step of its reasoning applies a strategy to select one or more 'good' candidates, that may be formalized as follows:

Select the subset of candidates whose *impact* is higher than a given threshold;
if no such candidates exist, then 'no candidates found';
else order candidates in decreasing degree of impact;
select the subset of candidates in this list whose *safety* is higher than a given threshold;
if no such candidates exist, then 'no candidates found';
else analyse candidates in the order of inclusion in this list;
if the *plausibility* of the examined candidate is higher than a given threshold, or
an 'informative source' exists to increase the credibility of this candidate, then
display it to the User, with an appropriate comment.
If 'no candidates found', then declare that
"a deception strategy aimed at inducing R to interpret what she knows in a convenient manner for your goal cannot be applied".

Example 11:

Let us consider again Example 6. As S wants to change R's belief about X's gender, it applies a measure of impact based on the *posterior goal distance* (Imp^1). S selects, in the list of candidates in the first column of figure 6, those whose Imp^1 exceeds a fixed threshold (for instance, .9):

"X is agile" (Imp^1=.94)

"X uses cosmetics" (Imp^1=.93)

It then selects, from this list, the items whose safety exceeds a given threshold (for instance, .8):

"X uses cosmetics" (Saf=.82)

As the plausibility for the selected candidate is quite low ($Plau^2$ = .6: local plausibility, in column 7 of figure 6), the system suggests mentioning an information source to improve it. For instance: *"Market surveys prove that both men and women use cosmetics"*.

Notice that the 'best' candidate depends on the order in which the three measures are considered, that is on the priority that is given to the criteria they represent.

5. Evaluation

To evaluate Mouth of Truth, we decided to study whether the criteria applied by the system when reaching its decision are 'cognitively plausible'; that is, whether they are similar to those applied by humans and what are the eventual differences. The advantage of this *structural* approach of evaluation is that, in addition to providing a global image of the quality of results produced by our system, it gives us some insight on the reasons for differences thereby enabling us to refine our method. In particular, we aimed at answering the following questions:

a. Are our *measures of impact, plausibility and safety* compatible with the subjects' evaluations of the same conceptual measures?

b. *Are these measures applied* in selecting a candidate move? Which of these measures is privileged?

c. Does the *influence goal* affect the selection criteria? That is: do results differ if the Receiver initially tends to be convinced that X is a man or a woman, or has no idea about X's gender?

5.1. Study design

The study employed a self-administered questionnaire that began with the description of the study goal:

> *"We ask you to participate in a simplified version of Turing's Imitation Game. Let us imagine that you are playing with a player G who has to guess the gender of a person that we will call 'X' by asking you to 'say something about' X. You know that X is a woman but want to leave G in doubt about X's gender, even by providing incorrect or vague information. You want to avoid, however, that G discovers, during the play (either immediately or in the long term) whether you are sincere or not.*
> *You know that G already knows something about X; now, you have to make your move, in which you will give to G one new item of information. What would you say? We propose some alternatives and ask you to evaluate each of them".*

Three Cases were then described:

Case 1 is the stereotype of a woman (*X uses cosmetics and cries easily*),
Case 2 is the stereotype of a man (*X is a manager and likes war movies*), while
Case 3 is a person with no clear gender identity (*X likes talking about private matters and can lift heavy weights*).

The order of presentation of the three scenarios was varied casually among the subjects.
For each Case, a list of 5 'candidate moves' was proposed. For instance:

- In Case 1, you might say: *'X usually wears trousers'*, *'X tends to show emotions'* etc;
- In Case 2, you might say: *'X likes reading love stories'*, *'X has a leading job position'* etc;
- In Case 3, you might say: *'X practices yoga'*, *'X tends to reason logically'* etc.

Subjects were asked to evaluate, on a scale from 1 to 5, the 'plausibility', 'impact' and 'risk of being discovered in a deception attempt' for each candidate move. The three questions were formulated as follows:

Plausibility: *Would G find this plausible?*
Impact: *Would you succeed in insinuating the doubt about X's gender in G?* (for Cases 1 and 2) or
Would you succeed in leaving G in doubt about X's gender? (for Case 3)
Risk: *Would G risk discovering that you are trying to deceive him or her?*

We therefore measured safety in terms of the (inverse) quantitative variable 'risk of being discovered' and we will show the results in terms of 'risk' rather than 'safety'. Subjects were asked to assign each question a score between 1 (answer = 'no') and 5 (answer = 'yes'). They were also asked to select, for each case, what they consider to be the *best* and the *worse* candidate move, among the 5 options.

5.2. Subjects involved

At least two categories of 'confounding factors' risk to distort the results of an evaluation study in this domain. First of all, if it is true that deception requires some form of intelligence, the deception ability of humans varies considerably among individuals and is influenced by their age, social and educational status. Secondly, the personal viewpoint concerning the characteristics of men and women influence (again considerably) the parameters in the domain model (OWN-S) which is employed in the decision process; as a consequence, differences between system and human decisions may be due to differences in these parameters rather than in the criteria applied. For this reason, we fixed the following study conditions:

- we selected subjects with the same age, social and educational status (graduating students in computer science), equidistributed according to their gender;
- we performed a *pre-test study*, in which we asked 8 subjects to subjectively estimate the probability that a list of persons (each described by a couple of characteristics) was a man or a woman. The subjects were of the same type as those subsequently included in the main study, although they were not the same. The questions were of the form: *"How likely is it, in your opinion, that a person who practices Aikido and is not aggressive is a man?"*. We employed the results of this pre-test to adjust the parameters in OWN-S so that they reflect, as much as possible, the viewpoints of subjects subsequently included in the main study.

Thirtysix subjects (18 man and 18 women) with an average age of 25.1 years (range 20-31) were involved in the final study. Each subject examined 15 questions (5 moves $*$ 3 cases), thereby permitting analysis of results from 540 candidate moves overall.

5.3. Main results

a. Subjective estimates of plausibility, impact and safety.

As the distributions of these variables are frequently asymmetrical, the percentage of cases within a one-unit interval around the mode was used as a measure of the *level of agreement* among subjects in the evaluation of that measure. The level of agreement is higher for plausibility (62%) than for risk (59%) and impact (55%). Some examples of good agreement concerning plausibility are the following: *'X tends to show emotions'* (high plausibility, in Case 1), *'X likes love stories'* (low plausibility, in Case 2). Examples of good agreement concerning impact: *'X plays football'* (low impact, in Case 1). Examples of good agreement concerning risk: *'X practices yoga'* (in Case 3). The reduced level of agreement about impact may be explained by considering that subjects had to apply two different conceptual measures in the three Cases: a 'posterior goal distance' in Cases 1 and 2 (when the influence goal was to change the Receiver's belief about the deception object) and a 'probability change' in Case 3, when the influence goal was to leave this belief invaried (Section 4.4.a).

b. Agreement between subjective and MoT's evaluations, for the three measures.

The 540 responses were classified as having low, medium or high plausibility, impact and safety, for subjects and for MoT. For MoT, we considered *local plausibility* for the three cases, *posterior goal distance* for Cases 1 and 2 and *probability change* for Case 3. Subjective and automatic evaluations of these variables are related as shown in Tables 1a, b and c. Agreement between subjects and Mouth of Truth is, again, very high for plausibility, where 51% of answers match ($\chi^2 = 95$; $p<<.001$). It is a bit lower for impact (match in 47% of cases; $\chi^2 = 52$; $p<<.001$) and risk (match in 46 % of cases; $\chi^2 = 41$; $p<<.001$). Subjects and MoT agree especially when the impact or the risk of the move is 'low', while they agree less in distinguishing between 'high' and 'medium' values.

Subj MoT	H	M	L
H	120	86	44
M	50	73	23
L	22	42	80

Table 1a: Plausibility

Subj MoT	H	M	L
H	51	61	34
M	37	63	42
L	38	74	140

Table1b: Impact

Subj MoT	H	M	L
H	39	61	44
M	18	60	66
L	24	77	151

Table 1c: Risk

c. Preferred selection strategies

Subjects seemed to have particular difficulty in answering this part of the questionnaire. Some of them changed their choice several times before taking a decision; others attempted to select more than one candidate in the 'best' and in the 'worse' category.

Case	Question	Best	Worse	Plausibility	Impact	Risk	Strategy
Case 1 *Uses cosmetics and cries easily*	Usually wears trousers	8	7	3.9	2.4	2.4	
	Has got medium-length hair	4	2	4	2.4	2.5	
	Tends to be aggressive	6	3	2.8	3.1	3.2	
	Tends to show emotions	4	***19***	4.2	***2.2***	2.5	*lowest impact*
	Plays football	**14**	5	2.7	**3.6**	3	highest impact
Case 2 *Is a manager and likes war movies*	Argues by 'appeal to emotions'	6	5	2.4	3.6	3.2	
	Works in Public Administration	**8**	3	**4.1**	2.2	**2.0**	high plausibility with low risk
	Likes love stories	6	***18***	2.2	3.4	3.7	*low plausibility with high risk*
	Has particular interests, to the gender	6	1	3.6	3.4	2.9	
	Has got a high responsibility position, at work	**10**	9	**4.1**	2.1	**2.0**	high plausibility wigh low risk
Case 3 *Likes talking about private and can lift heavy weights*	**Practices yoga**	**11**	5	3.7	3.7	2.0	
	Tends to apply logical reasoning	5	3	3.2	3.4	2.4	
	Argues by 'appeal to emotions'	3	***14***	3.6	3.1	2.5	
	Has got a moderately developed muscular apparatus	6	6	4.0	3.0	2.5	
	Cares for children and gets money to family	**11**	***8***	3.8	3.1	3.0	

Table 2: 'Best' and 'worse' candidate moves and selection strategies applied by the 36 subjects in the study.

This attitude has two (not necessarily alternative) explanations: maybe there was not a unique 'best' or 'worse' candidate or maybe some of the subjects did not have a clear strategy in mind.

The candidates which were finally identified, in the three cases, as 'best' or 'worse' are shown in Table 2. Agreement among subjects is higher in Cases 1 and 2 than in Case 3. Contrary to MoT, the strategies applied seem to vary according to the 'influence goal' (Section 3.1).

- In Case 1, in which the player initially tends to believe that X is a woman, the candidate move is selected so as to maximize its *impact*: the influence goal is to change the player's belief about X's gender, even by mentioning facts with a low plausibility and a high risk;
- in Case 2, in which the player initially tends to believe that X is a man (a less dangerous state of mind, for the deceiver), the candidate move is selected so as to maximize its *plausibility* by keeping the *risk* low, at the same time;
- Case 3 is the situation in which the variability in the subjects' choice is highest. One of the moves *(Cares for children and provides money for the family*) is considered as 'best' and as 'worse' by a large proportion of subjects. A clearly prevailing strategy does not seem to exist.

5.4. Comments

The *measures of plausibility, impact and safety* defined in Mouth of Truth are reasonably compatible with those used by the subjects involved in the study. The variability in the estimates of the three variables may be explained in terms of what Tversky and Kahneman called *availability effect* in their famous work on how people assess the value of uncertain quantities. According to these authors, rather than applying probability theory and concepts, humans

tend to employ some heuristic principles which reduce the complexity of this task but sometimes lead them to systematic errors. For instance, they tend to assess the probability of events by the *ease with which instances or occurrences can be brought to mind* (that is, by their availability); reliance on availability may be biased, however, by the 'retrievability of instances' (Tversky and Kahneman, 1974). In our simplified Imitation Game, familiarity with people having unusual living habits may bias the estimates of plausibility, impact and risk of these situations. In addition to this confirmation, the main finding of our study concerned the *candidate selection strategy*. The strategy initially employed by MoT was revised so as to diversify it according to the influence goal: candidates with maximum impact (and with high plausibility and safety only if possible) are now selected when the Receiver's belief is close to what the message Sender wants to avoid; , on the contrary, candidates with high plausibility and safety are selected when the situation is less riskful (belief already close to the desirable). Still, we are not convinced that this strategy is ideal. One of the factors that characterize the conversation and playing style of individuals is the strategy they tend to adopt: knowing the opponent's preferred strategy is a condition which favours the winning of games or the discovering of deception attempts. So, a deception attempt might be also successfull if the strategy is varied during the dialog, so as to confound the Receiver.

6. Final Remarks

Although (as we said in the Introduction) deception simulation is a relatively new topic in human-computer interaction and multiagent systems, our study is built upon several previous experiences. Reasoning on other agents' beliefs and goals is an important research topic in User Modeling and Multiagent Systems. In particular, uncertain reasoning is required in a number of cases, and belief networks have been applied successfully to these cases. For example: student models (Conati et al, 2002), plan recognition (Horvitz and Paek, 1999), argumentation (Zukerman et al, 1999) and dynamic emotion activation (de Rosis et al, in press). As we stated in Section 3.1, deception is a special case of communication aimed at engendering belief in a particular proposition. It therefore shares several concepts and methods with communication (and in particular with persuasion). The problem of how to build a model of the interlocutor and how to select the most appropriate argument have been the object of study of argumentation systems. Several authors have discussed the meaning and proposed a measuring method for concepts like 'argumentation strength' (Sillince and Minors, 1991), 'probative weight', 'dialectical relevance' (Walton, 2000) or 'impact' (Zukerman, 2002). Others have studied how to define the 'level of activation' of belief about a proposition, for instance as a function of the number of times the proposition was assessed and the time elapsed since the last access (Zukerman, 2002). Much has been debated around the concept of plausibility, as well. For Quaresma and Lopez (1997), a property is plausible when the hypothetical world it brings to is non contradictory. Friedman and Halpern attach a generic meaning of uncertainty to this term (Friedman and Halpern, 1995) while, for Castelfranchi and Poggi (1993 and 1998), "*A plausible -to an agent A- fact is a 'credible' fact, to A, a fact that is 'acceptable' in A's world of beliefs because it is 'compatible' with them*". This definition considers the context in which assessment of plausibility is made by including, in the world of beliefs which contribute to evaluating compatibility, only those facts which are active when the given fact is examined. Dempster and Shafer introduced what is probably the most popular definition of plausibility; the authors start from a vision of 'partial knowledge' to define the plausibility of a fact p_h as the "*maximum value of credibility that p_h might assume to an agent A, should all other facts on which A has no knowledge reveal to be in favour of p_h.*" (Shafer, 1986; Barnett, 1991). In this vision, the degree in which R may believe in p_h lies in an interval between its 'belief function' and its 'plausibility'. Finally, Demolombe proposed a measure of *trust* in a modal logical framework, from which our definitions of source informativity do not differ greatly (Demolombe, 2001). Treating this problem probabilistically, as we

propose in this paper, has the advantage of grading all the measures mentioned, including source informativity. We are not completely convinced, however, that our probabilistic combination of informativity is cognitively plausible. When information received from various sources is conflicting, Receivers might decide to exclude some of them rather than combining all of them by applying various selection strategies: (i) believing the source that provides the most 'plausible' data, (ii) believing only the most 'informative' one, (iii) believing the most 'assertive' one or (iv) the most 'sceptical' one. The last two options correspond, respectively, to a personality trait of the Receiver (being 'trustful' or 'prudent') which we plan to include in our future models.

In this paper, in modeling how agents can deceive within a probabilistic framework for representing mental states, we challenged the so called 'sincerity assumption' in HCI and MAS. 'Deception' was distinguished from its special case of 'lie' and different forms of deception were characterized: for example, we discussed how one may deceive by avoiding to say anything or by saying the truth. Several criteria for deciding how to deceive were identified and, in particular, a model of information impact on the Receiver's mind was provided. Since the message Sender must plan its strategy so as to be believed, some of the criteria that S might apply in making this decision are discussed: content plausibility, communication safety and source informativity. A prototype is described for simulating deceptive strategies, in which the mental states of the Sender and the Receiver are represented as belief networks. Although all examples included regard a simplified version of Turing's Imitation Game, the tool is domain-independent; an application to a 'Murder Scenario' proposed by Zukerman (1999) is described, for instance, in (Carofiglio et al, 2001c). We feel that our work contributes meaningfully to deception simulation by clarifying the following aspects:

- how complex, but also interesting, is a cognitive model of deception and how useful it may be for simulating 'natural' dialogs;
- how appropriate is a probabilistic framework for modeling these notions and strategies. The advantage of belief networks over logical reasoning is, in particular, to enable simulation of the various kinds of deception that we have examined in this paper in a rather 'natural' way: active, passive, direct, indirect, on causes, on effects, 'by diversion' and so on.

The next step necessary to integrate our model within a dialog simulation framework, is to consider deception as a dynamic process. This problem was presented in Section 3, where we mentioned the 'horizon effect' in candidate selection strategies; a candidate move should be selected not only in terms of its immediate advantage, but also of its future potential effects. Viewing deception as a dynamic process requires the ability to extend strategies in time, by keeping a history of the dialog and adapting communication to past events (so as to avoid contradiction and repair possible failures) but also by modeling the possible suspicion of the Receiver more in depth than we did so far.

Acknowledgements

We thank the three anonymous reviewers for their careful reading and their insightful comments and Mrs Pauline Butts for her help in the revision of the manuscript.

References

J. A. Barnett, "Calculating Dempster-Shafer Plausibility". *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 13, 6, 1991.

D. Berry, I. Michas, T. Gillie, and S. Foster, "What do patients want to know about their medicines and what do doctors want to tell them? A comparative study". *Psychology and Health*, 12:467-480, 1998.

H. Brody, "The lie that heals; the ethics of giving placebos". *Annals of Internal Medicine*, 97, 1982.

D.B.Buller, J.K.Burgoon, A.Buslig, and J.Roiger, "Testing interpersonal deception theory; the language of interpersonal deception". *Communication Theory*, 6, 3, 1996.

J. K.Burgoon, D.B.Buller, K.Floyd and J.Grandpre, "Deceptive realities". *Communication Research*, 23, 6, 1996.

S.Carberry and L.Lambert, "A process model for recognising communicative acts and modeling negotiation subdialogs". *Computational Linguistics*, 25, 1999.

V.Carofiglio and F.de Rosis, "Ascribing and Weighting Beliefs in Deceptive Information Exchanges". *M. Bauer, P.J. Gmytrasiewicz, J.Vassileva (Eds.), User Modeling 2001, LNAI 2109, 222-224, Springer*, 2001a.

V.Carofiglio and F.de Rosis, "Exploiting Uncertainty and Incomplete Knowledge in Deceptive Argumentation". *V.N. Alexandrov, J.J. Dongarra, B.A. Juliano, R.S. Renner, C.J.K. Tan (Eds.), Computational Science - ICCS 2001, LNCS 2073, 1019-1028, Springer*, 2001b.

V.Carofiglio, F.de Rosis and R.Grassano, "An Interactive System for Generating Arguments in Deceptive Communication". *F. Esposito (Ed.), AI*IA 2001: Advances in Artificial Intelligence, LNAI 2175, 255-266, Springer*, 2001c.

C.Castelfranchi, and I.Poggi, "Lying as pretending to give information. Pretending to Communicate", *H. Parret (Ed), Springer Verlag*, 1993.

C.Castelfranchi and I.Poggi *Bugie, finzioni e sotterfugi*. Carocci Publ Co, 1998..

C.Castelfranchi, F.de Rosis and F.Grasso, "Deception and suspicion in medical interactions; towards the simulation of believable dialogues". *Machine Conversations, Y Wilks (Ed), Kluwer Academic Publishers*, 1999.

C.Castelfranchi, R.Falcone and F.de Rosis, "Deceiving in GOLEM: how to strategically pilfer help". *In Castelfranchi, C. and Tan, J-H (Eds): Trust and deception in virtual societies. Kluwer Academic Publishers*, 111-124, 2001.

P.R.Cohen, and H.J.Levesque, "Persistence, intention and commitment". *Intentions in Communication*. P.R.Cohen, J Morgan and M.A.Pollack (Eds), M.I.T. Press, 33-69, 1992.

C.Conati, A.Gertner and K.VanLehn, "Using Bayesian Networks to Manage Uncertainty in Student Modeling". User Modeling and User-Adapted Interaction. 12, 4, 371-417, 2002.

R.Demolombe and A.Jones, "Modal-logical specification of Security Requirements". *FAPR '96 Tutorial*, 1996.

R.Demolombe, "To trust information sources; a proposal for a modal logic framework". In Castelfranchi, C. and Tan, J-H (Eds): *Trust and deception in virtual societies*. Kluwer Academic Publishers, 111-124, 2001.

F.de Rosis, F.Grasso and D.Berry, "Refining instructional text generation after evaluation". *Artificial Intelligence in Medicine*, 17, 1999.

F.de Rosis, C.Castelfranchi and V.Carofiglio, "On various sources of uncertainty in modeling suspicion and how to treat them". *Workshop on Deception, Fraud and Trust in Agent Societies. Autonomous Agents*, 61-72, 2000.

F.De Rosis, C.Pelachaud, I.Poggi, V.Carofiglio and B.De Carolis. "From Greta's mind to her face. Modeling the dynamics of affective states in a Conversational Embodied Agent". *International Journal of Human-Computer Studies*, in press.

B.S.Firozabadi and A. J. I..Jones, "On the characterisation of a trusting agent; aspects of a formal approach". *Workshop on Deception, Fraud and Trust in Agent Societies. Autonomous Agents*. 1998.

N.Friedman and J.Y.Halpern, "Plausibility measures: A User's Guide". *Proc. Eleventh Conference on Uncertainty in AI*. 1995.

E.Horvitz and T.Paek, "A Computational Architecture for Conversation". *Proc. of the 7th Inthernational Conference on User Modeling.* Springer Wien, 201-210,1999.

A.Jameson, "Numerical uncertainty management in user and student modeling: an overview of systems and issues". *User Modeling and User-Adapted Interaction.* 11, 1-2, 129-158, 1995.

A.N.Kaplan and L.K.Schubert, "A computational model of belief". *Artificial Intelligence*, 120, 2000.

D.Kashy and B.DePaulo, "Who lies?". *Journal of Personality and Social Psychology.* 70, 5, 1996.

M.Lee and Y.Wilks, "Eliminating deception and mistaken belief to infer conversational implicature". *IJCAI Workshop on 'Collaboration, Cooperation and Conflict in Dialogue Systems*', 1997.

R.Lelouche and S.Doublait, “Qualitative reasoning with bluff and beliefs in a multi-actor environment”. *International Journal of Man-Machine Studies*, 36, 149-165, 1992.

R.E.Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algoritms.* Wiley-Interscience Publ, 1990.

T.Paek and E.Horvitz, “Conversation as Action Under Uncertainty”. *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, Stanford, CA, June, 2000.

J.Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufman Publishers, 2000.

P.Quaresma and J.Lopez, “Modeling credulousness and cooperativeness in a Logic Programming framework”. *IJCAI Workshop on 'Collaboration, Cooperation and Conflict in Dialogue Systems'.*

C.A.Reed, "The Role of Saliency in Generating Natural Language Arguments" in *Proc. of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, Morgan Kaufmann, Stockholm, 876-881, 1999.

E.Reiter and R.Dale, *Building natural language generation systems*. *Cambridge University Press*, 2000.

G.Shafer, “Probability Judgement in Artificial Intelligence”. *Uncertainty in Artificial Intelligence*. L. N. Kanal and J. F. Lemmer (Eds), Elsevier Science Publishers B. V. (North Holland), 4, 23-44, 1986.

J.A.A.Sillince and R.H.Minors, “What makes a strong argument? Emotions, highly-placed values and role-playing”. *Communication and Cognition*. 24, 3-4, 281-298, 1991.

D.J.Spiegelhalter, “Probabilistic Reasoning in Predictive Expert Systems”. *Uncertainty in Artificial Intelligence*. L. N. Kanal and J. F. Lemmer (Eds), Elsevier Science Publishers B. V. (North Holland), vol. 4, 47-68, 1986.

A.M.Turing, *Computing Machinery and Intelligence*. Mind, London, N.S., 59, 1950.

A.Tversky and D.Kahneman, “Judgement under uncertainty: heuristics and biases”. *Science*. 185, 1124-1131, 1974.

J.Vincent and C.Castelfranchi, “On the art of deception; how to lie while saying the truth. Possibilities and limitations of pragmatics”. H Parret, M Sbisà and J. Verschueren (Eds). *Studies in Language Companion Series*, 7. John Benjamin's, 1981.

D.N.Walton, “Dialectical relevance in persuasion dialogue”. Informal Logic. 19, 2-3, 119-143, 1999.

D.N.Walton, “Argumentation and theory of evidence”. *In New Trends in Criminal Investigations and Evidence*. Vol II. Antwerp, Intersentia. 711-732, 2000.

M.C.Weinstein and H.V.Fineberg. *Clinical Decision Analysis*. W.B.Saunders Company Press, 1980.

Y.Wilks, *Machine Conversations.* Kluwer Academic Publishers, 1999.

G.Zlotkin and J.Rosenschein, “Incomplete information and deception in multiagent negotiation”. *XII International Joint Conference on Artificial Intelligence*, Sydney, 1991.

I.Zukerman, N.Jinath, R..McConachy and S.George, “Recognizing intentions from rejoinders in a Bayesian Interactive Argumentation System”. *Proc. of PRICAI 2000*, 2000..

I.Zukerman and D.Litman, “Natural language processing and user modeling: synergies and limitations”. *User Modeling and User-Adapted Interaction*, 11, 1-2, 129-158, 2001.

I.Zukerman and S.George, “Towards a noise-tolerant, representation-independent mechanism for argument interpretation”. *Proceedings of COLING 2002. 1170-1176*, 2002.