Modelling Conflict-Resolution Dialogues

Fiorella de Rosis, Floriana Grasso, Cristiano Castelfranchi, and Isabella Poggi

1. Introduction

In dialogue modelling research, human-computer or human-human dialogues are simulated by applying experience from human-human interaction analysis (Carletta et al., 1997). This has led to several dialoguing agent architectures (among others, the Belief-Desire-Intention model in Rao and Georgeff, 1991), where a cognitive model of interacting agents is built up, and a conversational strategy is assigned to each of the agents, in the form of either an algorithm or a set of decision rules. However, knowledge representation and treatment methods in these models usually privilege purely rational aspects of reasoning, whereas several studies have suggested that dialogue is strongly influenced by the relationships among participants, their role, their personality and their emotional status (Elliott and Ortony, 1992).

The need for giving more weight to emotional and personality aspects is emerging as a new and promising research objective (Cañamero, 1998, de Rosis, 1999, Prevost and Churchill, 1998), aimed at modelling *believable* agents. A believable behaviour would be shown by an agent which avoids answering questions which it is not able or willing to answer, is able to react to unexpected questions or attitudes of other agents, and assigns weights to its goals and beliefs according also to its emotional state, and not only to its knowledge (Reilly and Bates, 1995).

Emotional and personality aspects of the mental state play a particularly relevant role when agents have to face a conflicting situation: *impatient, aggressive, polite, persistent, rude*... people show radically different ways of managing conflicts with others.

In this work, we investigate how these aspects can be taken into account in the modelling of conflict-resolution dialogues. We start by examining (in Sect. 2. and 3.) the types of conflict that may exist between two interacting agents, and how conflict-resolution dialogues may be undertaken to try and solve them. To support our analysis, we consider, in Sect. 4., two examples of doctor-patient dialogues, in which several kinds and forms of conflict exist, which are solved only in part during the dialogues. We then analyse (in Sect. 5.) the reasoning process that the agents perform throughout such dialogues, and show how the sequencing of reasoning steps and the way each of them is performed are linked to personality factors. In Sect. 6., we briefly illustrate Xanthippe, a multiagent dialogue simulation system that was designed to test our theories, and we finally reflect (in Sect. 7.) on the interest and the limits of our approach.

2. Conflicts: Nature and Kinds

Two agents engage in a conflict-resolution dialogue (CRD) when at least one of them wishes to solve a conflict which has arisen with the other. A CRD is then initiated as soon as an agent believes that a conflict exists between its plan and the plan of another agent. The way the CRD is carried on depends, among other factors, on the nature of the conflict arisen.

2.1 Objective vs Subjective Conflicts

A first classification may be made with respect to the perception that the agent initiating the dialogue has of the conflict. The belief that a conflict exists can be right, and in this case there is also an *objective* conflict between the two agents, or mistaken, and in this case the conflict is merely *subjective*. CRDs aimed at solving merely subjective vs objective conflicts are different, at least from the moment when one of the agents suspects or realises that the conflict is only subjective:

- dialogues undertaken to solve conflicts based on erroneous beliefs are aimed at *clarifying misconceptions or misunderstandings*, on the grounds that eliminating these mistaken beliefs will eliminate the conflict itself;
- dialogues undertaken to solve an objective conflict are aimed at *changing the goal of the other agent*, with, typically, two distinct strategies:
  1. inducing the other agent to drop part of his conflicting goals, in order to come to a *compromise* (in this case, both agents relaxe their goals and renounce to part of them) or
  2. inducing the other agent to completely drop its conflicting goals (*unilateral renounce*).

  In both cases, the interaction is not limited to the clarification of a misconception, but is a classical case of influencing, that is of attempting to change the goals and the intentions of the other, through argumentation and persuasion.

Of course, in real life CRDs, these two cases can coexist (we shall see some examples of this coexistence in Sect. 4.):

- agents can pass from a perspective to the other: they can start believing that there is an objective conflict, and then discover that it was just a problem of misunderstanding;
- the two agents may have different opinions about the objectivity of the conflict: one agent believes that there is just a misunderstanding, while the other believes that there is an objective conflict;
- the conflict might be in part (on certain goals) merely subjective and due to misconceptions, and in part (on other goals) objective.

A further distinction has to be made, to clarify the difference between subjective and objective conflicts:

a. *merely subjective* conflicts originate from some mistaken belief of one agent, concerning the two conflicting goals or their relations: the agent believes that there is a conflict, while such a conflict does not really exist, because the two goals are in fact not incompatible. This case has, in turn, two different sub-cases:

a.1. *false, apparent conflicts*, based on a misunderstanding about the other agent's goals:

X believes that Y wants p, while Y, in fact, wants Not-p.

In this case, the conflict does not exist at all objectively, but only subjectively (in X's mind), and is due to a mistaken belief of X about Y's goals.

For example:
Adam would like to go to the restaurant, tonight; he believes that Eve would not like it, while in fact she was just going to reserve a table at a nice restaurant in Trastevere.

a.2. *false indirect conflicts*, based on a misconception about the world:

X wants q and believes that Y wants p and that p implies Not-q, while this implication is false.

In this case too the conflict does not exist objectively, but only subjectively (in X's mind), as it is due to a mistaken belief of X about the world[1].

For example:
Adam would like to go to the restaurant tonight, while Eve would like to go to the theatre. Adam would revise his belief that this is a conflict if Eve convinces him that the two plans are not incompatible (for instance, because the show starts late).

These kinds of conflict are very different from a real conflict:

b. *misconception based real conflict*:

X wants p because it (mistakenly) believes that q, while Y wants Not-p because it believes that Not-q.

In this case, the two agents do have opposite goals, but (at least) one of them is based on wrong beliefs, about the world or about the other; if this wrong belief is retracted, the goal will be abandoned and the conflict will be resolved.

For example:
Adam would not like to go to the theatre tonight, because he believes that there is nothing interesting being played, at the moment; Eve attracts his attention to the new Dario Fo's play, that she would like to see, and maybe he will change his mind.

[1] In some logical systems, a.2 is a subcase of a.1, in the sense that if an agent has the goal that p, and believes that p implies q, then it has also the goal that q. A conflict of type a.2 would always imply, in these systems, also a conflict of type a.1 (but not vice-versa). The two types of conflicts have, in any case, different foundations.

It is important, therefore, to understand that: (i) not all conflicts are due to misconceptions about the other (and about the conflict itself) or about the plans and the domain and (ii) misconceptions do not necessarily produce conflicts. Thus, only some CRDs are dialogues about a misconception and are aimed at clarifying it.

2.2 Conflicts about Ends vs Conflicts about Means

In principle, conflicts about end-goals, that is about motives (e.g. desires or needs) are different from conflicts about sub-goals, that is about instrumental goals in a plan. In simple situations, abandoning an end implies not achieving it, and therefore renouncing to something important. On the contrary, dropping a mean, a sub-goal, may simply imply substituting it with another equivalent mean: in this case, there is no renunciation, no dissatisfaction, since the mean is not important *per se*, but only in view of the higher order goals it guarantees. Thus, a conflict about the ends can be solved only through a complete or partial renounce by at least one of the conflicting agents, while a conflict about the means can be solved without substantial renunciation by any of the two agents: the problem is just to find some reasonable alternative plan for the end-goals and to inform (and possibly to convince) the other of such an alternative.

Of course, this is an oversimplified picture. In fact, usually the alternative plans or sub-goals are not completely equipollent: if we really choose according to some preference, this means that in general two alternative means/plans P_a (e.g. "to fly") and P_b (e.g. "to take a train") to achieve a goal G_1 (e.g. "to go to Naples"):

1. either are not perfectly equivalent as for G_1: they achieve G_1 with different probabilities or in a different measure,
2. or are equivalent as for G_1, but one of them is "preferable" because other independent ends are involved (e.g., G_2="to save time", and G_3="to save money") and P_a and P_b are not equivalent with respect to these ends. For example: P_a achieves also G_2 while P_b achieves also G_3, therefore the choice depends on whether G_2 is more important than G_3.

In the latter situation -which is the most frequent- the choice among two alternative means (for G_1) implies the choice between other ends (G_2 and G_3); renouncing to a means, therefore, implies in fact renouncing to some of these ends.

Even with these complications, it remains true that a conflict between end-goals is different from a conflict between instrumental sub-goals. In fact, even if, by choosing among sub-goals, an agent gives up some ends, these are just side-goals, they are not those which were motivating the agent's plan and action (we are not travelling to save money or to save time, but in order to be in Naples). Sacrificing side-goals (even end-goals, which however do not motivate one's action) is less difficult than sacrificing real motivations, and also the types of argumentation to employ in the two cases are different.

2.3 The Compromise

For a possibility of a *compromise* to exist, agents must have goals of a special kind: goals that can be attained "partially". There are dichotomic goals (for example, to pass an exam, to die, etc.): either you achieve them or you fail. There are other goals that can be achieved only in part:

- either gradually, since they have some quantitative property (for example, to gain a lot of money, to eat, etc.),
- or partially, since they are complex goals which include several, less complex, goals: for example, to go to the theatre and to a restaurant are subgoals of the more general goal of "enjoying oneself".

In both cases, one can achieve a subset, a part of the original -dimensional or complex- goal. Let us call these goals *composite.* These kinds of goal have two important, although obvious properties: they can be "partially conflicting" and they can be "partially dropped". For a compromise to be made, then:

- each of the two agents, A_i and A_j, must have a composite goal, G_i and G_j, and the conflict must be about G_i and G_j; this means that some parts of G_i are incompatible with some parts of G_j;
- both agents have to renounce to the complete achievement of their goal, and have to abandon some conflicting part of their goal.

3. Conflict-Resolution Dialogues: the Role of Personality

Human beings show a variety of behaviours, due to very diverse factors (knowledge, goals, preferences, etc.); adaptive systems try to embody these factors in a User Model, in order to customize their behaviour to them. In addition to more common factors related to the knowledge state of the user (the classic distinction between *expert* and *novice* is an example), new classes of factors are now being considered in the intelligent social systems community, which are often referred to as "personality traits": these factors are close to representing the "aims" of an agent and the ways to pursue them (Carbonell, 1980, Castelfranchi et al., 1998).

In conflict resolution conversations, as in any kind of conversation, personality traits can play various roles.

They can influence the choice of the **strategy** to employ in a conflicting situation; for instance a *fussy* agent will tend to argue about every single goal, without bothering with trying an alternative plan, and without considering that reaching an agreement on a minor goal can be time consuming and useless. On the contrary, an *accommodating* agent will always try to find different ways to achieve its own higher order goals in order to keep everybody's higher goals unchanged. Similarly, a *polemic* agent will argue even about not interfering goals; a *fair* agent will not use trick options (he will not lie), etc.

Personality factors also influence the **degree of attachment** an agent has towards its own goals. Many authors have addressed the problem of goal endorsement: Carbonell, 1980 makes a distinction between *cooperative* and *competitive* predisposition, and terms like *selfish*, *helpful* or *benevolent* have often been used to describe different attitudes with respect to goal endorsement (Cohen and Levesque, 1990).

Finally, personality traits influence the agent's **reaction to the success or the failure** of one of its strategies: a *persistent* agent will try the same plan many times before abandoning it, an *ambitious* one will try every possible plan to achieve its goals, and so on.

If we assume that personality traits influence human communication, then also when conversing with a computer it has sense to consider these factors in the user model: there is by now agreement on the fact that a real "friendly" computer system has to be able to perceive the personality (and even the emotions, see Picard, 1996) of its users to tailor its behaviour to them. But it should also be admitted that an "intelligent" system has, more or less implicitly, its own behaviour and idiosyncrasies, which can be allegedly seen as a "personality". It is time then to represent this model explicitly, by examining how it influences the algorithms behind system-user dialogues.

4. A Case Study in Doctor-Patient Dialogues

In this section we analyse two CRDs from the perspectives outlined so far. We have chosen two dialogues in a Doctor-Patient scenario, for several reasons. Doctor to patient interrelations are influenced by a number of underlying conflicts: the degree of information that patients should receive about their treatment (Berry et al., 1997) and the level of autonomy they should have in contributing to establish the therapy (Savulescu and Momeyer, 1997) are just two of a very large set of samples of potential conflicts. Dialogues in medical setting thus provide significant examples of the types of conflict that we classified in the previous Sections, and of how they are frequently only in part raised, and only in part solved, in the dialogue. The conversations shown in this Section include several of the conflicting situations mentioned before. Both dialogues have, as a trigger situation, a *purpose declaration* by one of the participants, which the other one realises to be in conflict with one of its goals. The conversations consist of a set of *turns*; in each turn, a participant can make a claim (with or without justifying it), ask a question, or make a proposal (if it takes the initiative). Alternatively, it can answer a question, acknowledge a claim, or declare agreement or disagreement to a proposal (if it replies to a stimulus from the other agent). The dialogues have the aim to

Turn 0 P:	*I would like you to tell me about the side effects of these drugs, because I don't know them.*
Turn 1 D:	*I know: but I'm not sure that it is really necessary that you know them.*
Turn 2 P:	*However, I have to inform you that, if I know them and I know, at the same time, the way that they occur (whether they are frequent, serious and so on), my motivation towards the treatment will increase.*
Turn 3 D:	*I understand now. I thought you were an anxious person, who might have been demotivated from knowing these aspects of treatment.*
Turn 4 P:	*All right. I therefore would like you to tell me about side effects, and the way they might apply to my case.*
Turn 5 D:	*All right, I will do it.*
Turn 6 P:	*Thank you.*

Fig. 4.1. A clarification dialogue

identify, step by step, the reasons for the original conflict, and to solve it, if possible, or to conclude that it cannot be solved.

4.1 A Clarification Dialogue

The first example (in Fig. 4.1) shows a computer simulation of a dialogue between an intelligent explanation system and a user (de Rosis et al., 1995). Both the system and the user are simulated by the computer system. In the figure, the explanator-agent's turns are labelled with D, for Doctor, as this is the role this agent would play in a session with a real user, and the user-agent's turns with P, for Patient. The conversation is about a piece of text D has just generated to describe the drug therapy that P should undertake. P is not satisfied with the text produced, because it does not mention the side effects of the therapy, and this creates a conflict between the two. It has to be remarked that the choice of this particular conflict is not fortuitous: experimental studies have shown that this type of situation is very common in real life doctor/patient relationships (Berry, Michas and de Rosis, 1998).

The conflict triggering this CRD is clearly concerned with the doctor's communicative goal: "Informing about the side effects of the drug". This conflict turns out to be only subjective, and is due to a misconception of the doctor about the personality of the patient. The two agents have a cooperative attitude and are both interested in solving the conflict: each of them tries to convince the other to change its opinion but, at the same time, is prepared, if necessary, to change its mind; the main part of the dialogue aims therefore at discovering the reasons for the conflict.

Let us examine this conversation from the personality traits point of view.

– D initially hypothesised that P had an *anxious* personality: mentioning the side effects of the therapy could then have weakened P's will to undertake it. Therefore, D decided to omit them from the explanation of the therapy.
– On the other hand, P mistakenly hypothesises that D believed that P already knew the side effects of this drug. P therefore proposes to D to mention them, and justifies this proposal with the claim that in fact they are not already known (Turn 0).
– D communicates that P's hypothesis is not true, and that its goal was actually to avoid mentioning them at all (Turn 1).
– P tries to find another explanation for the conflict (D's hypothesis about P's personality) and manifests its real attitude towards performing the action (Turn 2).
– D acknowledges this communication, changes its belief about P's personality and changes its mind about the goal, by accepting P's proposal (Turn 3).

The dialogue, although attempting to represent a real life conflict, is artificious, and especially the behaviour of the doctor-agent is probably unrealistic. In the next section, a real dialogue is presented, to give a flavour of the complexity of the behaviour that believable agents should, instead, show.

4.2 A Compromise-Finding Dialogue

The second dialogue (Fig. 4.2) is the English translation of part of a real conversation between a patient (P) and a gynaecologist (D) in an Italian public surgery, reported in Petrillo, 1994. The conflict, in this case, concerns the contraceptive to be selected as the most appropriate to the specific patient, but also some related, higher order goals. This is, therefore, an example of objective conflict.

P's unique goal is to take a "suitable" contraceptive, where suitable means, to her, to take the pill; she therefore needs a prescription for that.

D's goal is more complex: she wants to promote a correct use of contraceptives, which does imply prescribing a suitable contraceptive, but also being sure, at the same time, that the patient is convinced about the chosen method and that she is informed about contraceptive methods in general. The first subgoal can be in turn specified by one of two -alternative- subgoals: prescribing the pill or let the patient have a coil fitted. Although the doctor seems to prefer, in principle, the second one, she believes that a high consideration should be given to the patient's preferences.

There are, therefore, in this conversation, both objective and subjective conflicts between D and P. The objective conflicts are about the doctor providing a complete and well-balanced vision of contraceptive methods, tailored to the patient's needs and beliefs, and the patient providing information about the reasons behind her preference for the pill. In contrast, the subjective conflict concerns what the two participants to the dialogue want to select as a contraceptive.

Turn 0 P:	*I'd like a prescription for the pill, please.*
Turn 1 D:	*You'd like the pill? And... you are already aware of other methods... What are your reasons for the pill?*
Turn 2 P:	*Oh, the coil! I don't wanna use that!*
Turn 3 D:	*You don't feel like it.*
Turn 4 P:	*....and I think this is the only solution...*
Turn 5 D:	*Why wouldn't you have a coil fitted?*
Turn 6 P:	*I don't know why... ...My gynaecologist advised me against it.*
Turn 7 D:	*Your gynaecologist advised you against it... did he say why?*
Turn 8 P:	*He says it's something foreign to the body, after all...*
Turn 9 D:	*OK, it's all right that you take the pill now, ...it's the right time, ...especially as you're young.*

Fig. 4.2. A compromise-Finding Dialogue.

These conflicts might be solved in several ways:

To solve the *objective conflicts*:

a.1) P might accept D's goals: she might inform the doctor about the reasons why she prefers the pill, and might leave to the doctor the opportunity of providing information about contraceptives in general;

a.2) or, D might accept P's goal to avoid talking about herself, and renounce to inform her about contraceptive methods.

To solve the *subjective conflict*:

b.1) P might revise her belief about the selection of the pill, and accept a coil-based contraception; in this case, the conflict, though only subjective, would cause P changing her plan anyway;

b.2) or D might clarify that she is not against the use of the pill, and prescribe it; the correction of the misconception would then make the subjectivity of conflict manifest.

When, as in this example, goals are complex and several alternative ways to solve the conflict exist, the personality of the two interacting agents plays a crucial role in establishing a selection criterion: agents will select which goal to abandon in case of conflict, and which goal to try and achieve (by means of argumentation, clarification and so on), also according to their personality, and therefore to the value they attach to these goals.

To come back to our example, at the beginning of the dialogue D tries to push P towards solution a.1, though in a "soft" way. However, P is very determined not to renounce to her goal, and also not to adopt D's higher order goal. D, therefore, moves towards solution a.2, by means of a *unilateral renounce*.

Throughout the dialogue, the suspect of a misconception (that the coil is dangerous) arises to D, and she tries to clarify it; however, P misinterprets D's attempt, identifies a (subjective) conflict (that the doctors wants to convince her to have a coil fitted) and consequently refuses the clarification attempt of D, by adopting a deceptive strategy (that her gynaecologist advised her against the coil).

Remarkably, D is the agent who guides the conflict-resolution dialogue (probably because of her role).

It should also be noticed that, to induce D to her unilateral renounce, P does not employ any kind of sophisticated argumentation: she only manifests that she is closed to any offer of compromise. In this example, the personalities of the two agents (as well as their role, and the type of structure in which the dialogue takes place) contributed to establish the form and level of solution for their conflict.

We shall turn back to this dialogue in Sect. 5.2, for a more detailed analysis of the two agent's forms of reasoning at each turn.

5. Forms of Reasoning in Conflict Resolution Dialogues

To simulate CRDs in a convenient way, the reasoning process that each participant performs at each turn before issuing a "communicative act" needs to be modelled. This model can be based on the application of *conversational rules*, like, for example, in Schlegloff, 1988. However, if the dialoguing agents are represented as BDI agents, it is probably more reasonable to represent the reasoning process in terms of sequential activation of several *forms of reasoning*, applied to the agent's mental state. Examples of representation of such a mental activity are the recursive *Propose-Evaluate-Modify* cycle for modelling collaborative planning in Chu-Carroll and Carberry, 1995, and the model of proposal/acceptance (or rejection) sequences in Sidner, 1994. We go one step further, to examine forms of reasoning employed in the main phases of this activity.

5.1 The Reasoning Cycle

In our model, the conversation can be seen as the result of applying recursively the following steps: (a) interpreting the communication received, (b) revising one's own mental state, (c) deciding how to react, (d) answering. The sequencing of these steps, and the way each of them is performed, are linked to personality traits and to extralogical factors, such as emotions and higher order motivations.

a. interpreting the communication received: *"what did the other agent say?"*
We shall not consider here the linguistic aspects of interpretation, and shall assume that any statement is always interpreted "correctly", and translated into a standard communicative act. We make a distinction between two types of communicative acts: *questions* and *proposals*, which activate distinct forms

of reasoning and reaction. The argument of both types of speech act can be either a belief or a goal of one of the interacting agents.

b. revising one's own mental state: *"what do I believe, and what does the other agent -presumably- believe?"*

b.1. *after a proposal:* the agent receiving the proposal has to decide whether it conflicts or not with its own belief and goal structure (by reasoning on its own mental state), and whether to update its model accordingly. The same reasoning process can be made, as well, on the other agent's image. The decision on whether to reason only on itself or also on the other agent's mental state is traditionally regulated by the agent's attitude towards its partner, which is defined as *cooperative* or *non-cooperative*. In cooperative dialogues (which are the majority, in the literature: see, e.g., Chu-Carroll and Carberry, 1995), the agent would accept any non-conflicting proposal without reasoning on the other agent's mental state, while such a form of reasoning would be activated only to discover weaknesses in the other agent's viewpoint, to be exploited in argumentation (Sycara, 1991). We maintain that this dichotomy between cooperative and non-cooperative agents is not realistic, and that believable agents show various *levels of cooperativity* as to specific goals, even within the same dialogue (Castelfranchi, 1992). Cooperativity levels depend not only on the context, and therefore on the agent's higher order goals, but also on the agent's personality. For example: an *altruistic* agent would behave in a "truly cooperative" way by systematically considering the other agent's viewpoint, to verify whether the received proposal really matches this agent's presumed interests. Our assumption implies that, after interpreting any proposal, an agent may decide whether to accept it by either reasoning only on its own mental state or also on the other agent's one.

b.2. *after a question:* in order to provide an appropriate answer, the agent has to verify the truth value of some of its beliefs or goals (in case of Yes/No questions), to instantiate one of its beliefs or goals (in case of Which questions), or to find out which data support one of its beliefs or goals (in case of Why questions). However, it may also ask itself why the other agent made such a question in the first place, especially if it has a "defensive" attitude ("what is it *really* asking me?") or a "highly cooperative" attitude ("how can I help?").

When the agent reasons on the other agent's viewpoint and finds out several candidate explanations for a question or a proposal, traditional, rationality based approaches suggest to select either the "most plausible", or the one which "most quickly resolves the conflict". This idea corresponds to a vision of "optimality in problem solving", which is prevalent in Artificial Intelligence.

We argue that this process is highly influenced by emotional or personality factors of *both* interacting agents, and that the optimality criterion has to be conditioned to the personality of the reasoning agent and to its hypotheses about the other agent's personality. For example: doctor's conclusions when hearing the sentence: "*...My gynaecologist advised me against it.*" (at Turn 7) would vary depending on whether D has reasons to believe that P is *trustful* towards her, or whether they had an old or a recent relationship. Another example: a *selfconfident* agent will tend to ascribe the reasons of a conflict to the other agent's errors rather than to errors of its own accord; the opposite will be for an *insecure* agent.

c. deciding how to react: *"what is really convenient, to me and to the other agent?"*
Once the agent has come to a full understanding of the other agent's message, its reaction will be conditioned by its own interests and beliefs. In deciding how to react, the agent will consider not only the subject of the message, but also whether other (higher order) goals exist, which deserve consideration. Let us suppose, for example, that the agent concluded in the previous step that the conflict may be due to a divergence about a specific belief. It has thus to decide whether to change its mind about that belief or try to convince the other agent to do it: this decision will depend on how *persistent* the agent is. Another example: an *aggressive* agent might be willing to hold on a conflict even after the original causes have been -partially or totally- clarified. An *acquiescent* agent might decide, on the contrary, to pass over the conflict even though it has not been solved. Personality traits, attitudes, roles and higher order goals are strictly interrelated issues.

Similar considerations can be applied in deciding whether no more conflicts exist, and the conversation can be terminated with a declaration of "agreement". In our example dialogue the doctor clearly agrees with the patient's contraceptive choice: therefore, she might well have just answered "OK", at Turn 2, and written down the prescription. On the contrary, she wants to verify whether the patient's decision is really informed and conscious, and carries on the dialogue accordingly. In a different context, agents might show a different attitude, corresponding to different higher order goals, and would therefore behave differently.

d. answering: *"what shall I say?"*
Several decisions may need to be made:

d.1. *when to argue and how to argue*: the agent may reply to a proposal by declaring to accept it, or by making a counterproposal, and arguing for it. In the simulation of purely rational dialogue, argumentation is typically a function of the strength of the arguments available. In a believable agents context, strong or highly plausible arguments are not necessarily employed in the replies.

In the example dialogue, the doctor never argues "by discredit of a contraceptive", but prefers "positive" arguments: even if she seems to think that the coil is preferable to the pill in the specific case, she never argues against the pill. If the doctor did, because of her role and authority, the patient might let her pursue her informative goals. The reason why she renounces to this is probably that her goal to avoid encouraging negative attitudes towards any contraceptive has higher priority.

d.2. *when to make a question*: in purely rational negotiation dialogues, information is requested only when the agent has not enough data to decide whether to accept the other agent's proposal. When believable agents are modelled, this decision is also conditioned to the agents' higher order goals. Some of these goals are linked to the context and to the interrelations between the agents, whereas others may come from previous phases of the dialogue, or may even be active *a priori*.
Taking these higher order goals into account is therefore essential to establish how to reply: as we have discussed before, this decision may presume different reasoning styles in the previous phases of the dialogue (only on one's own mental state, or also on the other's). Conversely, higher order goals may also lead to different decisions about asking: in our example dialogue, the doctor maybe wonders why the patient made a specific statement, guesses one or several reasons for that, but decides to abstain from asking about them, because of her higher order goals.

d.3 *which question to make*: in our example dialogue, the doctor makes either *open* (Turns 1 and 5) or *closed* (Turn 3) questions. The first question type is not necessarily preferred only when the doctor has not enough information to provide a list of possible replies. The decision can be influenced, as well, by higher order goals, such as "to encourage the patient to assume an active role", or "to manifest an enlistening attitude". In selecting a closed question, the doctor might, on the contrary, be willing "to show her competence", or "to manifest her intention to place the patient on a parity position", and so on.

d.4. *when to avoid a question, to answer in a doubtful or elusive way, or to lie:* when an agent is on the defensive (maybe because it is aware of the weakness of its argumentation ability), it will be induced to select "elusion", "reticence" or "deception" as an alternative to a sincere assertion (see, for example, the patient at Turn 6). Conversely, the intention "to avoid polemizing" may induce to answer in a "slightly doubtful" way or to pretend not to notice a lie (as the doctor at Turn 9).

5.2 The Reasoning Cycle in the Compromise-Finding Dialogue

In this section we try to apply ideas sketched out in Sect. 5.1 to the analysis of the reasoning behind the doctor and patient dialogue about contraceptives.

Turn 0 P: *I'd like a prescription for the pill, please.*
Turn 1 D:

1. By reasoning on her model, D checks whether she agrees with P's goal

`Goal(P, Prescribe(D, P, pill)).`

She discovers to share the same goal, for "positive reasons" (the patient's age); however, she has no specific reason to exclude the alternative goal either:

`Goal(D, Prescribe(D, P, coil)).`

She therefore adopts a "non directive", slightly doubtful answering strategy, by saying:
You'd like the pill?

2. D then examines her own goals, and concludes:

   `Goal(D, Promote a correct use of contraceptives).`

   She also verifies that this goal is conditioned to the following sub-goals:

   `Goal(D, P receives the correct contraceptive),`
   `Goal(D, P is convinced about the method),`
   `Goal(D, P knows about contraceptives in general).`

   The first two could be achieved by just prescribing the pill. However, D privileges the third one (maybe it has the priority) and, as she does not know whether

   `Know-About(P, Contraceptives),`

   she asks:
   And... you are already aware of other methods...

3. To satisfy the second goal, D needs to understand the reasons of P: she then puts herself "from P's point of view", by reasoning on P's model and trying to guess why P holds such a goal (i.e., why she prefers to take the pill). However, rather than suggesting possibly unknown worries about other contraceptives, she just asks:
   What are your reasons for the pill?

Turn 2 P:

1. P has a "defensive" attitude. She then tries to guess why

   `Goal(D, Know-why(Goal(P, Prescribe(D, P, Pill)))`

   and she concludes that:

   `Goal(D, not Prescribe(D, P, pill))`

   possibly because:

   `Goal(D, Prescribe(D, P, coil)).`

   P, therefore, presumes that a conflict exists, with D, about which contraceptive to use. As we shall learn later on, this conflict is only apparent.

2. By reasoning on her model, P checks whether she may agree with the (presumed) "real" goal of D, and realizes that

   `not Goal(P, Take(P, coil)),`

   therefore she just says so:
   Oh, the coil! I don't wanna use that!
   This is also an implicit reply to D's second remark ("yes, I do know that a coil is an alternative to the pill!").

Turn 3 D:

1. By reasoning on her model, D discovers that she does not share completely P's goal
   `not Goal(P, Take(P, coil)),`
   but decides to avoid manifesting her disagreement.
2. To pursue her plan, she needs to know why P has such a strong view against the coil. She then reasons on P's model, and tries to guess which beliefs supported that goal. D selects one of the candidate reasons:
   `Bel(D, Bel(P, PsychologicalArgumentsAgainstCoil))`
   and tries to suggest the answer:
   You don't feel like it.

Turn 4 P:

1. By reasoning on her model, P discovers that she agrees with D; however, she does not want to talk about this issue (this is a real, objective conflict!) and then gives an elusive answer, by turning back to her first request
   `Goal(P, Prescribe(D, P, pill)):`
   ....and I think this is the only solution...

Turn 5 D:

1. D interprets the communication as an implicit admission that her guess was correct.
2. To achieve her goal about "Correct information on contraceptives", D decides to go deeper into the reasons of P's bias about the coil, in order to correct possible misconceptions. She then adopts the goal:
   `Goal(D, Know-why (not Goal (P, Take (P, coil)))),`
   which is pursued by asking:
   Why wouldn't you have a coil fitted?

Turn 6 P:

1. By reasoning on her model, P discovers that she does not want to talk about it; she therefore gives an "elusive" answer:
   I don't know why...
2. However, P immediately realises that such an answer is implausible and, maybe in order to discourage D from arguing in favour of the coil, she lies by "recurring to the authority":
   `Goal(P, Bel(D, Source(Gynaec, ArgumentsAgainstCoil)))`
   ...My gynaecologist advised me against it.

Turn 7 D:

1. By reasoning on her model, D wonders whether the communication is plausible: she doubts about it, but does not want to manifest her doubts; she therefore reformulates P's statement:
   Your gynaecologist advised you against it...
2. She then decides to ask for more information:
   did he say why?

Turn 8 P:

1. By reasoning on her model, P decides to carry on with the lie, by ascribing to her gynaecologist maybe a belief of her own:
   `Bel (P, Bel (Gynaec, ForeignBody(coil)))`
   He says it's something foreign to the body, after all...

Turn 9 D:

1. By reasoning on her model, D considers implausible this statement, but once again she decides not to disclose her disagreement.
2. Maybe D understands that P is elusive and lies, probably to hide her negative "fantasies" about the coil. She decides that there is no point in carrying on with this conflicting situation, a compromise is not possible, and she is forced to a unilateral renounce:
   – she concludes that P definitely does not want to use the coil, and abandons the goal of suggesting it, as opposed to prescribing the pill, as a way to achieve her first goal, of providing the best contraceptive;
   – she also concludes that it is difficult to convince P that the coil is not a bad contraceptive method, and therefore abandons her third goal, of providing information on contraceptives in general; in doing so, she also renounces to reveal P's lie and to correct her misconceptions;
   – she therefore privileges her second goal, of insuring that the patient is convinced about the chosen method, and, as she herself is convinced that the pill is a suitable -to P- contraceptive, she decides to reinforce positive arguments for it:
   OK, it's all right that you take the pill now, ...it's the right time, ...especially as you're young.

5.3 Asymmetry of the Reasoning Process of the Two Participants

The reasoning cycle that we described in Sect. 5.1. is the "generic" reasoning that a "generic" agent would perform. However, interacting agents do not necessarily reason in the same way: in the majority of cases, their relationship is not symmetrical, as personality, emotion and role factors are not the same.

Our compromise-finding dialogue is a significant case of such an asymmetry: the doctor makes questions, the patient answers, and rarely takes the initiative. Moreover, as far as the reasoning process is concerned, the patient reasons prevalently on her own model, whereas, at almost every turn, the doctor puts herself from the patient's viewpoint.

Such a dialogue cannot, consequently, be simulated by a recursive model in which an agent's behaviour is supposed to be unique (like in the artificial clarification dialogue in Sect. 4.1): reasoning strategies will rather have to be tailored to the two agent's characteristics and to their personality traits.

6. Xanthippe: A Dialogue Simulation System

To test our hypothesis about reasoning in CRDs between believable agents, we designed a system, Xanthippe, which enables us to simulate these dialogues in different conditions. Agents are programmed, in the system, by an agent-definition language which draws ideas from AGENT-0 (Shoham, 1993) and adapts them to our needs. The system is made up of two components:

- an *agent creation* component, that enables defining agents with a given "mental state", a set of "reasoning tasks" that they are able to perform (implemented as C++ and Lisp programs) and a set of "commitment rules", to decide on when these tasks have to be activated;
- a *dialogue simulation* component, that enables simulating the dialogue between those agents.

6.1 Agent Creation

Agents are designed so as to hold specific goals, beliefs and personality traits, and to follow conversational strategies which are not necessarily unique and are not uniformly cooperative or non-cooperative throughout the dialogue.

6.1.1 Mental State. The mental model of each agent is made up of three main components:

1. *agent's own mental state*: this is a belief-goal tree structure representing the conditions under which a goal is held and the way higher order goals can be achieved, given a set of "basic" beliefs and personality traits, and as a function of presumed personality traits of the other agent;
2. *image of the other agent's mental state*: this default image has the same components as the first one, although a not necessarily identical structure. We assume that every agent may accept that the other agent can reason in a different (from its own) way. This attitude towards the others depends, again, on the agent's personality: an *egocentric* agent would expect that all agents reason in the same way as it does;
3. *shared beliefs and goals*: a set of common sense beliefs and goals (for instance: concept definitions) which, in the agent's view, are common to all the agents with which it interacts.

6.1.2 Forms of Reasoning. Our agents are able to perform various reasoning tasks, each implemented by a specific *knowledge processing module*:

1. *does a proposal conflict with my beliefs or goals?*
   When an agent reasons on itself to verify whether to adopt a proposed goal or belief, it applies a goal-directed reasoning to its own mental state; if neither the candidate formula nor its negation can be derived from this knowledge base, the agent concludes with an "indifference" towards the proposal; otherwise, with an "agreement" or a "conflict".

2. *does a proposal really fit with the other agent's interests?*
   A similar type of reasoning is made on the other agent's mental state, to verify whether a received proposal matches that agent's presumed higher order goals.
3. *why does the other agent believe a communicated fact (or hold a communicated goal)?*
   When the agent investigates on the reasons of a communicated belief or goal, it makes an *abduction* on the other agent's mental state. An abduction applies a stepwise backward reasoning by checking, at each step, which of the -presumed- supporting beliefs and goals of the other agent conflict with its own components. When several candidates to explanation of the conflict exist, the selection of the most plausible of them is made by attaching to each hypothesis an endorsement value, which is a function of the emotional state and of the personality of the agent which is performing the abduction process.
4. *why do I believe a fact (or hold a goal)?*
   A backward reasoning on the agent's own mental state is made to answer know-why questions, in order to retrieve the basic beliefs which support the inquired fact.
5. *which is the answer to a Yes/No question?*
   This requires a simple goal-directed reasoning.
6. *should I change my mind, or try to convince the other agent to change its mind (in case of conflict)?*
   This task applies personality-dependent criteria to decide what to do when the agent discovers that a divergence of beliefs about a fact exists.
7. *which are the consequences of my change of mind?*
   When an agent reasons on the consequences of changing its mind, it makes a data-directed reasoning on its mental state, to revise its beliefs and goals, and to check, in particular, whether the original conflict is solved or still exists.

Other forms of reasoning, which have not yet been implemented in Xanthippe, are the following:

1. *do I hold other higher order goals?*
   This form of reasoning requires examining all "pending" goals, that is all higher order goals which have not yet been achieved.
2. *(in case of agreement) should I stop the dialogue or carry on and consider other goals?*
   In this case, the pending goals retrieved are ranked according to their priority, by applying context-based criteria, in order to establish whether to try to achieve them or to abandon them.
3. *should I argue or avoid polemizing?*
   Higher order goals are applied to select the strategy in this case.
4. *should I make an open question or a closed one?*
   Again, the decision is based on higher order goals.

5. *how should I reply to a communication of "change of belief" by the other agent?*
   Such a reply can be just an "acknowledgment", or a more complex reply, with the decision based once again on higher order goals.
6. *should I answer sincerely or be elusive, evasive or other?*
   When an agent decides how to react to a communication or an answer, it may decide to apply various types of "insincere" behaviour (it may lie, elude a question, be evasive and so on), again according to its personality, and therefore depending on the set of its higher order goals (Castelfranchi, Falcone and de Rosis, 1998).
7. *should I believe the communication received, or should I hypothesize an insincere assertion?*
   The possibility of insincere behaviour of some agents implies the need -for all agents- for a careful examination of the communication received, in order to discover whether a misleading behaviour can be hypothesized.

Some of the reasoning modules of Xanthippe apply ordered resolution in a forward or a backward way, with true-false or fill-in-the-blank question types. Others are based on an abduction algorithm which sets its endorsement attachment criteria according to the personality of the reasoning agent. All modules exploit an *ATMS* (de Kleer, 1986), both to insure that truth is maintained in both components of each agent's model and to speed up the various forms of reasoning.

6.1.3 Communication Capability. Agents communicate in a standard language which is similar to that proposed in Sidner, 1994. Communicative acts are generated at the end of each reasoning turn, as a function of the result of reasoning.

6.1.4 Reasoning Strategies. When an agent receives a communication from the other agent, it starts performing a reasoning process. The steps of this process are the result of *commitment rules* which establish the reasoning task to activate, as a function of the communication received, the agent's mental state (and therefore, also emotions and personality) and the results of previous reasoning steps. In the case of the compromise-finding dialogue in Sect. 4.2, for example, the commitment rules defined for the two interacting agents would simulate, for the patient, a "non cooperative" behaviour (in which the agent reasons only on itself) and, for the doctor, a more "cooperative" one, (in which the agent also reasons on the other's mental state). Similar rules can also help to establish which communicative act to produce at the end of the reasoning process.

6.2 Dialogue Simulation

Once both agents have been defined, the dialogue simulation may start: the user chooses an agent and inserts the communicative act constituting the first turn, in which the agents makes its purpose declaration. The system shifts the control to the other agent, which starts reasoning on the communication received. When a reasoning turn is completed, the communication act produced is displayed, after having been translated into a simplified natural language, to favour an easier understanding of the dialogue. The control is then passed to the other agent, and so on, until the conflict is solved, or it is agreed that it cannot be solved. Changes of mind occurring in the two agents at each reasoning step are also displayed in a separate window, to dynamically monitor how the agents' mental states evolve.

7. Conclusions

In this paper we discussed how the design of conflict resolution dialogues among believable agents should take account of factors like personality traits and emotions, in addition to purely rational attitudes. We also sketched the design of a system, Xanthippe, able to simulate such dialogues.

The fist prototype of Xanthippe generated the clarification dialogue that is shown in Sect. 4.1, but does not yet allow us simulating completely the type of believable behaviour that is shown in Sect. 4.2. Agents, in Xanthippe, consider one goal at a time, and act according to the "sincere assertion" and the "confidence" hypotheses. To simulate the more plausible behaviours that we have exemplified, our dialoguing agents should be empowered with at least three new abilities:

1. *ability to consider several goals at the same time*, and to assign a priority to each them; as a consequence, ability to select, in each reasoning phase, which goal to pursue, which goal to abandon and so on, according also to some personality-based criteria. At present, this is not possible, mainly because reasoning is purely logical, and the strength of goals, beliefs and relationships among them is not represented. This prevents modelling the graduality of the persuasion process, the dynamic selection and abandoning of goals and the simultaneous activation and consideration of several goals at the same time, which are typical of believable agents;
2. *ability to reason about the other agent's communication*, by wondering the reasons behind a question and by considering the possibility of "insincerity";
3. *ability to consider the possibility to avoid answering* (also when the answer is known) or to answer insincerely if needed.

This research turns the dialogue simulation focus towards considering the personality and emotional factors which may intervene in situations of conflict: this analysis will enable bridging the gap between argumentation and negotiation dialogues, and unifying them into a unique, more flexible theory and modelling. Forms of insincere behaviour (such as lie, elusion, or others), which are especially relevant in case of conflict, will be introduced in a rather "natural" way in such models. Management of unexpected questions or attitudes (that we mentioned in the introductory section as one of the distinguishing capabilities of believable agents) will be introduced by enabling the agent to deal with messages that it cannot explain with its image of the other agent (due to an error in this image or to an incomplete knowledge). Effects of messages on the emotional status of the receiving agent will be simulated by revising dynamically these aspects of the user model within the process of revising the mental state. This new view therefore promises to further enrich the very interesting results that have been obtained so far in the field of dialogue simulation (Sycara, 1991, Chu-Carroll and Carberry, 1995).

Our modelling approach can be applied in several domains: conflicts between a computer system and its user may hold in any phase of interaction, due to a number of reasons (different views about which information the user really needs, how results should be presented, and so on). The possibility of conflicts among software agents has been contemplated by many authors, and we believe that the sincere assertion should probably be relaxed also in these contexts, to introduce some form of "defense", in the society of agents, against unfair attitudes of some individuals (Castelfranchi, Falcone and de Rosis, 1998).

References

D.C. Berry, I. Michas and F. de Rosis. Evaluating Explanations About Drug Prescriptions: Effects of varying the nature of information about side effects and its relative position in explanations. *Psychology and Health*, **13**:767–784, 1998.

D.C. Berry, I. Michas, M. Forster, and T. Gillie. 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, 1997.

D. Cañamero (editor). Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium. Menlo Park, CA: AAAI Press.

J.R. Carbonell. Towards a Process Model of Human Personality Traits. *Artificial Intelligence*, **15**:49–74, 1980.

J.C. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty Sneddon, and A. Anderson. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, **23(1)**:13–31, 1997.

C. Castelfranchi. No More Cooperation, Please! In search of the social structure of verbal interaction. In A. Ortony, J. Slack, and O. Stock, editors, *Communication from an Artificial Intelligence Perspective*, Computer and Systems Sciences. Springer-Verlag, 1992.

C. Castelfranchi, F. de Rosis, R. Falcone, and S. Pizzutilo. Personality Traits and Social Attitudes in Multi-Agent Cooperation. *Applied Artificial Intelligence*, **12(7-8)**:649–675, 1998.

C. Castelfranchi, R. Falcone, and F. de Rosis. Deceiving in GOLEM: How to Strategically Pilfer Help. In *Autonomous Agent '98: Working notes of the Workshop on Deception, Fraud and Trust in Agent Societies*, 1998.

J. Chu-Carroll and S. Carberry. Communication for Conflict Resolution in Multi-Agent Collaborative Planning. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, 1995.

P. Cohen and H. Levesque. Rational Interaction as the Basis for Communication. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, Chapter 12, pages 221–255. MIT Press, Cambridge (Mass.), 1990.

J. de Kleer. An Assumption-Based Truth Maintenance System. *Artificial Intelligence*, **28**:127–162, 1986.

F. de Rosis (editor). UM99 Workshop on Attitude, Personality and Emotions in User-Adapted Interaction, Banff Centre, Banff, Canada, June 1999.

F. de Rosis, F. Grasso, D.C. Berry, and T. Gillie. Mediating Hearer's and Speaker's Views in the Generation of Adaptive Explanations. *Expert Systems With Applications*, **8(4)**:429–443, 1995.

C.D. Elliott and A. Ortony. Point of View: Modeling the Emotions of Others. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, pages 809-814. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

G. Petrillo. Relazioni Asimmetriche e Strategie di Influenza Sociale: analisi di conversazioni medico-paziente. In F. Orletti, editor, *Fra Conversazione e Discorso*, pages 99–120. La Nuova Italia Scientifica, 1994.

R.W. Picard. Does HAL Cry Digital Tears?: Emotion and Computers. In D. Stork, editor, *HAL's Legacy - 2001's Computer as Dream and Reality*, Chapter 13, pages 279–303. MIT Press, Cambridge (Mass.), 1996.

S. Prevost and E. Churchill (editors). Proceedings of the First Workshop on Embodied Conversational Characters, Tahoe City, CA, October 1998.

A.S. Rao and M. Georgeff. Modeling Rational Agents within a BDI-Architecture. In J.F. Allen, R.E. Fikes, and E. Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann Publishers, 1991.

W.S. Reilly and J. Bates. Natural Negotiation for Believable Agents. Technical Report CMU-CS-95-164, Carnegie Mellon University, School of Computer Science, June 1995.

J. Savulescu and R.W. Momeyer. Should Informed Consent Be Based on Rational Beliefs? *Journal of Medical Ethics*, **23**:282–288, 1997.

E. Schlegloff. Pre-Sequences and Indirection: Applying Speech Act Theory to Ordinary Conversations. *Journal of Pragmatics*, **12**:55–62, 1988.

Y. Shoham. Agent-Oriented Programming. *Artificial Intelligence*, **60(1)**:51–92, 1993.

C. Sidner. An Artificial Discourse Language for Collaborative Negotiation. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI94)*, pages 814–819. AAAI press, 1994.

K. Sycara. Pursuing Persuasive Argumentation. In *Working notes of the AAAI Spring Symposium on Argumentation and Belief*, Stanford University, March 1991.