



Deliverable 5.2

" First specification of the components of the Companions' Minds "

Contract number: **FP7-215554 LIREC**

Living with Robots and intEractive Companions

Start date of the project: 1st March 2008

Duration: 54 months

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 215554.



Identification sheet

Project ref. no.	FP7-215554
Project acronym	LIREC
Status & version	Final "D5.2"
Contractual date of delivery	31 of May 2009
Actual date of delivery	3 of July 2009
Deliverable number	D5.2
Deliverable title	First specification of the components of the Companions' Minds
Nature	Report
Dissemination level	PU
WP contributing to the deliverable	WP5
WP / Task responsible	"WP5/T5.1.2"
Editor	"Ana Paiva"
Editor address	INESC-ID / Instituto Superior Técnico - Tagus Park Av. Prof. Dr. Cavaco Silva, 2780-990 Porto Salvo, Portugal
Author(s) (alphabetically)	João Dias (INESC-ID), Iolanda Leite (INESC-ID), Carlos Martinho (INESC-ID), Samuel Mascarenhas (INESC-ID), Ana Paiva (INESC-ID), André Pereira (INESC-ID), Rui Prada (INESC-ID), Patricia A. Vargas (HWU), Andreas Wichert (INESC-ID)
EC Project Officer	Pierre-Paul Sondag
Keywords	Emotional Agents, Social Relations, Agent Architecture, Theory of Mind
Abstract (for dissemination)	First specification of the companion's mind architecture, focusing on building models of others and using knowledge about social relations and emotions in order to create and maintain long-term relations with users.

CONTENTS

Section	Title	Page No.
1	Introduction	5
2	First Specification of the Companion's Mind	7
2.1	<i>Social Relations</i>	8
2.2	<i>Personality and Emotions</i>	10
2.3	<i>Empathy</i>	12
2.4	<i>Theory of Mind</i>	15
2.5	<i>Memory</i>	17
2.6	<i>Reasoning and Action Selection</i>	17
3	Scenario	23
3.1	<i>Applying the Companion's Architecture to the scenario</i>	23
3.2	<i>Limitations of the scenario</i>	24
4	Conclusion	27
5	References	29

1 Introduction

The main goal of WP5 is to develop adjustable models and mechanisms that support autonomous decision-making, influenced by the companion's social setting, internal state and past experiences. In Deliverable D5.1 we identified the following cognitive competences necessary in order to have our companion interacting in an intelligent and social appropriate manner: social intelligence, emotions and personality, theory of mind, memory and adaptation. From this study, a first architecture was designed, and a set of key components and their interconnections were proposed.

Further work on the analysis of the competences and social behavior allowed us to identify another competence (which was only briefly mentioned in D5.1 as a subcompetence of social intelligence) relevant for long-term interactions: Empathy. Therefore, we will incorporate an additional model of empathy into the companion's architecture, based on the work of Rodrigues (Rodrigues et al 2009).

This deliverable provides a first specification of the components studied in task 5.1 together with the specification of the empathic model, and its integration into the architecture. Our initial goal (as described in the technical annex) for the deliverable was to focus only on how the proposed components change with the perception of actions, i.e., how does the user's and agent's actions affect the companion's emotional state, social relations, theory of mind, etc. However, we have realized that providing a more complete description of the architecture, including the influence of the several components in the companion's behavior, will make it easier to understand what kind of functionalities and interconnections will be required amongst the components. We also believe that this will benefit the reader, by presenting him a clearer idea of the workflow within the architecture. As such, we will include in the deliverable a general description of how the designed components are expected to influence behavior.

Thus, we will begin by presenting the architecture's overall diagram, and will then describe each of the identified competences, specifying how the competences are mapped into the architecture's components. Afterwards, we will illustrate the application of the architecture in a particular scenario, aimed at evaluating some of its core components. Finally, we end with some brief conclusions.

2 First Specification of the Companion's Mind

In this section we will describe the proposed model for an artificial companion architecture, encompassing the following cognitive competences: social intelligence; emotions and personality; empathy; theory of mind; memory and adaptation. We believe that the architecture proposed (see Figure 2-1) will be addressing these issues.

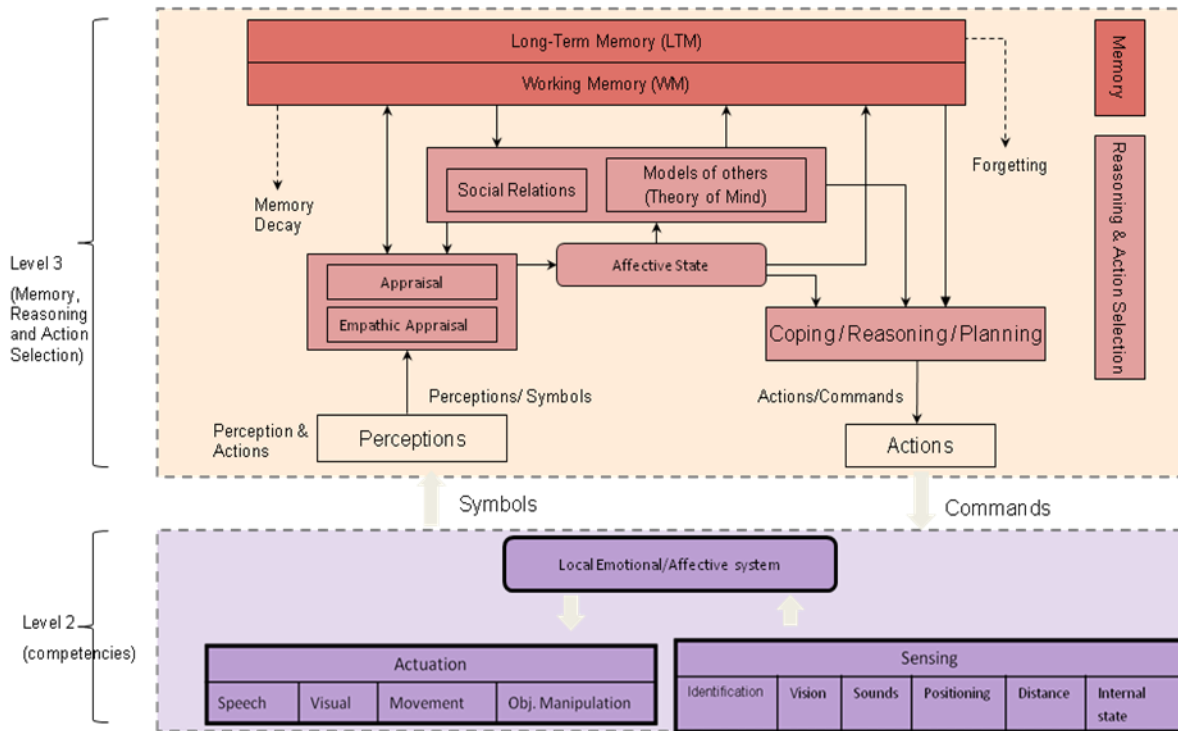


Figure 2-1 – Companion Architecture

This architecture will correspond to the level three of the general architecture for the LIREC companions (see Deliverable 9.1). In this level, the agent/robot receives a set of perceptions that were pre-processed by level two. For example, the agent can receive the perception “user’s positive feeling”, pre-processed by the vision module. The perceptions are then sent to an appraisal process. This process corresponds to a subjective evaluation of an event or situation that can result in an affective state. Following the previous example, the companion after appraising the “user’s positive feeling” perception feels, for instance, “Joy”. Note it is not imposed that the result is an emotion but rather an affective state that guides the actions of the agent.

In terms of memory, it is considered that the companion has a working memory and a long term memory. The first is used to store the agent’s current goals, plans or action rules. The second is responsible for storing information about past interactions. For example, if the companion is playing chess, the status of the goal of winning that particular game is stored in the working memory, while the number of times the companion has won against a certain user is stored in the long term memory.

Furthermore, the agent has a model of the other agents (including users) and is capable of performing some reasoning with that model (theory of mind). This allows the companion to better adapt his behaviour towards different agents/users. There is also a social relations component which defines the types of social relations established and their dynamics. As we will describe later in this document, there is a bidirectional influence between this component and the appraisal process.

Finally, after the affective state is determined and the ToM model and social relations updated, the mechanisms of coping, planning and reactive behaviours decide the next action the agent will perform. The information about the selected action is sent to the lower level in the architecture. This lower level (level 2) will, depending on the current companion's embodiment, choose a competence suitable to that action.

One of the design requirements taken into consideration was to achieve a general enough architecture to be used in different scenarios, with different requirements in terms of functionality, and even with different implementations of the modules. As such, there was an effort to generalize some of the components so that they can be implemented by different systems. For instance, in the appraisal component we do not compromise ourselves with a particular appraisal theory. The restriction is that the appraisal process must follow a two-stage process, in which a subjective evaluation of the event takes place generating a set of appraisal variables, and those appraisal variables are then mapped into particular emotional/affective states.

2.1 Social Relations

As argued in D5.1, the creation of social relations is crucial to maintain the user engaged in long-term interactions. Therefore, the companion needs to have an explicit model of social relations. He also needs to use explicit strategies and goals to create and maintain them. We will focus only on the relation of *social attraction*¹, which reflects the affective ties that one person (or agent) establishes with the others (can be positive or negative). We were also considering to model *social influence*² (also known as social power), which relates to the influence exerted by a social agent on a person, where the social agent can be another person, a social role, a norm or a group. However, it is still not clear how to use *social power* to influence the companion's behavior.

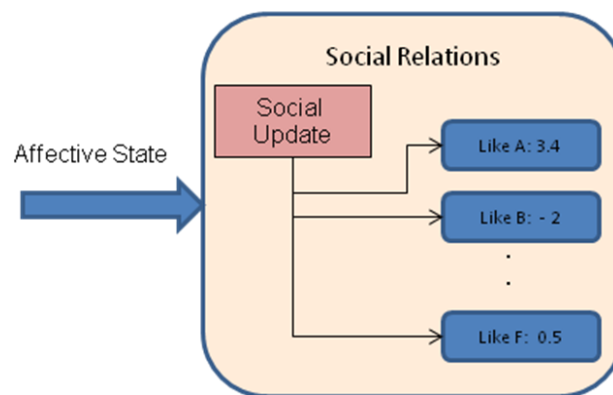


Figure 2-2 Social Relations Component

The social relations component stores the companion's social attraction towards other agents (including the user). Figure 2-2 shows an example of three relations: a very positive like relation with agent A, a strongly negative relation with B, and a slightly positive relation with F (with values ranging between -5 and 5). Since interpersonal attraction is not necessarily reciprocal, other's social attraction towards the companion need to be modelled as well. However this estimation is done in the Theory of Mind component.

¹ Definition taken from D5.1

² Definition taken from D5.1

In the architecture proposed, the dynamics of social relations (which defines how social relations change with time) follow Heider's Balance theory (Heider 1958). The Balance Theory is centred on the concept of a POX triple where P is a person, O is another social actor and X an object, which may be a third person, an idea or anything else. This triplet represents a cognitive configuration built by P, which relates P's beliefs in O's attitude towards X, and P's own attitude towards X and O. The Balance Theory hypothesis is that people avoid unstable cognitive configurations and that they mobilize their efforts to resolve it and change it to a stable state. To better understand this theory, suppose that a person P does not like X but likes O, and additionally believes that O likes X. This cognitive state is unbalanced, and thus P will try to recover the balance. According to Heider's model, P has three main strategies: (1) change her/his attitude towards O and develop a dislike for O, (2) reconsider her/his attitude towards X and start liking X, or (3) make O dislike x.

In our model, we apply the same principle to the perception of events. Suppose that the user has performed an action undesirable for the companion, such as insulting him (as shown in the example of Figure 2-3). If the event was performed by the user, then we assume that the user intended to perform that event and as such, the event is desirable to the user.³ Therefore, if the user likes the event, this corresponds to an unstable state. Then the social relations are updated in order to converge to a more stable state.

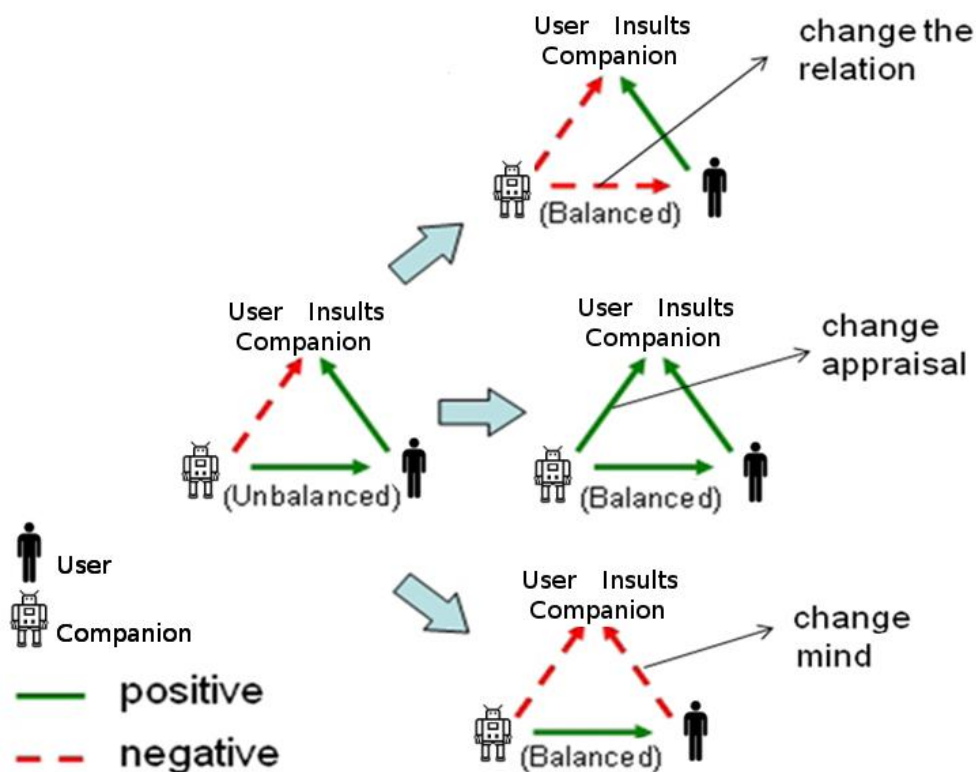


Figure 2-3 Example of balance theory applied to the perception of events

There are two strategies⁴ used in our model: (1) lowering the relation between the companion and the user, or (2) changing the appraisal of the event so that it becomes also

³ This is a simplification, since the user may have done the action unintentionally or may have been coerced to do so.

⁴ The strategy of changing other's mind is much harder to do in agents, so we are not considering it in our model.

desirable (or neutral) to the companion. The two strategies are processed in different components of the architecture. The first one to be applied is strategy 2, which is performed in the appraisal process. It works by biasing with a small weight, the desirability of an event, according to the relation that the companion has with the agent that performed the action. For instance, if the event is not very undesirable, a positive relation might make the event to be considered neutral or slightly positive, and thus reach a stable cognitive state. However, if the event is very desirable or very undesirable, the existing social relation will not be able to perform a big change in the desirability value. In this particular case, this strategy will not be effective in reaching a stable state.

After the appraisal process, the first strategy takes place, in the social update component. Whenever a new emotion caused by another agent is added to the emotional state, the social update component will analyze the valence and intensity of the emotion and update the corresponding relation proportionally. The rationale is that positive emotions increase the interpersonal relation, while negative emotions decrease it. Stronger emotions cause bigger changes in the relation.

Finally, in order for our companion to present intelligent social behaviour, he needs to use knowledge about the dynamics of social relations and how others appraise events, so that he can make others attracted to him and create/modify social relations. To do so, he needs to have explicit goals, activated for instance when the estimated user's attraction towards the companion drops below a given threshold. Then, to achieve such goals the companion must be able to reason about how the social update and appraisal processes work.

2.2 Personality and Emotions

Unlike the remaining concepts, personality does not appear as a component in the architecture, i.e. it is not a process per se. Still it is possible to create agents that appear to have different "personalities" by changing the parameterization of their emotional profiles. For example, a "fearful" companion can be modeled by attributing a low resistance to feeling the emotion "Fear". Furthermore, the definition of the companion's goals (or the importance it attributes to goals) and wired emotional reactions can also reflect the companions' personality traits, as they will have a strong impact in the companion's observed behavior.

Emotions or affective states are a result from the appraisal process, which consists on a subjective evaluation of the significance of what is happening that has impact in the agent's well being (Lazarus 1991). In other words, the appraisal process involves subjective judgements of how good or bad something is to the agent.

Therefore, in our model, the appraisal component generates affective states⁵ (which include emotions and mood) based on the events perceived. This is done using a two-stage process as shown in Figure 2-4. In the first stage, the subjective evaluation takes place, and it determines a set of appraisal variables that represent the event's impact to the agent and others. We will not specify exactly how the subjective evaluation is done and what are the resulting variables, since we do not want to compromise with a particular implementation of an appraisal theory. As example, in FATiMA (Dias 2005, Dias & Paiva 2005) this is done by matching the event against a set of predetermined rules that specify the desirability, praiseworthiness and desirability for others of the event.

⁵ Affective states may also correspond to sensations instead of emotions. For instance, the emotivevector model used in iCat does not generate emotions, but sensations instead. Given that for the purpose of the discussion in this document, sensations or emotions have the same effect, we will use the term emotion to represent both emotions and affective sensations.

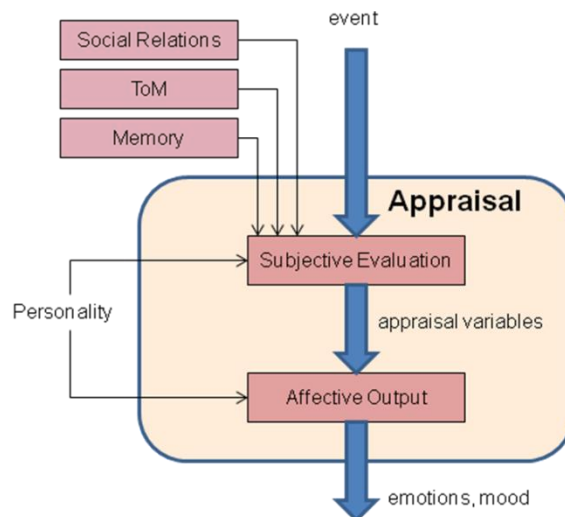


Figure 2-4 Two-stage Appraisal Process

After the appraisal variables are determined, the second stage is responsible for determining the resulting emotions, and corresponding intensities. Once more, the process of how emotions are generated and what types of emotions can be created is left open. Just to provide an illustrative example on how this could be done, the OCC theory (Ortony et al 1998) states for instance that a desirable event generates an emotion of type *Joy*, and that a desirable event that is undesirable to someone else generates an emotion of type *Gloating* towards that person.

It is important to acknowledge the influences that personality and other components have in the appraisal process. Existing social relations have a biasing factor when determining the desirability of an event, as mentioned before. The influence that the social relations have on the final appraisal can be a personality parameter, for instance we can model one agent who just cares about himself and not about others or an agent that is more influenced by what others think or feel. The ToM is also required in order to have social relations influencing the final appraisal. Memory is another component that affects appraisal. The idea here is that past experiences influence the desirability of events, in an associative manner. If an event has triggered negative emotions in the past, it may be considered negative even if it no longer presents a threat or negative condition to the agent.

Although the ToM, Social Relations, and Memory components influence the subjective evaluation process, they do not influence the second stage of appraisal directly. Only personality parameters have some effect in this process by making it harder or easier to experience an emotion of a given type.

The result of the appraisal process is stored in the agent's affective state. The affective state stores the agent's current emotions and mood. This component is also responsible for a decay process that makes the emotions and mood fade out with time. Following Oatley's model (Oatley et al 2006), emotions decay much faster than mood. The affective information stored in the affective state is then used to influence all other components, from decision making and memory to social relations.

2.3 Empathy

Empathy is one of the strategies proposed by Bickmore and Picard (Bickmore & Picard 2005) in order to maintain relationships in virtual agents, and we've argued in D.5.1 that the companion should be able to establish an empathic relation with user, being able to detect and respond sensitively to the user's affective state. However, at the time we were focusing only on high-level cognitive empathic responses (e.g. how to make a friend become happy), which is more related to social intelligence, while neglecting affective processes of empathy such as emotional contagion (feeling an emotion because of someone else is feeling an emotion). Indeed, even though Empathy has no consensual definition, many agree that "Empathy is an affective response more appropriate to another's situation than one's own" (Hoffman 2000), and as such it involves the ability to perceive, understand and experience others' emotions. Taking this into account, we now want to incorporate an additional model of affective empathy into the companion's architecture.

The creation of agents that are capable of eliciting or showing empathy has been a strong area of research. Still, most often in current agent systems, empathy is supported by empirical approaches, mostly based on pre-scripted rules. We believe this approach is not very well suited for long-term interaction scenarios, because it greatly limits the set of all possible empathic responses.

For the companion's mind, we aimed at using and adapt the general analytical model of empathy proposed in (Rodrigues et al 2009). This model is grounded on two neuropsychological theories of empathy: the Perception Action Model (PAM) (Preston & Wall 2002) and the Empathic Brain model (Vignemont & Singer 2006). These theories argue that the perceptions of others' emotional states are linked with our own somatic and autonomic responses via our own neurological representations. Also, to explain why we are not constantly empathizing with each other, several modulation factors of the empathic response are also proposed.

In accordance to these theories, we view empathy as a process. Our process involves the reactive perception of others' affective state (Empathic Appraisal) and the subsequent generation of an Empathic response (also congruent with the aforementioned theories), which then might lead to an empathic action.

2.3.1 Empathic Appraisal

Performed in the appraisal component, the empathic appraisal (as illustrated in Figure 2-5), is the main part of the empathic process. To illustrate it let us consider a situation where the companion is watching one of his friends playing against someone else. The empathic appraisal is initiated whenever the companion perceives an event that raises an emotional cue in another agent. For example, when the companion witnesses player A (his friend) smiling after winning a game against player B. After the perception of such emotional cues, the companion uses an emotion reading component to associate them to possible emotions that are being felt. For instance, in the case of A's smile the candidate emotions are likely to be Joy, Love, Pride, or Satisfaction, with for instance Joy being set as the default one⁶.

⁶ Once more, this example is based on OCC Theory. The determination of the candidate emotions will depend on the particular appraisal theory used.

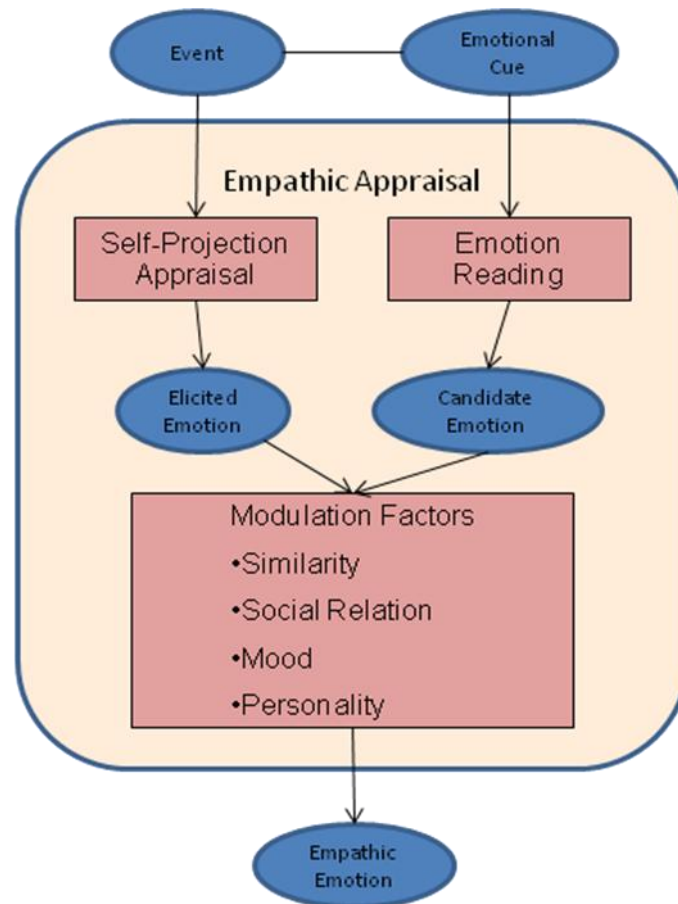


Figure 2-5 Empathic Appraisal Process

Simultaneously, the Self-Projection appraisal takes place. In this appraisal process, the companion uses his own appraisal component to evaluate the same event that caused the emotional cue, but assuming the other agent's situation by self-projection. In the previous example the companion will project himself into A's situation and appraise the same event (winning the game) but as if he was the one who had won the chess game. The emotion elicited by this appraisal is compared with the candidate emotions in order to decide the potential empathic emotion (the one that results from the empathic appraisal).

The selection for the potential empathic emotion is based on the following criteria: if the elicited emotion is contained in the group of candidate emotions, then the elicited emotion is selected. But if not, the default emotion from the candidate emotion list is selected instead. For example, imagine that the companion elicits Pride when simulating the appraisal of him winning a chess game. Since Pride is congruent with a smile, the companion presumes player A is feeling Pride. As such the potential empathic response of the companion will be to feel Pride as well. However, before the potential empathic emotion is added directly to the emotional state, its intensity is determined by the following modulation factors:

Similarity - represents the existent overlap between the agents of the empathic interaction, specifically in their emotional appraisals. It is determined by the degree to which the emotion elicited by the self-projection appraisal is congruent with the candidate emotions. The higher/lower the similarity, the stronger/weaker the empathic emotion will be.

Social Relation - represents the social bond that the empathic agent has with the other agent, namely how much he likes and cares for him. Like similarity, it enhances (in the case

of a positive bond) or decreases (in the case of a negative bond) the intensity of the empathic emotion.

Mood - represents an overall valence (positive or negative) of the agent's affective state. A negative mood increases the potential of a negative empathic emotion, and decreases the potential of a positive one. On the other hand, a positive mood works in an opposite manner.

Personality - indicates the agent's resistance to feel certain emotions. Empathic emotions to which the agent has a weaker/stronger resistance will be more/less likely to be added to the emotional state.

To exemplify the interplay of these factors, let's consider again the same scenario of player A winning the game against the player B and feeling Joy. What can happen to the companion in terms of empathy towards A? If when it projects himself into A's situation he elicits Joy as well, then their similarity will be high which will likely cause the companion to feel Joy for A. However, imagine that the companion strongly dislikes A; i.e. he has a negative social relation with him. In this case, it's much more unlikely that the companion will empathize with A. Finally, this is also true if the companion is in a really negative mood or his personality has a strong resistance towards feeling Joy.

2.3.2 Empathic Response

In our model, an empathic response (see Figure 2-6) starts with an emotion generated by the empathic appraisal. These emotions can also trigger empathic actions in the same way that other emotions trigger specific reactive behaviours. For example, when agent A feels Joy for B, it can potentially trigger the empathic action of smiling and congratulating B. Since the appropriateness of these actions can be highly dependent on the situational context, they are defined by a set of action rules that are domain-dependent. The empathic response process is performed in the Action Selection component of the architecture using Action Tendencies.

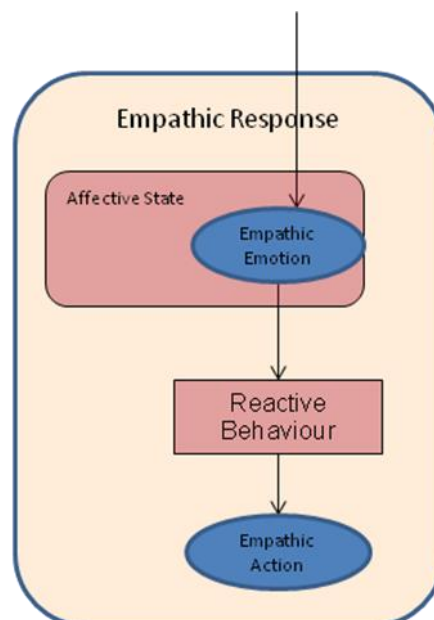


Figure 2-6 Empathic Response Process

2.4 Theory of Mind

As part of the requirement to have intelligent social behaviour, and to be able to establish and maintain social relations, the companion must have a theory of mind about the user and others. More concretely, it needs to have an idea of other's goals, the state of their current relation, the emotional state of others, and past experiences together. However, having a model of other's internal state is not enough. If we want our companion to reason about how future events affect the user's emotional state and social relations, we also need to model other's internal processes, such as appraisal, social update and reasoning.

Figure 2-7 shows a diagram of the model about others stored in the ToM component. The model includes data structures (such as Memory and Affective State), but also internal processes of others (appraisal, reasoning, social update). By examining the diagram, one can realize that the model of others corresponds to a new instance of the original architecture, with some small simplifications. The companion only has a one-level theory of mind, which means that the companion has a model of what agent B thinks, but does not know the model that B has of agent C. As such, the ToM model of others does not include the ToM component. Furthermore, we are not considering modelling the process of empathic appraisal in others.

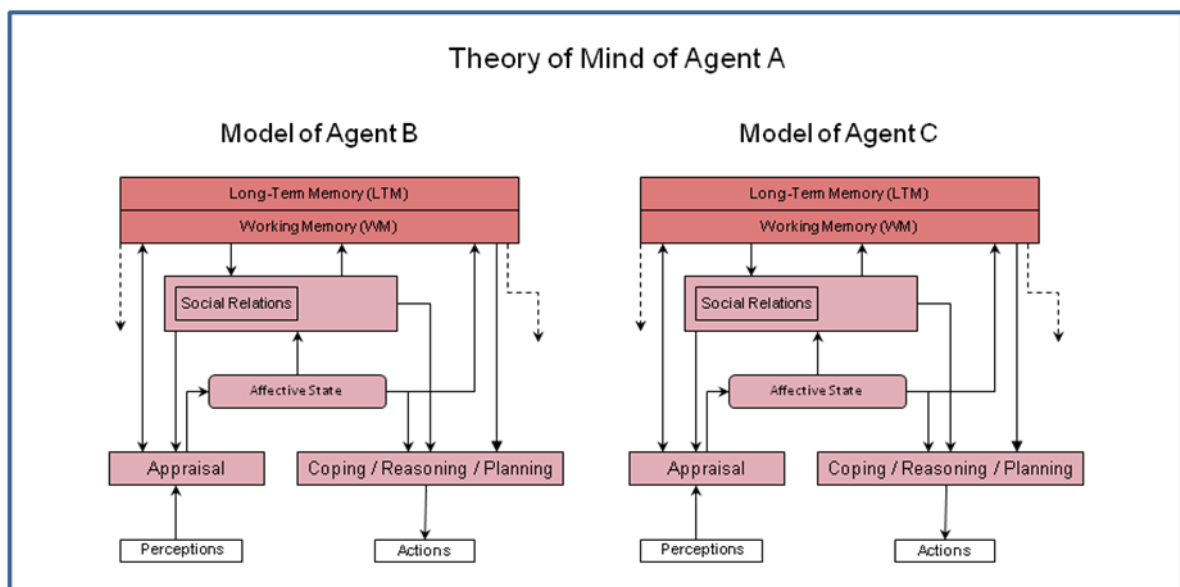


Figure 2-7 Theory of Mind models data structures and processes of others

Creating a model of others that includes processes in addition to data structures of others has two main uses. The first one is that by applying the processes we can determine the internal state of others (current values stored in data structures). When an event is perceived, the agent determines if other agents have perceived the same event (for instance, looking at which agents are near it). If so, it sends the event to the ToM component in order to simulate how others appraised the event. Then, the result of each individual appraisal is used to update the affective state, memory, and social relations of the corresponding agent's model.

The second function is to use the same processes to predict how others will appraise and react to a given future event. This is of paramount importance in order to reason and build up plans of actions that involve other's emotions and social relations. For instance, if the companion wants a user to increase his relation towards him, he considers actions that will be appraised positively according to the user's ToM.

2.4.1 Building up Models of Others

Initially, the companion starts with a basic model of others, where he assumes that others behave similarly to him. So, if the companion finds a given event undesirable, it assumes that someone else will also consider it undesirable. As time goes on, it will incrementally improve the models of others by observing their reactions to events. Whenever an event happens, the agent observes other's reactions, and performs a reverse process of appraisal as shown in Figure 2-8.

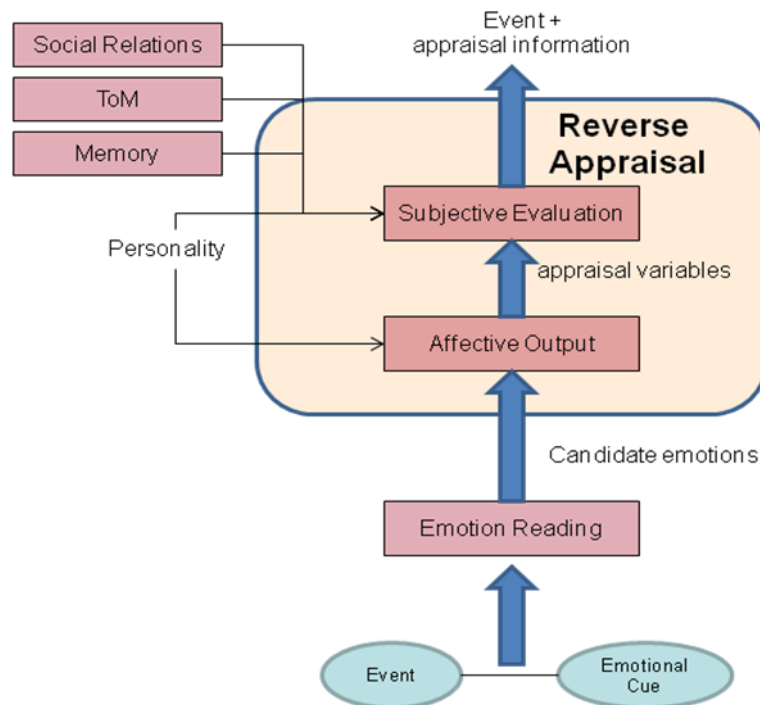


Figure 2-8 Reversing the Appraisal Process

Before the reversed appraisal takes place, the perceived emotional reaction of others is associated with the event. This information is used by the Emotion Reading module, which is the same module applied in the Empathic Appraisal, to determine a set of candidate emotions that could have caused the perceived reaction. As an example, if the user smiles, a possible candidate emotion is Joy.

After the emotional state is determined, the reverse appraisal (the appraisal component must support this mechanism) is used to calculate the value of appraisal variables. These appraisal variables represent how someone else appraised the event (desirable, undesirable, etc). For instance, if the user appears to be happy after an event, we can deduce that the event is desirable for the user. This information is used to update the model of others, by first associating the event to the determined appraisal information in Autobiographic Memory (AM). The idea is that in order to build up the model of others, it is important to remember their reactions (in terms of emotional reactions) to events. For instance, if an event happens and the agent perceives a smile from Agent B, then he stores that event in memory together with the information that agent B found the event desirable.

There are two advantages in using the AM component to build up the model of others. The first one is that it allows us to explore the effects of forgetting, given that the AM will implement forgetting mechanisms, in building up models of others. It will be possible for the companion to forget that an action was undesirable for a given user, and to repeat the action in future. We believe that, if used properly, this creates more believable behaviour. The second, and most important advantage, is that by using AM's spreading activation

mechanisms we can use a generalized model of others when not enough information is available. Consider the following illustrative scenario: the companion wants to estimate how agent A appraises an event E, however the companion has never seen agent A reacting to such event. Using only agent A's model it would not be possible to estimate it. However, the companion remembers seeing agent B, C and D reacting negatively to the same event. Therefore, it can generalize this knowledge and estimate that agent A will likely also find it undesirable. More details on how this is performed can be found in Deliverable D4.2.

2.5 Memory

Memory is a fundamental component and therefore is linked with all other processes of the companion. This is in part because of the requirement of a longer-term interaction where the agent needs to adapt itself to the user and to the history of interactions with the user. The long term memory is responsible for storing information about past interactions, and this information is used by the agent to influence the processes of appraisal, creating social relations, decision-making and the theory of mind, as mentioned before. The working memory is used to store the agent's current goals, current plans or action rules. Since the memory component is properly described in D4.2, we will not detail it in this document.

2.6 Reasoning and Action Selection

Reasoning and action selection are divided into two components in the architecture. The first component, Action Tendencies enables our companion to quickly react to particular affective states, whilst the deliberative component deals with the companion's goal based behaviour.

2.6.1 Action Tendencies

Action Tendencies represent the companion's innate reactions to particular affective states. For instance, crying when distressed or insulting someone when angry. The Action Tendencies component is then specified as a set of action rules (as seen in Figure 2-9).

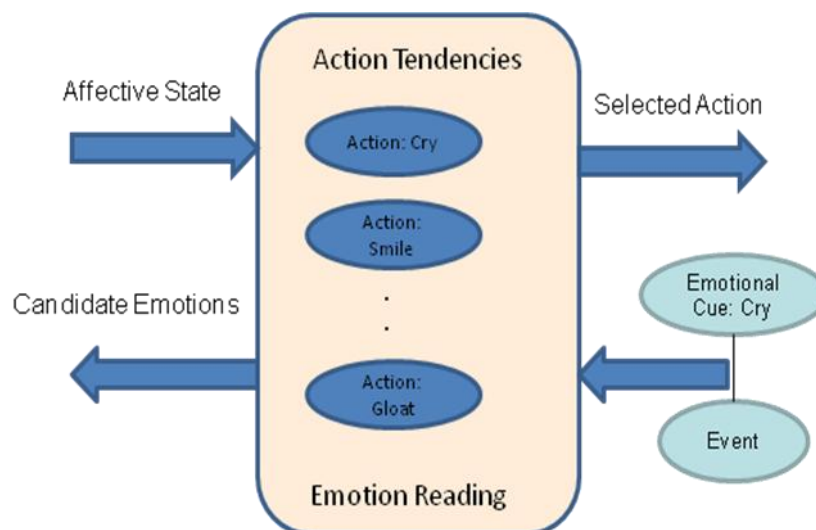


Figure 2-9 Action Tendencies component, also used for Emotion Reading

Action rules are usually domain-specific (depending on the type of actions available) and are authored beforehand. An action rule is defined with the following properties:

Attribute	Description
<i>Action</i>	The action performed if this action tendency is triggered
<i>Preconditions</i>	A set of preconditions that must be true in order to execute the action
<i>Eliciting Emotion</i>	The emotion that triggers this action tendency.

The action attribute identifies the action performed when the action rule is selected, the preconditions attribute specifies a set of preconditions that must be verified in order to activate the action rule, and finally the emotion attribute specifies the type of emotion that will trigger the action rule. The emotion attribute usually also specifies a minimum intensity of the emotion required to activate the rule.

Implementing the activation of action rules is straightforward; the Action Tendencies component starts by activating the rules that match the current affective state; the resulting rules' preconditions are then tested; finally, if more than one action rule is active, the one activated by the strongest emotion is selected as the action to be performed.

Given that the Action Tendencies component maps particular affective states into emotional expressions, it seems to be strongly related to the functionality of emotion reading: mapping emotional cues to affective states. As such, this component must also perform Emotion Reading by reversing the activation process (as shown at the bottom of Figure 2-9). When an emotional cue is received together with an event, this component tests the corresponding action rules that could have caused that emotional cue. After checking for preconditions, it returns the set of emotions (and minimum intensities required) that activate the remaining action rules. However, it is important to notice that when perceiving an emotional cue from others, the companion should not use his Action Tendencies component for emotion reading, but the Action Tendencies component of the corresponding ToM model. The idea is that different persons or agents might express affective states differently.

2.6.2 Deliberative Behaviour

The architecture's deliberative behaviour is based on the activation and pursue of predefined goals similarly to FAtiMA. Deliberation (deciding what state of affairs the agent wants to achieve) takes place first by checking which goals have their preconditions satisfied and then selecting the most relevant one according to the current state of the world. After a goal is selected, means-ends reasoning is performed in order to build up a plan of actions to go from the current state of the world to the desired state. The details of these processes will be described later in an implementation document. Here, we will focus on describing how social behaviour is achieved by integrating explicit knowledge about emotional appraisal and social relations into means-ends reasoning.

Additionally to ordinary goals, the deliberative component has explicit goals about social relations and affective states. These goals may specify preconditions related to the current state of social relations and affective states. For example, one can specify a goal activated when the user's like relation drops below a given value or when the user's affective state is sadness. Then, the goal will be to increase the user's social relation or to make the user become happy.

In order to achieve such goals, the planner must be able to reason about how the social update and appraisal processes work. Therefore, we need to model these processes into STRIPS like operators to be used by the planner. Figure 2-10 shows several examples of how appraisal rules and action tendencies can be mapped into STRIPS operators. An appraisal rule maps an event to appraisal variables (it corresponds to the first step of the appraisal process). In the first example provided, when the companion perceives an event

where he wins the game to agent [y], this is considered very desirable for the companion and very undesirable for agent [y]. This appraisal rule is modelled as a special operator *Self:perceives(Self,win,[y])*, with the event *Event(Self,Win,[y])* as precondition and the resulting appraisals as the effects, *Appraisal(Self,Self,Win,[y],VeryDesirable)*⁷ and *Appraisal([y],Self,Win,[y],VeryUndesirable)*⁸. An operator's name is preceded by the name of the agent who will execute the operator. In this case, the operator corresponds to a perception of the companion (hence the name Self).

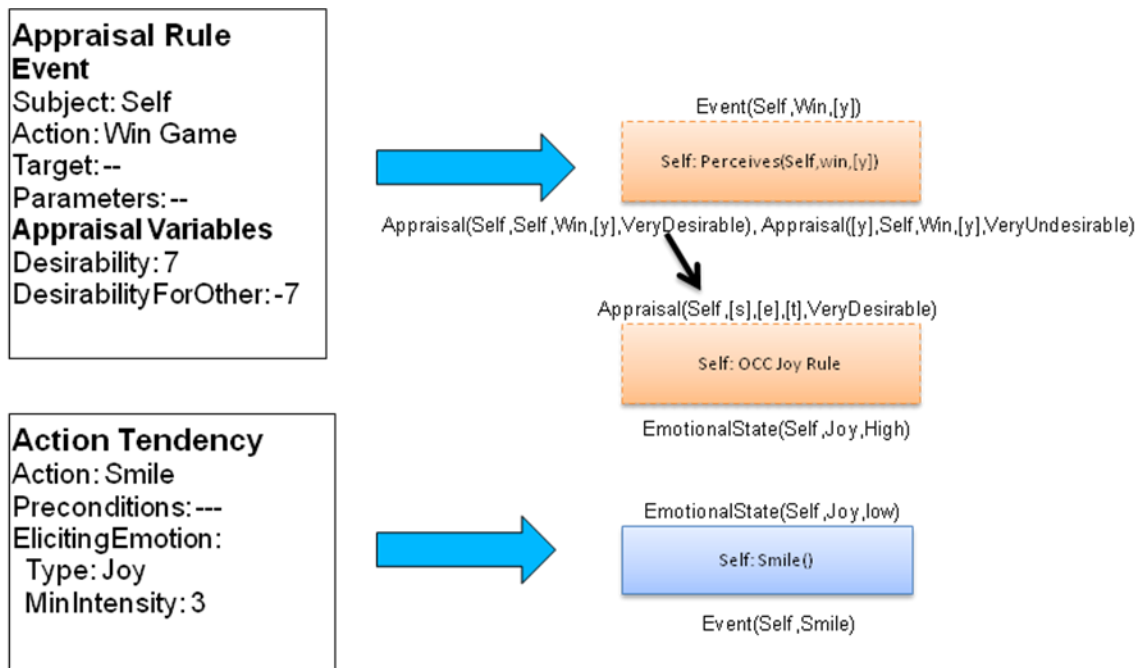


Figure 2-10 Mapping appraisal and action tendencies into operators

It is also necessary to model the second stage of the appraisal process as planning operators. The second example shown in Figure 2-10 represents an OCC rule that maps desirable events (whatever the event is) into a joy emotion. The precondition *Appraisal(Self,[s],[a],[t],VeryDesirable)* means that the companion must have appraised an event performed by a subject [s], referring an action [a] to a target [t], and the resulting appraisal was very desirable. The arrow between the first two operators represents a causal link in planning terms, meaning that the first operator achieves the second operator's precondition by replacing the variables [s],[a],[t] with *Self,Win,[y]*. The effect of this operator, *EmotionalState(Self,Joy,High)*, represents that a new emotion of type Joy with a high intensity will be added to the companion's emotional state.

Both the appraisal rule and the emotion generation rule operators are represented with a dashed line, meaning that these operators do not correspond to real executable actions, but to internal processes of the architecture.

⁷ This effect represents that the companion (Self) appraises the event *Self,Win,[y]* as very desirable.

⁸ This effect represents that agent y (who lost the game) appraises the same *Self,Win,[y]* as very undesirable. The notation used is temporary and may be changed in a future deliverable. One open question is whether we should use discrete values for appraisal variables and emotion intensities. Using discrete values facilitates the planning procedure, but we lose information in the process. On the other hand, numeric values will require us to extend the planning algorithm used.

Action tendencies are also mapped into operators. The rationale is that the companion must include emotional expressions of others in its plans, since they are the only way for the companion to be sure whether his plan of making someone happy succeeded or not. The third example in Figure 2-10 shows a smile action tendency mapped into an operator *Self:Smile()*. The preconditions of the action rule and the triggering emotion are put as the operator's preconditions (in this case *EmotionalState(Self,Joy,low)*) and the effect corresponds to the execution of the action. Unlike the previous two operators, Action Tendencies represent executable actions, and as such, if the plan execution algorithm encounters an operator of this type, it sends it for execution (in the case that the action is supposed to be executed by the companion), or waits for the action to happen (in the case that the action is supposed to be performed by someone else).

Similarly to the previous components, the social update processes is also modelled. The deliberative component contains a set of predefined operators that map positive emotions into increased social relations, and negative emotions into lowered social relations.

Finally, it is also worth mentioning that other components also have a strong influence in the deliberative component. For instance, success and failure of goals and actions is stored in memory and later used to estimate likelihood of success. This value is used to help select between alternative goals and plans of actions. The current affective state also influences coping strategies applied, such as giving up unlikely goals. A positive mood makes the companion to give up goals more easily, while a negative mood makes it harder to give up goals.

2.6.3 Illustrative Example

We will give a brief example of what kind of reasoning an agent can do with this information. Suppose that the companion has the goal of making the user happy. In this case, the goal will be to make the user to smile. The corresponding plan is shown in Figure 2-11. Using the ToM model (in particular the Action Tendencies component) about the user, the companion knows that the user will smile if it is in a state of joy. The planner knows that joy is caused by a desirable event, and considers all actions desirable for the user (cheering, telling a joke, inviting to see a movie, etc). In the plan it is possible to see the special operators that predict how the user will react when perceiving a cheer-up event from the companion. Assuming that things will turn out as expected, the planner then selects the cheer-up action to be executed. If everything goes as planned the companion will succeed in making the user smile. However, if the user does not smile, the companion will update the user's model about the cheer-up action, and will either try something else or eventually fail/give up to make the user happy.

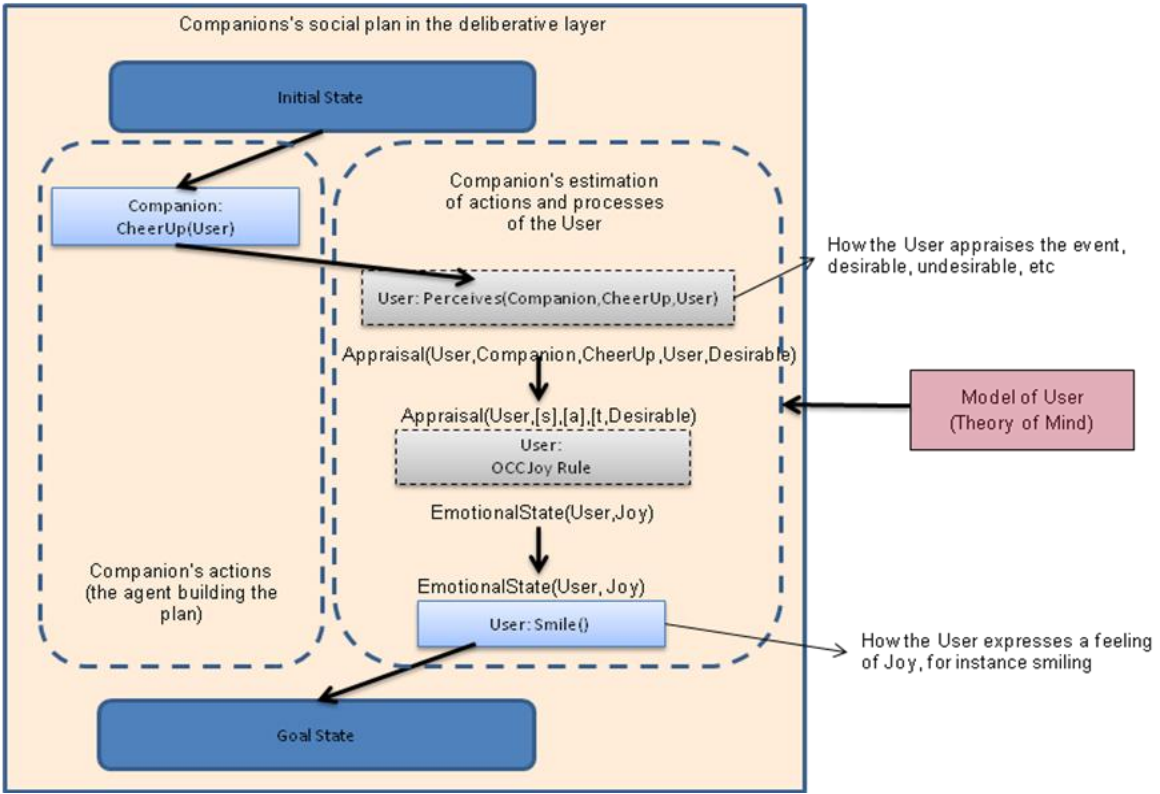


Figure 2-11 Building up a plan to make the user smile.

3 Scenario

Given the complexity of the architecture depicted in chapter 2, we will now describe a more complete example of how the architecture can be applied and tested in a particular scenario. The scenario we are going to describe is an extension to the MyFriend scenario. The main difference is that instead of having iCat as a chess player that plays against the user, we will have two users playing the game. iCat will then act as a friend of one of the users, watching the game and making comments about it to his friend (as depicted in Figure 3-1).

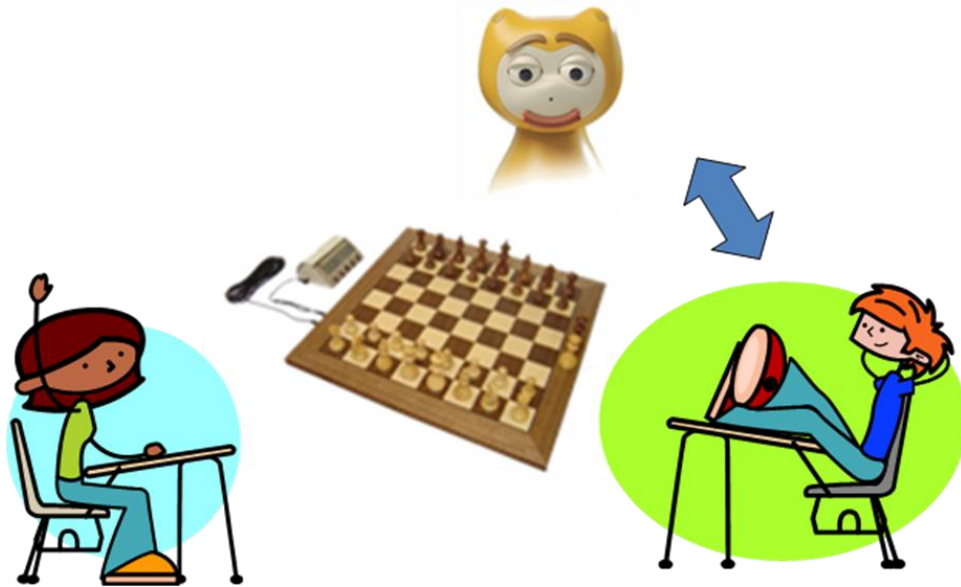


Figure 3-1: Extension to the MyFriend scenario. iCat watches the game and talks with his friend.

For simplification purposes the companion will only interact with and build up a model of one of the users. Although the architecture can cope with more than one user, we have limitations with the number of cameras we can install, and with the facial recognition component.

The advantage of this scenario is that it simple to build (the scenario) from what we have at the moment, and at the same time it is complex enough to test some of the proposed components. Moreover, it already poses interesting challenges, such as building up a correct model of the user based on the recognition of some visual cues and contextual features.

3.1 Applying the Companion's Architecture to the scenario

Our aim with this scenario is twofold: to evaluate the user's reaction to iCat as a game companion instead of an opponent; and to evaluate a first implementation of the architecture focusing in some components. In particular, we want to test the impact of the empathic model of appraisal in the user's relation with the companion. We will create a version of the companion where he makes comments about the game, without expressing emotions (since he is not playing he does not consider any move desirable or undesirable) and compare it with another companion that experiences empathic emotions according to what the user is feeling. A second test will focus on the companion's capability to deliberately change his friend's affective state⁹, and whether the companion is correctly adapting the strategies used to particular users. We also want to test the impact of this capability in the user's relation.

⁹ For ethical reasons, we are considering only positive changes in the user's affective state.

The companion perceives the state of the chess game and the visual cues from his friend's face through one or two cameras pointed at him. This information is sent to a multimodal affect sensitivity module (for more details on this module refer to Deliverable D3.2), which then sends two measures to the level 3. The first measure is the user's engagement level with the iCat (whether the user is looking at the iCat much or not). The second one, tell us if the user is having a positive or negative emotional feeling. In order to correctly follow the companion's architecture, it would be better to detect particular facial expressions (happiness, frowned face), and then use the emotion reading component to assess the candidate emotions. Unfortunately, it is still difficult to detect the user's exact type of expression with a high level of accuracy rates.

Appraisal follows iCat's nine sensations model (Leite et al 2008), which are generated from the state of the chess game, the move expected and the actual last move played. Since the companion is not playing the game directly, the appraisal of the state of the chess game will not generate any affective states. Instead, the empathic appraisal will take a preponderant role in this scenario. This empathic appraisal will use the sensations modeled in the companion's affective state and the emotional cues received from the user (mainly positive, neutral and negative emotion).

At the same time that the candidate emotions are determined, the companion performs self-projection appraisal and appraises the chess board to see how he would react if he was playing. And now something interesting may happen, it is possible for the iCat to appraise the state of the chess game differently: while the user thinks he's done a good move, iCat realizes that it was a very bad move indeed. This will influence the resulting empathic emotion. With the candidate emotion and self-projection determined, the empathic emotion is determined by applying the modulation factors as mentioned in chapter 2. A set of predefined action tendencies is then used to map the affective sensations into particular iCat's facial expressions and/or speech.

There are two types of events stored in Autobiographic Memory: the moves played, which are associated with the corresponding chess state; and the companion's interactions with the user. These events are associated with the user's affective state experienced at the time. As an example, the AM will remember the user being sad for losing a queen to his opponent, or becoming happy because the companion said something positive to him. After several games, and after applying forgetting mechanisms, the companion will only remember the most relevant moves and events.

As for the deliberative component, there are two main goals modeled. The first one is to increase/maintain the relation with the user, by doing small talk, making positive comments about a user's move, and even talking about similar past moves/past games stored in Autobiographic Memory. The second goal is activated when the user appears to be sad because he is losing the game. The companion will try to apply a set of strategies to make the user become happier. The ToM component will be very important in order for these goals to work properly. Suppose that the user does not like to be patronized when loosing. After trying to cheer up the user a few times, the companion understands that the user responds even more negatively to that interaction, and thus will stop trying to apply that particular strategy. The companion learned that such particular action will not make the user happier, and thus will not achieve his goal. From that moment on, he will either use alternative strategies, or if none work, just give up on cheering him up. We believe that this kind of adaptive behavior is very important in long-term interactions with users.

3.2 Limitations of the scenario

Unfortunately, this scenario also poses some potential problems and limitations. One of the most concerning ones, is that a high error rate in the affect sensitivity component will have

strong effects in the results obtained, thus making it difficult to build up a proper model of the user, or even to detect the correct affective state to generate an appropriate empathic emotion.

One additional problem is that there is only one external entity modelled, the user, which cannot talk or interact directly with the companion (and thus change the companion's relation towards him). As such, we cannot correctly explore the dynamics of social relations. Also, we cannot properly test negative relations, with agents performing negative actions to others.

Taking into account the limitations of the scenario proposed, and in order to properly test all the components of the companion's architecture, we are currently studying how to apply the architecture in other scenarios.

4 Conclusion

With the aim of developing companions that can act and respond to the human user, in an intelligent and socially appropriate manner within a long-term period, we have put forward a first specification of the companion mind's architecture. In order to properly address our goal, we identified several cognitive competences that needed to be mapped into the architecture: social intelligence, emotions and personality, empathy, theory of mind, memory and adaptation.

These cognitive competences, previously studied in D5.1, were then mapped into the several components of the architecture, and a set of interconnections between the components was identified. There has been a special concern with the agent's social behaviour, in particular the capability of experiencing empathic emotions and the capability of establishing and maintaining long-term social relations with users. To do so, the companion needs to model explicit social behaviour, or in other words, it needs to have explicit social goals and knowledge of available social actions and the corresponding effect they have in a social relation, and use that knowledge and the current social relations to achieve such goals. Moreover, the companion must adapt its behaviour according to the user over time, and as such, it must build up a model of how the user behaves and reacts to events, and use that model to select the more appropriate social actions.

Finally, we have presented a test scenario, where we plan to use the companion as a friend of a user playing a chess game against another user. The companion does not play the game, but interacts with his friend talking about the game. The goal is to test an initial version of the components of the architecture, and evaluate the user's reaction to the companion's empathic and social behaviour.

5 References

Bickmore, T. & Picard, R. (2005) Establishing and maintaining long-term human-computer relationships, *ACM Transactions on Computer-Human Interaction*, ACM Press, 12, 293-327.

Dias, J. (2005) Fearnot!. *Creating emotional autonomous synthetic characters for empathic interactions*. Master's thesis, Universidade Técnica de Lisboa, Instituto Superior Técnico, Lisboa.

Dias J., Paiva A. (2005). Feeling and reasoning: a computational model for emotional agents. In *Proceedings of 12th Portuguese Conference on Artificial Intelligence, EPIA 2005*, pages 127-140. Springer.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. NY: Wiley. 1958.

Hoffman M. L. (2000), *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press, 2000.

Lazarus, R. (1991). *Emotion & Adaptation*, Oxford University Press.

Leite, I., Pereira, A., Martinho, C., Paiva, A. (2008). Are emotional robots more fun to play with? 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008), pp.77-82, IEEE Computer Society.

Oatley, K., Keltner, D., Jenkins, J. (2006). *Understanding Emotions*, Blackwell Publishing.

Ortony, A., Clore, G., Collins, A. (1998) *The Cognitive Structure of Emotions*. Cambridge University Press, UK.

Preston, S., Waal F. (2002) Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1):1–71.

Rodrigues S., Mascarenhas S., Dias J., Paiva A.. (2009) I can feel it too!: Emergent empathic reactions between synthetic characters. To appear in proceedings of Affective Computing and Intelligent Interaction (ACII) 2009, Springer

Vignemont F., & Singer T. (2006), The empathic brain: how, when and why?, *Trends in Cognitive Sciences*, vol. 10, pp. 435-441.