Chapter 1

DECEIVING IN GOLEM: HOW TO STRATEGICALLY PILFER HELP

Cristiano Castelfranchi and Rino Falcone
Institute of Psychology,
National Research Council,
Rome, Italy
cris,falcone@ps.rm.cnr.it

Fiorella de Rosis
Department of Informatics,
University of Bari,
Italy
derosis@di.uniba.it

Abstract We first introduce GOLEM, a Multi-Agent world that we designed in order to study:
- different social attitudes and personalities in interaction and, in particular, attitudes in delegating tasks to other agents (search for help) and in adopting goals of others (help);
- reasoning about delegating and adopting tasks and about the other agents' personality.

We then shortly illustrate the different forms of deception that are allowed in this world. This enables us to provide some basic ontology of deception and to clarify the relationships between deception and lie, deception and mistake and so on.

In the main part of the paper, we focus on a family of examples of deception, which are particularly relevant in GOLEM: *social* and *intentional deceptions*, about *the agent itself* (its personality, its capabilities, its goals) rather than about the domain, and in order to obtain help (instrumental deception), which is, at present, the most important social motive for GOLEM agents. All these forms of deception are not necessarily based on implicit or explicit communication: that is, they do not necessarily imply lie.

Finally, we consider some more advanced forms of lie that cannot be treated in the current implementation of GOLEM but are very interesting for their *strategic* character: in these forms of lie, the deceiving agent predicts and exploits the reasoning that the deceived agent will perform as a consequence of her (false) speech acts (request, refusal, etc.).

The above mentioned discussion has several purposes:
- to clarify basic problems of interaction among cognitive agents;
- to show how even a toy world and a simple agent platform may be theoretically productive and
- to point out which reasoning forms should be formalized and implemented in multiagent systems

Keywords: Deception, Lie, Cooperation, Help, Personality, Agents

1. INTRODUCTION

In this paper, we investigate the role that deception may play in multiagent worlds, and the forms of deception that agents may apply in the different circumstances. To illustrate our theory, we take examples from GOLEM, a Multi-Agent world that was designed in order to study the following topics:

- social attitudes and personalities that may occur in interaction and, in particular, attitudes in delegating tasks to other agents (search for help) and in adopting goals of others (help);
- forms of reasoning that agents may apply in delegating and adopting tasks and in inferring the other agents' personality.

We will briefly present the different forms of deception that are allowed in GOLEM: this will enable us to provide some basic ontology of deception and to clarify the relationships between deception and lie, deception and mistake and so on. Then, in the main part of the paper, we will focus on a family of examples of deception which are particularly relevant in GOLEM: social and intentional deceptions, about the agent (its personality, its capabilities, its goals etc) rather than about the domain, and aimed at obtaining help (instrumental deception): this is, at present, the most important social motive for GOLEM agents. All the mentioned forms of deception are not necessarily based on implicit or explicit communication, that is, they do not necessarily imply lie. We will finally consider some more advanced forms of lie, that we do not yet treat in the current implementation of GOLEM, although they are very interesting for their strategic character. In these forms of lie, the deceiving agent (that we will call Eve) predicts and exploits the reasoning that the deceived agent (that we will call Adam) will perform as a consequence of her (false) speech act (request, refusal, etc.). In the final Section of the paper, we will introduce exchange relations and personalities and will explore the role of deception in exchange, by showing that exchange can be seen as a complex form of delegation and adoption.

2. GOLEM: PERSONALITIES AROUND BLOCKS

GOLEM is a system aimed at formalising, implementing and experimenting different kinds and levels of social cooperation, represented by different social attitudes or personality traits. There are, in our view, two basic elements of any MultiAgent cooperative activity: Delegation and Adoption. Informally: in delegation, an agent A needs some action that another agent B may perform, and includes it in its own plan. In adoption, an agent B has a goal since and until this is the goal of another agent A. By analyzing types and levels of delegation and adoption, we come to define the agent's level of autonomy and of cooperativeness (Castelfranchi and Falcone, 1997). Designing a 'supporting' agent then equates to defining several aspects of its behaviour: (i) the circumstances in which the agent will activate itself, (ii) whether it will literally follow the received request or will interpret it to abduce what it is expected to do in the specific circumstance, (iii) whether it will respond critically to this request, and so on. These design features may be seen in terms of 'type of delegation' given to the agent and 'type of help' the agent is programmed to offer. In GOLEM, we formalize these levels of Delegation and Adoption in terms of 'personality traits' or 'attitudes' attributed to agents.

A personality trait of a GOLEM agent is modelled as a logically and cognitively consistent combination of reasoning rules. Different traits may affect the delegation and help attitudes: an agent's personality is seen as a plausible combination of a delegation and an help attitude; a set of reasoning rules that correspond to the two traits is attributed to the agent's mental state, accordingly. GOLEM includes, at present, only two agents, that play in a domain; each of them tries to achieve its own goal; the two agents' goals may be compatible or conflicting. Agents play 'in turns': at each turn, an agent may perform some domain action or may 'pass the hand' without modifying the domain; it may, at the same time, perform a 'communicative act' according to a defined protocol. When an agent achieves its end-goal, the result is frozen: that is, it cannot be destroyed by the other agent.

Agents may have a limited view of the domain and an incomplete knowledge of each other, with possibly wrong beliefs (see Figure 1).

This means that agents cannot directly access knowledge about personality, goals, plans, abilities and beliefs of other agents: this information is ‘private’ (Lelouche and Doublait, 1992); they have to infer it by observing the behaviour of the other agent or by trusting its direct communication. In addition, our agents have limited and different views of the domain.

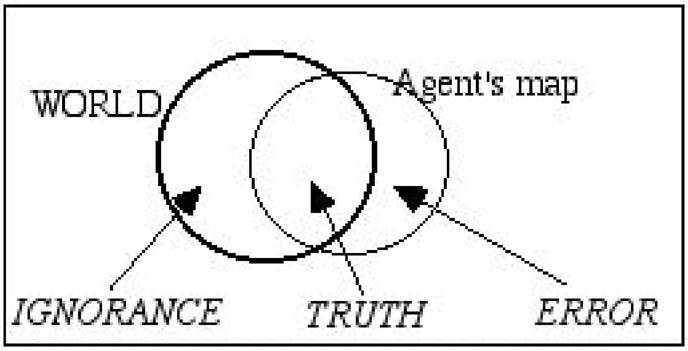

Figure 1.1 GOLEM agents’ knowledge about the domain

The user can set the conditions of simulation of a game in a particular domain and can follow how the game proceeds, through a graphical interface. At present, the domain of game is a "blocks world" including only four small and four big blocks; however, agents are built separately from the application domain: their behaviour is therefore described in a domain-independent way, so that games in any domain may be simulated. Table 1 describes some of the domain states that we will consider in our examples.

A particular simulation starts by selecting a 'world' (that is, a couple of agents with defined mental states) and by setting the initial domain state and the agent which moves first. The two agents introduce themselves by declaring their personalities and abilities; in this introduction, they may give partially incorrect or imprecise descriptions, or may even lie about them. For example: an agent might omit the description of its abilities, might say that it is able to 'make towers' without specifying whether they are big or small, or might tell that it is not able to do some action whereas it can do it. This self introduction is employed by the other agent to build an image of its partner’s mental state. The user can follow the progress of the play by means of a domain window, which shows a graphical representation of the domain state, with actions made at each turn; a dialogue window shows, as well, the communicative acts that the two agents exchange, translated into natural language.

State Name	*State Description*	*Symbolic Description*
Big-building	Two overlapping big blocks	(ON A B),(ON B C),CLEAR (A)
Small-building	Two overlapping small blocks	(ON a b),(ON b c),(CLEAR a)
Big-tower	Four overlapping big blocks	(ON A B),(ON B C),(ON C D), (TABLE D),(CLEAR A)
Small-tower	Four overlapping small blocks	(ON a b),(ON b c),(ON c d), (TABLE d),(CLEAR a)
Twin-tower	A big-Tower + a small Tower	(ON A B),(ON B C),(ON C D), (TABLE D),(CLEAR A),(ON a b), (ON b c),(ON c d),(TABLE d), (CLEAR a)
Bell-tower	A big-Tower with a small block (''a pinnacle") on top on it	(ON A B),(ON B C),(ON C D), (ON A a) (TABLE D),(CLEAR a)
Stock	All blocks on the table	(TABLE A),(TABLE B),(TABLE C),(TABLE D) (CLEAR A),(CLEAR B),(CLEAR C),(CLEAR D) (TABLE a),(TABLE b),(TABLE c),(TABLE d) (CLEAR a),(CLEAR b),(CLEAR c), (CLEAR d)

Table 1.1 some states in the blocks-world in GOLEM

Though being quite a simple world, GOLEM is very good for reflecting about deception and for simulating various forms of it. In fact, a rich platform for studying deception requires the following features :

- several agents,
- with limited and possibly wrong "knowledge",
- interacting and practically interfering with each other in a common domain,
- with partially different views of the domain,
- not invisible to each other (i.e. they can "see"/know when the other agent does a given action),
- with goals and intentions,
- with a representation of the mind of the other agent,

- endowed with some specialised communication channel and able to perform communication acts about the domain and -possibly- about themselves.

GOLEM has all these features.

3. A WORLD OF DECEPTIONS

Let us start, first of all, with a definition of what we intend as 'deception'. Some of the definitions of deception that may be found in the AI literature are, in our view, quite imprecise. For example, according to Lee and Wilks (1997): "An act of deception occurs when a speech act is performed intentionally to comunicate a false proposition". This is, actually, the definition of 'lie' (or at least of 'deceptive speech act'), which is only a sub-case of the category of deceptions; this definition is acceptable in the context of a paper about conversation and speech acts, but has not a general value. In addition, the definition is unclear because of the very vague assertion that "x intends to communicate a 'false' proposition". From which point of view is the proposition 'false'? If the proposition is 'objectively' false, as it seems from the text:

$$\textit{(Intends x (Believes y p))} \wedge \textit{Not p}$$

this looks like a wrong definition of 'intentional' deception: x might just be wrong by saying that she believes p (while p is false), without -therefore- having the intention to deceive.

We clearly prefer the "subjective" definition of deception that dates back to Augustinus (388-396 A.D.). For example:

$$\textit{(Intends x (Believes y p))} \wedge \textit{(Believes x (Not p)).}$$

Notice that, given this definition, it may even be that

$$\textit{(Intends x (Believes y p))} \wedge \textit{(Believes x (Not p))} \wedge \textit{p;}$$

in this case, x's act is an act of deception, although y is not deceived as for the "reality"!

To show how rich and complex is the concept of deception, we now classify the forms of deception that are allowed in our experimental micro-world. In GOLEM, deception may take the following forms:

a. Deception about the agents and about the 'physical domain'.

As we said, our agents have limited and different views of the domain: there are parts of the domain that they cannot see while another agent can: without this, any deception about the domain would be impossible. Limited and different views allow the agents to be also "sources" of information about the domain, not only about themselves.

b. Intentional, praeterintentional and accidental deception.

GOLEM agents are in fact intentional and planning agents: thus, they may 'intend' that the other agent believes or does not believe something; they can intentionally deceive the others. This happens when they have the goal that the other agent believes p while they personally believe not p (strong deception: see Augustinus, as already cited) or at least do not believe p (weak deception).

Let us call Eve (E) and Adam (A) the deceiving agent and its interlocutor. Let us denote, with GE g and BE p, respectively, Eve's goal about g and belief about p (and the same for Adam). Notice that g may denote a domain-state or some aspect of the other agent's mental state. Strong and weak deceptions may then be formalised, respectively, as follows:

$$G_E B_A\ p\ B_E \neg\ p\ (\vee\ G_E B_A\ p\ \wedge \neg B_E\ p).$$

Deception also holds when Eve has the goal that Adam does not believe p, while she personally believes it: $G_E \neg B_A\ p\ \wedge B_E\ p$ (see later).

However, GOLEM agents do not know (and therefore do not intend) all the effects of their actions. For example: they don't know how exactly the other agent will interpret their action; so, they might accidentally deceive the other agent. For example: Eve may ask Adam to perform some action without understanding that Adam will abduce, from this request, some wrong belief about her delegating personality (praeterintentional deception). Our agents may even know some effects of their actions without necessarily intending them; unpremeditated deception (or 'deception as a known side-effect') is therefore possible.

c. Social (due to other agents, induced) and non social deception (due to limited or wrong knowledge of a single agent). Social deception may be of two kinds: (a) intentionally planned deception about the domain, that may be exploited by another intentional agent, or (b) deception (may be accidental) about another intentional agent. Non social deception is the accidental deception of an agent about the domain, due to its limited or wrong perception or interpretation of the domain, or to its wrong or limited reasoning capability. The domain, in this case, deceives the agent.

An example of non-social deception: Eve's belief that (CLEAR B) is false because she ignores that (ON A B); she plans, on such a basis, to "put C on B" and fails.

An example of social deception of type (a): Eve knows that (CLEAR B); Adam hiddenly changes the world by putting A on B, in order to deceive Eve and to induce her to fail. By doing this, he induces Eve to believe something wrong. Notice that Adam induces a wrong belief in Eve by acting on the world, not directly on her beliefs!

An example of social deception of type (b): Eve believes that Adam has the goal that (ON A B), while Adam's goal is different.

d. Passive deception: the deceiving agent simply lets the other agent ignore something crucial for him, or believe something wrong. Active deception: the deceiving agent does something in order to induce the other agent to ignore something that is crucial for him, or to believe something that is wrong.

e. Deception by ignorance: the deceived agent ignores something crucial for him; for example: concealment, omission, etc., as in a "hidden letter" (Zlotkin and Rosenshein, 1991). Deception by falsity: the deceived agent believes something importantly false; for example: lies, bluff, etc., as in a "phantom letter" (same citation).

f. Deception with communication (ex. lies and bluff) and without communication.

Not every social deception requires communication. Let us look, for example, at Table 2: it might be quite difficult to call 'communication' the fact that Eve does nothing at all to obtain that Adam continues to ignore something (D), or continues to believe something wrong (C). It is also difficult to extend the notion of communication to the case in which Eve does something to hidden something else to Adam (B)! So, deception by communication seems to be limited to the box (A)[1] . GOLEM agents can deceive by hiddenly doing something on the domain (for instance, moving a block) or through communicative acts (see later).

	Mistake or Falsehood $G_E B_A\ p \wedge B_E \neg p$	*Ignorance* $G_E \neg B_A\ p \wedge B_E\ p$
Eve does something to obtain that Adam believes p, or to avoid that Adam believes p	A to falsifie, to fake or to lie about p	B to hidden p
Eve lets Adam (not) believe that p	to leave Adam be wrong C	to leave Adam in his ignorance D

Table 1.2 forms of deception in GOLEM

g. Deception *through implicit, behavioural communication* (ex. bluff) and by *explicit* and specialised communication acts - like speech acts - (ex. lies).

As we said in Section 2, GOLEM agents have some communication protocols: they can inform each other about domain facts or about their abilities; they can make a request, a refusal and so on. They can lie by deliberately giving false information with their communicative acts. But they are able, as well, to interpret the other agents' behaviour (including communicative acts) and to apply abductive reasoning on these findings. Thus, an agent can strategically exploit the interpretation of an action that, presumably, the other agent will do; it can "plan for" the understanding or misunderstanding of the other; it can perform a given action in order to obtain that the other agent gives a particular interpretation to this action. This means using that action as an implicit behavioural communication. So, GOLEM agents can recur to mimicry, fiction and bluff.

In this paper, we will focus on some forms of deception and on limited examples, in order to support the claim that deception is useful for delegating tasks (and then exploiting other agents' help). We will make examples of deception about the agent's 'internal', private information, and in particular its personality, capabilities, and plans, since agents' personalities play an important role in GOLEM (Castelfranchi and Falcone, 1997; Castelfranchi, de Rosis, Falcone and Pizzutilo, 1997). Thus, we will not argue about a general theory of deception (Vincent and Castelfranchi, 1981; Castelfranchi and Poggi, 1994; Galliers, 1988; Mitchell amd Thompson, 1986; Wiseman, 1996), nor will we examine all the forms of deception that are allowed in GOLEM.

4. WHY TO DECEIVE THE OTHER AND ABOUT WHAT

In GOLEM, agents pursue their goal about the blocks world, delegate some tasks to the other agent (for different reasons, depending on their personalities) and possibly help the other, by accepting the delegated tasks[2].

The delegation personalities -and then the reasons why an agent delegates- are the following:

- a lazy agent always delegates tasks if there is another agent which is able to take care of them; it acts by itself only when there is no alternative;
- a hanger-on never acts by itself;
- 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.

The help personalities -and then the reasons why an agent helps- are the following:

- a hyper-cooperative always helps if it can;
- a benevolent first checks that the request does not conflict with its goal;
- a supplier first checks that the other agent could not do the action by itself;
- a non-helper never helps, by principle;
- a literal helper restricts itself to considering the requested action;
- a overhelper goes beyond this request, to hypothesise the delegating agent's higher order goal, and helps accordingly;
- a subhelper performs only a subpart of the requested plan;
- a critical helper modifies the delegated plan.

Given this framework, it is clear that not all building or delegation attempts are successful, because they also depend on the other agent's decision; to avoid failures, agents must then become more strategic.

Till now, we excluded, in GOLEM, deceptions 'per se', for example for funny or for naughtiness, as we exclude wicked personalities or chronic deceivers. Thus, in GOLEM, deception is only instrumental to other goals: let's say, 'rational'. In instrumental deception, there are two general possible motives:

a) if Adam would know (believe) that p, there would be some negative consequence to Eve (contrary to Eve's goals). For example: Adam would do (or not do) a given action that Eve wants to avoid (or requires);

b) if Adam would believe that p, there would be some positive consequence to Eve (corresponding to Eve's goals). For example: Adam would do (or not do) a given action that Eve requires (or wants to avoid).

The problem is still more complex than this. To be more complete, one should consider several combinations of the two agents' beliefs. Given the framework (a), that is $G_A \neg B_A$ p, there are, in fact, several strategies that Eve might apply to achieve her goal, as shown in Table3. Only some of these situations are deceptive, and we will focus on the most typical ones. Given the framework b), there will be a similar table of possibilities.

	$B_E p$	$B_E \neg p$	$\neg B_E p$	$\neg B_E \neg p$	$\neg W_E p$
$B_E B_A p$	active deception	conviction	conviction	active deception	active deception
$B_E B_A \neg p$	passive deception	OK	OK	passive deception	passive deception
$B_E \neg B_A p$	passive deception	OK	OK	passive deception	passive deception
$B_E \neg B_A \neg p$	passive deception or attive deception	weak conviction	weak conviction	active deception	active deception
$B_E B W_A p$	passive deception or attive deception	conviction or OK	conviction or OK	passive or active deception	passive or active deception
$B_E \neg B W_A p$	passive deception or attive deception	weak conviction or OK	weak conviction or OK	passive detection	active detection

Table 1.3 a more general framework of deception types

To summarise: in GOLEM, deception is mainly due to conflicts. There are two main kinds of conflicts:

- conflicts between the domain plan/goals of the agents: if Eve achieves her goal, Adam cannot achieve his own goal. For example: the domain-goal of Eve is bell-tower , whereas the domain-goal of Adam is twin-tower[3]
- conflicts between the social goals, in particular between Eve's intention that Adam does some domain action (delegation) and Adam's possible decision of not doing it (refusal).

In order to avoid failure in achieving their goals, agents may recur to deception. We will examine in detail a specific case of deception from the motivational point of view.

5. DECEPTION FOR OBTAINING HELP

If the other agent knows that p, it will not adopt my request (Zlotkin and Rosenshein, 1981). This motive-based deception will be crossed with three content-based cases:

- deception about capabilities
- deception about personality
- deception about goals and plans.

Let us see some examples of these combinations.

5.1 DECEPTION ABOUT CAPABILITIES

Eve is a 'lazy', and she wants to delegate to Adam an action a that she is able to perform. Eve knows that Adam is a 'supplier': thus, if he knows Eve's abilities, he will refuse to adopt her goal. So Eve, in order to induce Adam to accept the request and help her, needs that Adam ignores that she is able to do the delegated action. Given this intention of deceiving, there are several possibilities:

1. Adam believes that Eve cannot do p: Eve leaves Adam in his error and exploits it. Notice that Adam can have such a wrong belief about Eve if she never performed that action in front of him, or if she performed it in some part of the domain that is hidden to him. There is also another interesting possibility based on personality: Adam wrongly believes that Eve is a 'delegating-by-necessity'; thus, from Eve's delegation (and ignoring Eve's capacity) he infers Eve's inability. In this case, Eve may let Adam be wrong about both her personality and her capabilities.

2. Adam believes that Eve can do p: Eve has to change Adam's beliefs. She can directly declare that she is not able to do a (and hope that Adam will believe her), or can convince Adam that she is a 'delegating-by-necessity': this implies that she is not able to do what she delegates. All these are 'lies'. Eventually, she can act 'as if' she was not able to do a, by simulating her inability. For example, if Adam believes that she is a 'delegating-by-necessity', she can simply delegate a and hope that, after observing this action, he will revise his belief about her ability rather than his belief about her personality. If, on the contrary, Adam believes that Eve is a 'lazy', she will have to renounce to do the action after Adam refusal, in order to obtain that Adam revises his belief about her abilities: in this case, she will possibly ask him again to perform the action, later on.

3. Adam has no idea about Eve's capacity. Eve can use the same strategy mentioned in 2.

We may represent the described situation in our agent implementation language[4] , as follows:

i) (B_E (Lazy Eve)),

ii) (B_E (Supplier Adam)),

iii) (G_E g),

iv) (B_E (Achieve a g))

v) (G_E (Done a)) follows from (iii) and (iv)

vi) (B_E($CanDo_E$ a))

vii) (B_E($CanDo_A$ a))

viii) ($B_E \neg$(BW_A($CanDo_E$ a)))

ix) (B_E rule-lazy)

x) (B_E rule-supplier)

xi) (G_E($IntToDo_A$ a) follows from (i, v, vii, ix)

xii) ($Delegates_E$ Adam a) follows from (xi)

xiii) (G_E($B_A \neg$($CanDo_E$ a))) from (xi, xii, x)

where:

rule-lazy is:
given that X is a lazy:
if (B_X($CanDo_Y$ a)) and (G_X ($Done\ a$))) then (G_X ($IntToDo_Y$ a))

and

rule-supplier is:

given that Y is a supplier and that ($Delegates_X$ Y a),
if (B_Y($CanDo_X$ a)) then $\neg$(G_Y($IntToDo_Y\ a$))
Given (vi), Eve's goal in (xiii) is a deception goal.

In summary: in order to induce Adam to believe that $\neg$($CanDo_E$ a), Eve can:

- declare this explicitly;
- induce Adam to believe that she is a 'delegating-by-necessity' and delegate a;
- induce Adam to believe that she is 'lazy' and avoid doing a after a refusal;
- induce Adam to believe that she is 'never-delegating' and never do a, even if needed;

In any case, she must hidden to Adam her ability about a, by not performing a in a 'public' area.

5.2 DECEPTION ABOUT PERSONALITY

Let us suppose that, if Adam recognises the real personality of Eve, he will not help her. We already saw an example of this situation when we said that Adam wrongly believed that Eve was a 'delegating-by-necessity' and that, from Eve's delegation (and by ignoring Eve's capacity) he inferred Eve's inability. In that case, Eve left Adam be wrong about both her personality and her capabilities.

Let us now introduce a different example, based on a specific attitude of Adam in front of some particular social personalities. Let us suppose that Adam is not a 'supplier' (that is, someone who doesn't help agents that can do by themselves), but that he is against 'lazy' or 'hanger-on' people. In fact, in GOLEM we have the possibility not only of rigid and invariant personalities but also of flexible social 'attitudes' that vary according to the personality or the attitude of the other agent. Then, an attitude is a trait conditional to circumstance (Castelfranchi, Falcone, de Rosis and Pizzutilo, 1998). For example, the agent is not always a lazy or a supplier but it assumes such an attitude depending on the personality or the attitude of the other agent: if Adam is hyper-cooperative or benevolent, then Eve assumes the lazy attitude; if Eve is delegating-if-needed, then Adam assumes the supplier attitude, and so on.

To go back to our example, let us suppose that Adam does not help 'on principle' certain kinds of agents: for example 'lazy' agents or agents that 'love towers'. If he knows that Eve has such a personality, he will reject any request she does. Thus, Eve has to deceive Adam about her personality. Let us assume that she is in fact a 'lazy'; in order to mislead Adam, she might perform some action without trying to delegate it. Of course, this presumes that her 'lazy' trait is not rigid and permanent (otherwise she could not believe differently), but is just a preference, a tendential attitude.

5.3 DECEPTION ABOUT GOALS AND PLANS

In GOLEM, Adam refuses to help Eve if he believes that her domain-goal is in conflict with his own goal. The only exception is when Adam is a 'slave' or a 'hyper-cooperative', that is, someone who helps Eve even if this is against his own interests, provided he believes that Eve cannot do the requested action by herself. Let us suppose, for instance, that Eve wants to come to have a twin-tower: if she knows that Adam's goal is bell-tower, she will foresee that Adam will refuse to do a big tower for her, if he comes to know her goal. She might then declare, generically, that she likes 'towers' (by leaving Adam in his ignorance) or that her final goal is big-tower (by lying); she might, as well, make a small building so as to induce, in Adam, a false hypothesis about her real goal.

6. STRATEGIC LIES

Let us now say something more about lie, i.e. deception through speech acts. Currently, lies in GOLEM are quite limited. As we saw in the previous Section, our agents may lie about their personality, capabilities, or goals; however, they are not allowed to use deceptively their speech acts or protocols. In particular, they cannot lie with requests, with refusals or with acceptances. These lies would be possible in principle; however, they require quite a complex strategic reasoning, that is beyond the capabilities of the current implementation of our system. More precisely, in order to instrumentally and rationally use such a lie, Eve should reason about the reasoning of Adam in response to her speech act, and about the consequent behaviour that Adam would adopt. Let us give some concrete examples of these kinds of lie, since they are quite interesting, although not yet included in the current system.

6.1 FALSE REQUEST

We have a "false request" *when an agent does not really desire that the other agent performs the requested action: it rather intends other effects of its act.* An example:

Eve and Adam have conflicting goals: Eve's goal is 'bell-tower' while Adam's goal is 'twin-tower'. The current domain state is 'stock'.

Eve knows Adam's goal; she also knows that Adam believes that she is 'delegating by necessity', and that he knows her goal. She is not able to move big blocks (Adam might ignore this or not), so she needs that Adam builds the big-tower in order to add the pinnacle on top of it. However, Eve knows that Adam will never build the big-tower before building the small-tower, because he wants to avoid that Eve completes her bell-tower and wins.

Eve moves first. She asks Adam to perform some action involving small blocks, in order to induce him to believe that she is not able to move them (remember that she is a "delegating by necessity", in Adam's view). Notice that she does not really need this action: she just cares about Adam's belief about her capability! Thus, her request is fictitious: she does not matter whether Adam really performs it or not.

Adam refuses to make the requested action, and examines his own plan. Since he now believes that Eve cannot move small blocks, he is not afraid that she will complete his big-tower with a pinnacle, so as to obtain a bell-tower. He therefore builds the big-tower ....

At the next turn, Eve wins, by putting a small block on top of it!!!

6.2 FALSE REFUSAL

We have a "false refusal" when an agent refuses to perform the requested action while in fact it does it (for its own plan). Of course, this form of deception presumes that the other agent cannot "see" the action, either because it does not look at the domain (it is 'distracted') or because the action occurs out of its view, in a hidden place.

Why should this deception be useful? For several reasons: Eve might deceive Adam about her capabilities or her helping personality, or might avoid Adam's interference in her domain plan[5]. Let us see an example of this last case.

Eve's goal is bell tower; she is not able to move big blocks; Adam's goal is twin tower. Eve asks Adam to make a big tower; Adam knows Eve's goal and ability and understands that, if he does the big tower and Eve knows it, she will add a pinnacle and will win (here, again, Adam's and Eve's final goals are in conflict). Thus, Adam responds negatively to Eve's request, so that Eve believes that there is no big tower in the domain; at the same time, he builds a big tower and plans to later build also a small tower (and to win).

6.3 FALSE ACCEPTANCE

We have a "false acceptance" when the agent accepts to perform the requested action while in fact he does not intend to perform it. Of course, to be a deception, this presumes that the other agent cannot "see" the action (again, because it does not look at the domain or because the action occurs out of its view, in a hidden place). Why should this deception be useful to Eve? Once again, for several reasons. She might deceive Adam about her capabilities, and let him believe that she is able to perform the action; she might deceive him about her helping personality or attitude; or she might prevent him from interfering in her domain plan[6].

Let's see an example of this. Eve's goal is bell tower; she is lazy; Adam's goal is twin tower; the domain state is big-tower. Eve asks Adam to put a small block on top of the big tower; Adam knows Eve's goal and ability of moving small blocks, and understands that, if he refuses, Eve will succeed by herself. Thus he accepts to help her, and induces Eve to believe that her goal will be achieved; however, he makes a small tower instead, ... and wins!

7. CONCLUSIONS

A few final considerations. Our analysis of the various forms and types of deceptions shows, we believe, that, to apply this form of social action, agents must have some image of the other agent's mind and be able to reason on that image, in several ways. To decide whether and how to deceive Adam, Eve needs to have second-order beliefs and goals in her mental state (Cohen and Levesque, 1990; Taylor, Carletta and Mell(Cohen and Levesque, 1990; Taylor, Carletta and Mellish, 1996); she needs, as well, to imagine how Adam will interpret her action or communication by 'simulating', in some cases, the reasoning that he will perform on it. This form of reasoning is often uncertain: we saw some examples (for instance, in Section 5.1.2.) in which Eve thought that Adam might interpret her communication in two distinct ways, one of them being more 'convenient' to her goals. In other cases, we saw that Eve could select the form of deception to apply according also to the most likely (and most favourable) interpretation that she though Adam would give to her communication. Probability and convenience of interpretations are therefore important parameters to consider in selecting among alternative forms of deception: in (de Rosis, Covino, Falcone and Castelfranchi, in press) we describe how this problem might be formalised.

In a future step of this research, exchange relations and personalities will be introduced and deception in and for exchange will be explored; consider that exchange is just a more complex form of delegation and adoption. Exchange-delegation attitudes imply, in fact, that Eve delegates some action only if she believes that Adam will adopt it because he intends, in his turn, to delegate some other action in the future, and believes that Eve will adopt his delegation. The exchange-adoption attitude consists in adopting only if one believes that the other agent will adopt, in its turn, the delegated action. On such a basis, complex forms of deception can arise, about willingness, capability, etc. in order to obtain adoption. Some more work on GOLEM is needed not only for modelling strategic lies, flexible attitudes and exchange attitudes and deception, but also for modelling reasoning and decision about whether or not to deceive and about the deceptive strategy to adopt. Finally, some more work is needed, as well, to introduce in this system diffidence and ability to recognise or suspect deception.

Notes

1. We claim that also box (A) is not entirely based on communication. However, this argument is out of the focus of this paper.

2. In the current version of GOLEM, there is no spontaneous help: that is, agents do not deliberately help each other without any request.

3. These goals are incompatible, because the small-tower requires all the small blocks available in the domain, while also the bell-tower requires one small block (as a pinnacle); and because the big-tower is no more a big-tower when a pinnacle is added on top of it.

4. With the following notation:
g := domain-state; a := domain-action; X, Y := agent-names (E := Eve, A := Adam);
(Achieve a g) := "the action a enables achieving the state g"
(Done a) := "the domain-state, after a has been performed"
(CanDoX a) := "X is able to perform a"
(IntToDoX a) := "X intends to perform a"
(DelegatesX Y a) := "X delegates a to Y".

5. Or for other strange social goals, like "to make him a surprise".

6. Or to make him fail in his action and plan.

References

Agostinus, (388-396 A.D.), De mendacio.

Castelfranchi Cristiano and Falcone Rino (1997), Delegation Conflicts, in M. Boman & W. Van de Velde (eds.) Multi-Agent Rationality, Lecture Notes in Artificial Intelligence, 1237. Springer-Verlag pg.234-254.

Castelfranchi Cristiano, de Rosis Fiorella, Falcone Rino (1997). Social Attitudes and Personalities in Agents. In Socially Intelligent Agents, AAAI Fall Symposium Series 1997, MIT in Cambridge, Massachusetts, November 8-10.

Castelfranchi Cristiano, de Rosis Fiorella, Falcone Rino and Pizzutilo Salvatore (1997). A Testbed for investigating personality-based multiagent cooperation, Proceedings of the Symposium on Logical Approaches to Agent Modeling and Design, Aix-en-Provence, August, 18-22.

Castelfranchi Cristiano and Poggi Isabella (1993). Lying as Pretending to Give Information, in H. Parret (ed.) Pretending to Communicate, Verlag, Berlin.

Castelfranchi Cristiano, de Rosis Fiorella, Falcone Rino and Pizzutilo Salvatore (1998). Personality traits and social attitudes in Multi-Agent Cooperation. Applied Artificial Intelligence. 12, 7-8, 649-676.

Cohen Phil R. and Levesque Hector J. (1990). In Intentions in Communication, P.R Cohen, J. Morgan, M.A. Pollack (eds), 33-71. Cambridge, Mass.: MIT Press.

De Rosis Fiorella, Covino Emanuele, Falcone Rino and Castelfranchi Cristiano (in press). Bayesian cognitive diagnosis in believable multiagent systems. In 'Frontiers in Belief Revisions', M A Williams and H Rott (Eds), Kluwer Academic Press.

Galliers Julia R. (1988). A strategic framework for multi-agent cooperative dialogue', in Proceedings of the 8th European Conference on Artificial Intelligence, London: Pitman, 415-20.

Lee Mark and Wilks Yorik (1997). Eliminating deceptions and mistaken belief to infer conversational implicature, IJCAI'97 workshop on "Collaboration, Cooperation and Conflict in Dialogue Systems", August 25, 1997, Nagoya, Japan.
(http://www.cs.umd.edu/~traum/CCCinDS/Papers/Mark_Lee.ps)

Lelouche R. and Doublait S. (1992). Qualitative reasoning with bluff and beliefs in a multi-actor environment. Int. J. Man-Machine Studies, 36, 149-65.

Mitchell, R.W. e Thompson, N.S. (Eds.) (1986) Deception: Perspectives on Human and Nonhuman Deceit. Albany, N.Y.: SUNY.

Taylor, J. A, Carletta, J. and Mellish, C.(1996). Requirements for belief models in cooperative dialogue. User Modeling and User-Adapted Interaction, 6,1, 23-68.

Vincent Jocelin and Castelfranchi Cristiano (1981). On the art of deception: How to lie while saying the truth. In H. Parret, M. Sbisa' and J. Verschueren (Eds.), Possibilities and Limitations of Pragmatics. Studies In Language Companion Serles (Vol. 7). Amsterdam: John Benjamins.

Wiseman, R. (1996). Towards a psychology of deception, The Psychologist, 9 (2), 61-64.

Zlotkin Gilad and Rosenschein Jeffrey (1991). Incomplete information and deception in multi-agent negotiation. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence, 225-231, Sydney, Australia.