WASABI as a case study of how misattribution of emotion can be modelled computationally

Christian Becker-Asano[1] and Ipke Wachsmuth[2]

1. Intelligent Robotics and Communication Labs. Advanced Telecommunications Research Institute International, 2-2-2 Hikaridai, Keihanna Science City, Kyoto, Japan, 619-0288. Email: christian@becker-asano.de
2. Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany. Email: ipke@techfak.uni-bielefeld.de

Summary

Cognitive scientists, psychologists, neuro-biologists, and computer scientists achieved significant progress in understanding and modelling the fuzzy concept ‘emotion’ and more general ‘affect’. Accordingly, a variety of computational realizations, discussed by Marsella, Gratch, and Petta in **Chapter MGP** of this volume, stem from a number of different psychological theories and philosophical conceptions. As correctly classified in **Chapter MGP** the computational realization we propose, labelled WASABI ([W]ASABI [A]ffect [S]imulation for [A]gents with [B]elievable [I]nteractivity), performs a mapping of appraisal outcome into a three dimensional space of pleasure, arousal, and dominance or PAD space in short, and it thereby ‘breaks the link’ between the internal representation of affect and its external domain object. Accordingly, we will present and discuss here, how the phenomenon of post-hoc misattribution, i.e., a mismatch between an emotion’s objective and its subjective cause, can be modelled and explained by the WASABI architecture.

The central idea of this architecture is to combine two dimensions, namely emotional valence and valence of mood, such that their mutual influence generates a continuously changing, self-rebalancing internal state, which can be interpreted as constituting a very basic, non-relational, short-term memory of affect. Whenever some external or internal event (the latter, for example, resulting from cognitive reasoning processes) is appraised as having an emotional effect, this effect is translated into an impulse of emotional valence, which then disturbs the internal

emotion dynamics. At the same time internal cognitive reasoning further analyzes the event to determine, if it is a candidate for elicitation of an emotion. In the current state of the architecture this reasoning is limited to the generation and checking of expectations within the context of a well-defined interaction scenario serving as proof of concept. The emotions are represented in PAD space such that a particular emotion can only be elicited (or in more philosophical terms 'become aware to the agent') if the agent's current internal feeling state represented in PAD space allows for it.

Although this architecture is already considerably complex, we admit that this is only our first attempt to grapple with the complex dynamic interplay of cognitive and bodily processes from which emotions are assumed to arise. Accordingly, we hope that the WASABI architecture, on the one hand, provides fruitful impulses to the interdisciplinary endeavour of understanding human emotionality, and, on the other hand, can serve as one example of a blueprint for how to increase a conversational agent's affective competency.

1 Introduction to WASABI's core ideas

As pointed out by Marsella, Gratch, and Petta in **Chapter MGP** a variety of computational models of emotions stem from different psychological and philosophical theories. When we started developing our own computational model of affect, which later was entitled WASABI, it was tempting to follow the ideas and conceptions of the by then famous structural model of emotions (Ortony, Clore, & Collins, 1988), or OCC model in short, as many other computer scientists had done before. The limitations and problems of this model, however, had already become apparent (Bartneck, 2002) so that we decided to first concentrate on modelling the temporal dynamics of emotions instead (Becker, Kopp, & Wachsmuth, 2004). We furthermore limited ourselves to only simulate the temporal unfolding of an emotion's intensity, postponing the question of how to realize cognitive appraisal. We also did not follow the 'basic emotions' idea (Ekman, 1999), but instead combined simple hedonic valence with a very basic conception of positive versus negative mood. These two dimensions were arranged to form an orthogonal space, which is labeled 'emotion dynamics' and can be found in the lower left corner of **Figure 1**. Within this space the mutual interaction between hedonic valence (represented on the x-axis) and mood (represented on the y-axis) is simulated such that a so-called emotion dynamics continuously unfolds over time.

Insert Figure 1 about here.

Eventually we added a third, orthogonal dimension to this space to account for those cases when nothing emotionally relevant happens over a certain period of time. This z-axis is labelled ‘boredom’ to indicate that any value along this axis represents an agent’s level of boredom (see **Figure 2**).

Insert Figure 2 about here.

The temporal unfolding and mutual interaction of emotion and mood is realized as follows:

1. The x-value is interpreted as a gradient, in relation to which the y-value increases or decreases. The more positive the value on the x-axis, the faster the y-value increases; the more negative the x-value, faster the y-value decreases. Speaking in the language of affective sciences, this models a fortifying and alleviating effect that emotions are assumed to have on moods, which is graphically indicated by the white up and down arrows in **Figure 2**.
2. Any non-zero value on either of the two axes is constantly pulled back to zero by applying two simulated forces F_x and F_y, which are exerted by two independently simulated spring-mass systems virtually attached to the reference point, see **Figure 2**. In terms of modelling affect, the simulation parameters are normally chosen such that in case of equal displacements for x and y the reset force F_x is greater than the reset force F_y, because emotions are commonly considered to last less long than moods.

With this basic setup in place, which is described in more detail in (Becker, Kopp, & Wachsmuth, 2007), a single emotional impulse of hedonic valence (see **Figure 1**) is sufficient to start the emotion dynamics. The impulse causes an instantaneous displacement of the reference point and an agent’s internal emotional state will change dynamically over time (as indicated by the dashed line connecting the coordinate system's origin with the point of reference in **Figure 2**) until it reaches the point of origin again, if no further impulses arrived in the meantime. When for a longer period of time no emotional impulses disturb the emotion dynamics, the z-value changes linearly to simulate an agent's increasing level of boredom.

Next, the three values x (emotion), y (mood), and z (boredom) need to be integrated in order allow for mapping them on named emotions, which are finally transmitted back to the appraisal module (see **Figure 1**). At this point, we decided to map into pleasure-arousal-dominance space, PAD space in short (Mehrabian A. , 1995), as described by the following equations:

$$PAD(x_t, y_t, z_t) = \left(p(x_t, y_t), a(x_t, z_t), d(t)\right), with$$

Equation 1

$$p(x_t, y_t) = \frac{1}{2} \times (x_t + y_t) \; and \; a(x_t, z_t) = |x_t| + z_t \; (with \; z_t \leq 0)$$

Insert Figure 3 about here.

All variables in **Equation 1** are indexed with t, because the emotion dynamics is updated 25 times per second to achieve a seemingly continuous simulation of internal feeling state. In PAD space primary emotions are located as indicated by the crosses in **Figure 3** and secondary emotions occupy areas on the levels of high and low dominance. The smaller the distance between an agent's emotional state, as represented by the continuously updated PAD triple *PAD*(x_t, y_t, z_t) (see **Equation 1**), and any of the primary emotions, the more likely the agent gets aware of this emotion with an intensity that is inversely proportional to this distance. If the agent's emotional state enters a region representing a secondary emotion, which was triggered just before, then this secondary emotion gets aware to the agent with an intensity derived from its intensity distribution in PAD space; see (Becker-Asano & Wachsmuth, 2009) for details.

The motivation for this quite complex interplay between cognitive reasoning and emotion dynamics will be clarified in the light of the interdisciplinary background and it will be contrasted with related work in affective computing. Basically, we can explain our motivation in relation to Marsella et al.'s conceptions (see **Chapter**

MGP, this volume). The dynamic simulation explained so far makes detailed representational and process commitments for affect-derivation, but leaves open how an agent's relationship with the external world influences its appraisal of events, other agents, or objects within that world. An emotional impulse can be derived from any kind of cognitive appraisal process, but it might also by product of hard wired, reactive perception-action patterns. From an engineering point of view, this flexibility allows for the core emotion dynamics to be combined easily with different computational architectures as long as they feed the emotion dynamics with emotional impulses, trigger primary and secondary emotions whenever appropriate, and make reasonable use of the set of aware emotions they receive in return. From a psychological point of view, we naturally assure mood-congruency of emotions, because, e.g., only in case of bad mood negative emotional impulses have the effect of eliciting anger in the WASABI architecture. When the agent is in good mood, in contrast, negative impulses will only dampen the good mood first, before it might get negative enough to allow for the elicitation of negative emotions. Furthermore, by decoupling the domain object from the subjective emotional response, posthoc reasoning about what might have been causing an agent's current emotional state can reasonably be performed. In result, the subjective cause might not match the objective cause and, thus, an agent driven by the WASABI architecture is susceptible to misattribution.

2 Interdisciplinary background

Our approach to modelling affective competency for our virtual human 'Max' is derived from and relates to a multitude of different ideas, conceptions, and theories of psychologists, cognitive scientists, philosophers, and neurobiologists. WASABI's core of simulating an emotion dynamics in three-dimensional affect space can be traced back to the ideas of the philosopher Wilhelm Wundt (1922/1863), who claimed that any emotion can be characterized as a continuous progression in such a three-dimensional affect space. By now the validity of dimensional theories of affect is widely accepted and the interested reader might kindly be referred to (Becker-Asano C. , 2008) for an introduction to the history of this class of theories.

In the following, however, we will concentrate on explaining how our architecture relates to Scherer's Component Process Model (CPM, see **Chapter S** of this volume), from which it derives a number of ideas. Afterwards, the neurobiological background of the WASABI architecture will be outlined in order to explain the rationale for distinguishing two classes of emotions, namely primary and secondary emotions.

2.1 WASABI in relation to the CPM

Scherer distinguishes the following five functions for the theoretical construct of emotions in the context of the CPM (see **Chapter S** of this volume):

1. Evaluation of objects and events
2. Regulation of internal subsystems
3. Preparation for action
4. Signalling of behavioural intention
5. Monitoring of internal state and external environment

Although the WASABI architecture is not directly derived from Scherer's CPM, we follow the above distinction and believe that we can account for a subset of these functions as follows (cp. **Figure 1**):

<u>Evaluation of objects and events:</u> Appraisal processes that enable our agent to evaluate external objects and events are realized in a software module, which is based on the Belief-Desire-Intention (BDI) approach to modelling rational reasoning (Rao & Georgeff, 1991). In this module goals and plans are explicitly represented, expectations are generated, and current events are evaluated against previous expectations. This cognition module, which contains the agent's *Appraisal module*, will be explained in **Section 4.1** in the context of the general explanation of our virtual agent Max.

<u>Regulation of internal subsystems:</u> According to Scherer (2001) this function is served by the 'peripheral efference component' and in (Scherer, 1984) 'the physiological component of activation and arousal' is made responsible for this function. Therefore, we assume that our simulation of emotion dynamics, which is driven by external and internal forces and continuously updates an agent's arousal level, can—at least to some respect—fulfil this function.

<u>Preparation for action:</u> This function is realized in the WASABI architecture by letting our agent's breathing and eye blinking frequency be modulated by the simulated arousal, which might be interpreted as 'preparing for action' by an outside

observer. We have to admit, however, that we do not explicitly model ‘behaviour tendencies’, which are postulated by Scherer (1984) as being part of this function.

Signalling of behavioural intention: Our agent’s facial expressions are directly driven by the primary emotions of the WASABI architecture such that the ‘motor component’ is realized quite straight forward. Furthermore, in case of secondary emotions such as hope or relief the agent’s cognition generates appropriate verbal expressions, which are seamlessly combined with non-verbal expressions driven by primary emotions.

Monitoring of internal state and external environment: Although the WASABI architecture simulates a continuously changing internal state through the implementation of an emotion dynamics, we do not explicitly model a monitoring process that, according to Scherer (**Chapter S**, this volume), is necessary to achieve subjective feeling states. We believe, however, that our architecture is a promising candidate for extensions toward the simulation of such subjective aspects of emotions.

In addition to these functional similarities the WASABI architecture as it is outlined in **Figure 1** can conceptually be divided into two modules, which are comparable to two of the three CPM modules presented by Scherer in **Figure 1** on page **XX** of this volume.

Scherer’s *Appraisal module* consists of one sub-module labelled ‘*Multilevel appraisal*’, which is responsible for very sophisticated and detailed ‘*sequential evaluation checks’*. In contrast, the computationally realized *Appraisal module* of the WASABI architecture permits much less sophisticated appraisal than proposed by the CPM. Nevertheless, we believe that we can account for some of the proposed evaluation checks as will be detailed in **Section 3** where our agent’s cognitive reasoning abilities are described.

Although the *Component patterning module* is omitted in our architecture, the changes within the emotion dynamics part of the *Integration/Categorization module* (see **Figure 1**) can be understood as to simulate *physiological responses*, which are part of Scherer’s module. In addition, *Motor expression* of emotions is achieved within the *Integration/Categorization module* of the WASABI architecture as well, after emotion, mood, and boredom have been mapped into PAD space as described above.

By representing emotions in PAD space we also account for *Categorization/Labelling*, which is part of Scherer’s third *Categorization module* (see

Figure 1, Chapter **S** on page **XX**). We lack, however, a *Central representation of all components*, unless one argues that this representation is achieved by the dynamically changing, emotion-related belief-structures within the BDI-based *Appraisal module*.

2.2 Primary and secondary emotions

A major difference between the WASABI architecture and the CPM consists of the distinction of two classes of emotions in WASABI, primary and secondary emotions. These two classes are derived from neurobiological research findings of Damasio (1994).

Primary emotions are supposed to be innate and they are understood as prototypical emotion types, which can already be ascribed to one year-old-children (Damasio, 2003). Secondary emotions are assumed to arise from higher cognitive processes and to be acquired during ontogenesis through learning processes in a social context. Damasio (1994) uses the adjective ‘secondary’ to refer to ‘adult’ emotions, which ‘utilize the machinery of primary emotions’ by influencing the acquisition of ‘dispositional representations’, which are necessary for the elicitation of secondary emotions. These acquired dispositional representations, however, are believed to be different from the ‘innate dispositional representations’ underlying primary emotions.

In the WASABI architecture this representational difference is reflected in the following two ways:

1. The PAD space representation of secondary emotions is much less precise than that of primary emotions (see **Figure 3**), because the former require much more elaborate cognitive reasoning than the latter.
2. Appraisal processes do not necessarily need to trigger primary emotions, before they can be elicited in PAD space. For secondary emotions to be elicited, however, this triggering in PAD space is a necessary precondition, as will be explained in Section 4.2.

We follow Damasio’s distinction, because it allows us to start with a set of more ‘primitive’ primary emotions, which can already be elicited by fast, hard-wired perception-action patterns without the need for complex deliberation. This is, of course, mostly a rather technical motivation, but doing so might eventually allow us to investigate developmental aspects of emotions. In fact, the results of an empirical study confirmed the hypothesis that an agent simulating secondary emotions in addition to primary ones is judged older than one that only simulates primary emotions (Becker-Asano & Wachsmuth, 2009).

3 The virtual human Max

The virtual human Max (see **Figure 4, left**) developed at Bielefeld University's Artificial Intelligence Laboratory has been employed in a number of scenarios in which Max's conversational capabilities have been steadily extended. In a museum application, Max is conducting multimodal smalltalk conversations with visitors to a public computer museum. In this setting, the emotion dynamics leads to a greater variety of often unpredictable, yet coherent emotion-coloured responses, which add to the impression that the agent has a unique personality. Furthermore, the WASABI architecture has also been applied to a gaming scenario, in which secondary emotions were simulated in addition to primary ones.

In the following we give a brief overview of our agent's cognitive architecture, before we explain in detail how different levels of appraisal are realized inside of it.

4.1 The architectural framework of Max

Insert Figure 4 about here.

Max has been developed to study how the natural conversational behaviour of humans can be modelled for face-to-face encounters in Virtual Reality. The cognitive architecture of Max (see **Figure 4, right**) realizes and tightly integrates the faculties of perception, action, and cognition required to engage in such interactions (Leßmann, Kopp, & Wachsmuth, 2006). Although in general it employs the classical perceive-reason-act triad, all processes of perception, reasoning, and action are running concurrently within the architecture.

Reflexes and immediate responses to events are realized by a reactive connection between perception and action. Such fast-running stimulus-response loops enable Max to also react to internal events and his reactive behaviours include gaze tracking as well as focusing the current interaction partner in response to prompting signals. In addition, continuous secondary behaviours reside in this layer, which can be triggered or modulated by deliberative processes and by the emotional state of the agent. These behaviours let Max appear more lifelike and they include eye blink, breathing, and sway.

Perceptions are also fed into deliberative processes which are responsible for interaction management by interpreting input, deciding which actions to take next, and composing behaviours to realize them. This reasoning is implemented following the BDI approach to modelling rational behaviour and makes use of an extensible set of self-contained planners. The architecture further comprises a cognitive, inner loop, which feeds internal feedback information upon possible actions to take back to deliberation.

4.2 Three different levels of appraisal

Insert Figure 5 about here.

By exploiting the functionality of the architectural framework described above, three levels of appraisal are computationally realized for Max within the *Appraisal module,* see **Figure 5**. *Reactive* as well as *Cognitive appraisal* serve as input for the *Integration/Categorization module* whereas *Cognitive reappraisal* evaluates the resulting set of aware emotions in the light of cognitively represented situational context information.

a) Reactive appraisal: This sub-module realizes aspects of the first appraisal objective postulated by Scherer’s CPM (see **Chapter S**, this volume). Max uses a look up table to assess an event’s intrinsic pleasantness and checks, if the event complies with his expectations. The intrinsic pleasantness directly translates into an emotional impulse to be sent to the emotion dynamics. Only if the event is unexpected, the primary emotion ‘surprise’ is triggered in PAD space (see **Figure 1**), such that Max is not surprised by events he could expect to happen. A similar assessment lets the *Appraisal module* trigger the primary emotion ‘fear’, when Max expects some negative event to happen in the near future.[1]

[1] ‘Fear’ is a very special primary emotion within the WASABI architecture, because it is the only prospect-based emotion that does not belong to the class of secondary emotions. Accordingly, a strict distinction between primary and secondary emotions is sometimes problematic and we hope our architecture can serve as basis for further discussion.

b) Cognitive appraisal: Central to this sub-module is the evaluation of an event's goal conduciveness (or its obstructiveness, respectively). As described in **Chapter S** in the context of Scherer's CPM an intrinsically pleasant event (e.g., the delicious cake offered by a friend) can nevertheless be negative, if it hinders an individual from achieving a higher-level goal (e.g., sticking to a diet). With respect to secondary emotions, deliberative reasoning about goal-conduciveness of possible future events takes place in this module as well. This prospect-based deliberation might give rise to emotions such as 'hope' or 'fear' ('hope' being considered a secondary emotion) and after an undesired expected event is confirmed or disconfirmed, the secondary emotions 'relief' or 'fears-confirmed' are triggered in PAD space, respectively.

   These appraisal processes are both part of the second appraisal objective in Scherer's CPM. Another appraisal target of this sub-module—changing the agent's 'dominance'—is considered to be part of appraisal objective three in the CPM. We use an agent's 'dominance' similar to Scherer's conception of 'power' and 'control', in that it reflects our agent's level of control over the situation as well as his social status. For example, whenever it is Max's turn in a game, the *Appraisal module* changes his level of dominance to maximum and vice versa.

These appraisal mechanisms generate all necessary input for the *Integration/Categorization module* (see **Figure 1**). Emotional impulses are derived from reactive and cognitive appraisal, primary emotions are triggered in effect of reactive appraisal, and, finally, secondary emotions are triggered and the agent's level of dominance is changed as product of cognitive appraisal.

In result, the *Integration/Categorization module* eventually sends back to the *Appraisal module* a set of aware emotions with their respective intensities. These primary and secondary emotions are then reappraised in the *Cognitive reappraisal* sub-module (see **Figure 5**). One target of this reappraisal is the assessment of coping potential (cp. Scherer's third appraisal objective of the CPM, **Chapter S** of this volume). We have to admit, however, that our implementation of coping behaviour so far is rather simple. In the museum guide scenario Max leaves the scene whenever he got very angry and only comes back after he has calmed down again. This 'calming down' results from WASABI's internal emotion dynamics, which drifts back to zero automatically in the absence of emotional impulses. We believe that the WASABI architecture is well suited for more elaborate realizations of coping-related reasoning, as for example implemented by Marsella & Gratch (2006) for their EMA model (see Section 5).

Due to the cognition-independent emotion dynamics simulation the cause of any of the aware emotions arriving in the *Cognitive reappraisal* sub-module needs to be re-established. For secondary emotions the cognitive architecture keeps track of the emotion's cause by memorizing it explicitly. Thus, Max can tell, for example, what he is relieved *about*, or *why* he sees his fears confirmed. The causal reasons for experiencing primary emotions such as anger or happiness, however, cannot be memorized during the *reactive* or *cognitive appraisal* steps, because they might result from an accumulation of equally signed emotional impulses which might have originated from purely reactive appraisal alone. Accordingly, the WASABI architecture allows for misattribution of an emotion's cause to happen. The processes leading to this effect are detailed along the lines of an example interaction in the next section.

4 A case study of causal misattribution in WASABI

The dynamic interplay of the agent's appraisal and his emotion dynamics is best demonstrated along an example interaction between Max and a human opponent in the card game Skip-Bo. We decided to use a playful interaction scenario assuming that humans will more openly show their feelings and, thus, also more easily accept a virtual agent's direct way of expressing its emotions. In addition, a game provides well-defined boundaries to the set of plausible actions and its rules allow for the computational generation of meaningful expectations for the agent.

Insert Figure 6 about here.

The commercial card game Skip-Bo was adapted for our three-dimensional cave-like virtual reality environment such that humans can play it against Max (see **Figure 6**). A set of carefully crafted plans allows Max to follow the rules of the game based on simple heuristics and all human opponents agreed that it is fun to play against him although he is not a particularly strong player.

The WASABI architecture was employed to let Max react emotionally throughout the game and it led to believable interactivity as will be outlined next.

4.1 Technical realization and example of an interaction

Insert Figure 7 about here.

The virtual human MAX is based on a multi-agent system which encapsulates his cognitive abilities inside specialized software agents (see **Figure 7**). These software agents communicate with each other by passing messages.

The *Integration/Categorization module* is implemented as a so-called Emotion-Agent, which acts in concert with a number of other agents. In the Skip-Bo scenario the Emotion-Agent receives emotional impulses from the BDI-Agent, which is continuously being updated with the set of aware emotions. The reasoning processes within the BDI-Agent also derive the actual state of Dominance from the context of the card game, such that MAX feels dominant whenever it is his turn and non-dominant, i.e. submissive, otherwise. Thus, whenever the human opponent fails to follow the rules of the game, MAX takes the turn to correct her and accordingly feels dominant until giving the turn back to the human. Concurrently, the BDI-Agent keeps the Visualization-Agent updated about the actual primary emotions and PAD values.

Insert Figure 8 about here.

Figure 8 illustrates an example of an information flow within the WASABI architecture. In this sequence diagram the three agents BDI-Agent, Emotion-Agent, and Visualization-Agent (‘Vis.-Agent’) are represented as boxes in the top. In the top-left box, labelled BDI-Agent, three plans—generate-expectation (‘gen. exp.’), check expectations (‘check exp.’), and react-to-secondary-emotion (‘react sec.’)—are rendered as three white rectangles to show their activity below. The same rectangles are used to depict the PAD space as well as the emotions *fearful* and *Fears-Confirmed* (‘Fears-Conf.’) which all reside in the Emotion-Agent. The internals of the Visualization-Agent are not detailed here. In this example it only receives messages from the other agents, although in reality it also distributes information about the human player's interaction with the game by sending messages to the BDI-Agent (see **Figure 7**).

We will now explain the sequence of message communication between these agents for which the time-line runs from top to bottom in **Figure 8**. At first, the generate-expectation plan is called, e.g., after MAX ends his turn by playing one last card on one of his stock piles in front of him (see **Figure 6**). This plan, first, results in a negative impulse ('send impulse neg.') which is sent by the *Reactive appraisal* sub-module to the emotion dynamics of the Emotion-Agent thereby indirectly changing MAX's emotional state in PAD space (see **Section 1**). Subsequently, while following the same plan, the primary emotion *fearful* is being triggered ('trigger fearful') by the same *Reactive appraisal* sub-module of the BDI-Agent, because MAX expects the human player to play an important card that would hinder him to fulfil his goal of winning the game.

In the Emotion-Agent, however, the negative emotional impulse pushed the reference point in PAD space already close enough to the (not yet triggered) emotion *fearful* to let MAX experience fear with low intensity. This is possible, because we decided to set *fearful* to a slightly positive base intensity of 0.25; see (Becker-Asano & Wachsmuth, 2009) for details. In **Figure 8** this base intensity is indicated by a small double line along the dashed, vertical lifeline of *fearful*. Accordingly, *slightly fearful* is sent to the Visualization-Agent even before the BDI-Agent triggers the emotion *fearful*. Because the intensity of *fearful* in the Emotion-Agent abruptly changes with the incoming *trigger fearful* message, MAX's emotional state changes from *slightly fearful* to *very fearful*. This sudden change in intensity is reproduced in **Figure 8** by the two gray triangles drawn along each emotion's lifelines. Accordingly, in that moment Max shows a clear expression of fear in his face (see **Figure 6**).

The intensity of *fearful* decreases within the next ten seconds and the reference point changes its location in PAD space due to the implemented emotion dynamics. Thus, *very fearful* automatically changes to *fearful* (see right side of **Figure 8**) in the absence of any further impulse or trigger messages.

Next, the *Cognitive appraisal* module of the BDI-Agent uses the 'check expectations' plan to check, if a human player's action matches any of the previously generated expectations. In this example, the BDI-Agent, first, sends a negative impulse to the Emotion-Agent, because it is assumed here that such a previous expectation exists. The reference point's location in PAD space is thereby changed such that MAX gets *very fearful* again. This sequence of different emotion intensities (*slightly fearful*, *very fearful*, *fearful*, *very fearful*) can happen in case of every primary or secondary emotion, although it is exemplified here only for *fearful*. It results from the dynamic interplay of the *Appraisal module* and the *Integration/Categorization* module.

The 'check expectations' plan, then, triggers the secondary emotion *Fears-Confirmed* ('trigger Fears-Conf.') in the Emotion-Agent, thereby maximizing its base intensity. Together with the negatively valenced mood, *fears-confirmed* reaches the agent's level of awareness and is sent back to the BDI-Agent ('send Fears-Conf.'). In effect, the plan react-to-secondary-emotion is executed within the *Cognitive reappraisal* sub-module to process the incoming message. This results in an 'utter Fears-Conf.' message, which is sent to the Visualization-Agent letting MAX produce an appropriate utterance.

4.2 Misattribution of an emotion's cause

A human opponent would possibly explain Max's behaviour like this:

> After MAX ended his turn with playing a hand card to one of his stock piles, he seemed to realize within one or two seconds that I could now directly play one of my four stock pile cards. I could derive this from his fearful facial expression and the fact that he seemed to inhale sharply producing the characteristic sound of someone being afraid. When I then actually played that stock card, MAX admitted that he had been afraid of that before.

In the *Appraisal module* of Max's cognitive architecture, however, the event that caused the fear is disconnected from the finally elicited emotion itself. Thus, if Max were asked why he shows fear, he would have to recapitulate which events of the recent past could have caused his fear. In principle, a number of events could have influenced Max's emotion dynamics negatively rather directly through *Reactive appraisal* alone leaving no trace in form of cognitive representations. For example, if we additionally simulated an artificial hunger level for Max, his slowly getting hungry could result in small negative impulses, which are send repeatedly to the Emotion-Agent and could slowly worsen Max's mood. Assuming this process to be realized in parallel to the BDI-based reasoning module, Max would be unable to consider his being hungry as one factor causing his experience of fear (or any other negative emotion).

Even without modelling additional influences outside of Max's cognitive awareness misattributions could happen in situations in which a high number of events quickly succeed each other. For example, if Max cognitively perceived a new person entering the scene directly after realizing that the human opponent is likely to play a fear-inducing card (see example above), he might later misattribute this new person's appearance to be causing his fear. In fact, this can happen even if the new person's appearance itself had no emotional impact at all. Accordingly, Max being asked about why he shows fear in the above example could then be prone to the following misattribution:

I think I fear that person next to you who just entered the room.

Before we further discuss the pros and cons of our architecture, we will contrast it next with related work in the field of affective computing.

5 Related work

El-Nasr, Yen, & Ioerger (2000) present FLAME as a formalization of the dynamics of 14 emotions based on fuzzy logic rules. It includes a mood value, which is continuously calculated as the average of all emotion intensities to provide a solution to the problem of conflicting emotions being activated at the same time. Our conception of emotion dynamics in the WASABI architecture is quite similar to their realization of mutual influence of emotion and mood.

Marsella and Gratch (2006) focus with their EMA model of emotions on the dynamics of emotional appraisal. They also argue for a mood value as an addend in the calculation of otherwise equally activated emotional states following the general idea of mood-congruent emotions. Their framework for modelling emotions is certainly much better suited to explain the cognitive underpinnings of emotions than the WASABI architecture can possibly do. Furthermore, it has successfully been evaluated as to model human emotion dynamics quite accurately (see **Chapter MGP**, this volume). The strength of the WASABI architecture, however, seems to be the relative simplicity with which convincing emotion dynamics (at least of primary emotions) can be achieved. With EMA's rational reasoning approach it seems also much more difficult to explain misattribution of an emotion’s cause, because of the direct linking of a domain object and its emotional effect.

Central to the architecture proposed by Marinier and Laird (2006) is the idea of ‘Appraisal Frames’, which are based on the EMA model and eleven of Scherer’s sixteen appraisal dimensions (see **Chapter S** of this volume) and are modelled for integration in the Soar cognitive architecture (Laird, Newell, & Rosenbloom, 1987). They distinguish an ‘Active Appraisal Frame’, which is the result of a momentary appraisal of a given event, from a ‘Perceived Appraisal Frame’, which results from the combination of the actual mood and emotion frames. Thereby, they take Damasio’s distinction between emotion and feeling into account—similarly to the conception underlying the WASABI architecture. It has to be noted, however, that Damasio defines a feeling as ‘the perception of a certain state of the body along with the perception of a certain mode of thinking and of thoughts with certain themes’ (Damasio, 2003, p. 86). This definition seems to be even more difficult to operationalise than Scherer’s assumption that ‘feelings integrate the central representations of appraisal-driven response in emotion’ (**Chapter S, page YY**, this volume). Although it remains a challenging question, if we can ever reasonably ascribe feelings to autonomous agents, we believe that following a component approach to modelling affect is most promising and to that respect Damasio and Scherer seem to agree.

Aiming at the development of believable conversational agents Pelachaud & Bilvi (2003) are continuously improving their ‘Greta’ agent, which is capable of producing bodily as well as facial gestures that are consistent with the situational context. This consistency is guaranteed by BDI-based modelling of Greta’s ‘mind’ (Rosis, Pelachaud, Poggi, Carofiglio, & Carolis, 2003) resulting in an architecture, which allows for the inclusion of an emotion model. The latter builds upon a ‘Dynamic Belief Network’ to account for the inherent dynamics of emotional processes, which is also central to our work as outlined above. Recently, Ochs, Devooght, Sadek, & Pelachaud (2006) presented another BDI-based approach to implement OCC-based appraisal for Greta.

The layered model of affect ALMA uses PAD space to derive an agent’s mood from emotions that are themselves the result of OCC-based appraisal (Gebhard, 2005). ALMA is rooted in a purely cognitive approach to modelling affect, which was only later extended by a representation of all 22 OCC emotions (plus ‘liking’ and ‘disliking’) in PAD space. Accordingly, in evaluating ALMA, Gebhard & Kipp (2006) heavily rely on third person rational judgement of the believability of emotion and mood labels, which ALMA generates for two interacting conversational agents. For WASABI, in contrast, the emotional effect of Max’s affective behaviour in direct playful face-to-face interaction has been evaluated to be beneficial (Becker, Prendinger, Ishizuka, & Wachsmuth, 2005). In **Chapter SCHR** of this volume (Schröder, 2004) introduces his approach to emotional speech synthesis, which is based on PAD space as well.

6 Conclusions and discussion

In contrast to most existing computational affect simulation models the WASABI architecture focuses on capturing the temporal dynamics of emotions and mood and makes only very few commitments regarding how cognitive appraisal is to be realized. Furthermore, the WASABI architecture breaks the link between an emotion eliciting domain object and the resulting emotion itself and how this can be exploited to realize misattribution of an emotion’s cause.

We admit, however, that the conceptual decisions taken in designing the WASABI architecture are not without questionable consequences. A major challenge is the question of how WASABI can account for the occurrence of mixed emotions. For example, imagine yourself cueing up for taking a ride in a roller coaster, which is likely to produce an adrenalin rush. In such a moment happiness of expecting a pleasurable ride appears to be mixed with fearing possible negative consequences of the same event in case of an accident. Although both emotions, happy and fearful, might be triggered by the *Reactive appraisal* sub-module (see **Figure 1**), the distance between these two emotions in PAD space makes it impossible for Max to be aware of them simultaneously. In fact, the assured mood-congruency of emotions prevents in this case the simultaneous elicitation of positively valenced happiness and negatively valenced fear. It can be argued, however, that in humans these two emotions are also not experienced simultaneously, but in quick succession one after the other depending on a human’s focus of cognitive attention.

Another challenge to be addressed in future research is how we can model and test the emotion-related effects an agent’s personality. So far, we heuristically determined reasonable values for the parameters of the emotion dynamics simulation.

We believe, however, that by changing these parameters we can systematically change an observer's judgement of the agent's personality. Although the 'Big Five' personality schema with its proposed relation to PAD space (Mehrabian A., 1996) is commonly used to realize an agent's personality (Gebhard, 2005), we believe that our conception of an emotion dynamics is already powerful enough to allow for the creation of different personalities within WASABI-driven agents.

Finally, how to realize cognitive reappraisal is still an open topic for the WASABI architecture. The BDI-based architectures discussed in **Section 5** are much better suited to explain the *why* of an emotion, because their emotion elicitation is explicitly based on rational reasoning processes. Thus, we believe that combining WASABI's core ideas (see **Section 1**) with these more cognitively motivated affect simulation architectures would yield interesting results and in doing so we might also achieve a more complete picture of a human's emotional life.

In summary, we hope that the WASABI architecture contributes to the diverse theories and ideas within the emotion research community and also provides a valuable technical contribution to the question of how to endow embodied agents with affective competency.

Bibliography

Bartneck, C. (2002). Integrating the OCC Model of Emotions in Embodied Characters. *Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges.* Melbourne.

Becker, C., Kopp, S., & Wachsmuth, I. (2004). Simulating the emotion dynamics of a multimodal conversational agent. *Intl. Workshop on Affective Dialogue Systems* (pp. 154-165). Berlin / Heidelberg: Springer.

Becker, C., Kopp, S., & Wachsmuth, I. (2007). Why emotions should be integrated into conversational agents. In T. Nishida (Ed.), *Conversational Informatics: an Engineering Approach* (pp. 49-68). London: John Wiley & Sons Ltd.

Becker, C., Prendinger, H., Ishizuka, M., & Wachsmuth, I. (2005). Evaluating Affective Feedback of the 3D Agent Max in a Competitive Cards Game. *Proc. of Intl. Conf. on Affective Computing and Intelligent Interaction* (pp. 466-473). Beijing, China: Springer.

Becker-Asano, C. (2008). *WASABI: Affect Simulation for Agents with Believable Interactivity.* Amsterdam: IOS Press.

Becker-Asano, C., & Wachsmuth, I. (2009). Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems*.

Damasio, A. (1994). *Descartes' Error, Emotion Reason and the Human Brain.* Grosset/Putnam.

Damasio, A. (2003). *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain.* Harcourt.

Ekman, P. (1999). Basic Emotions. In *Handbook of Cognition and Emotion* (pp. 45-60). John Wiley & Sons.

El-Nasr, M. S., Yen, J., & Ioerger, T. R. (2000). FLAME - Fuzzy Logic Adaptive Model of Emotions. *Autonomous Agents and Multi-Agent Systems, 3,* 219-257.

Gebhard, P. (2005). ALMA - A Layered Model of Affect. *Proc. of Intl. Conf. on Autonomous Agents & Multi Agent Systems* (pp. 29-36). ACM.

Gebhard, P., & Kipp, K. H. (2006). Are Computer-Generated Emotions and Moods Plausible to Humans? *Proc. of Intl. Conf. on Intelligent Virtual Agents* (pp. 343-356). Springer.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: an architecture for general intelligence. *Artificial Intelligence, 33*, 1-64.

Leßmann, N., Kopp, S., & Wachsmuth, I. (2006). Situated interaction with a virtual human - perception, action, and cognition. In *Situated Communication* (pp. 287-323). Mouton de Gruyter.

Marinier, R., & Laird, J. (2007). Computational Modeling of Mood and Feeling from Emotion. *Proc. of CogSci,* (pp. 461-466).

Marsella, S., & Gratch, J. (2006). EMA: A computational model of appraisal dynamics. *Proc. of Agent Construction and Emotions.*

Mehrabian, A. (1996). Analysis of the big-five personality factors in terms of the pad temperament model. *Australian Journal of Psychology, 48,* 86-92.

Mehrabian, A. (1995). Framework for a Comprehensive Description and Measurement of Emotional States. *Genetic, Social, and General Psychology Monographs* (121), 339-361.

Ochs, M., Devooght, K., Sadek, D., & Pelachaud, C. (2006). A Computational Model of Capability-Based Emotion Elicitation for Rational Agent. *Intl. workshop on emotion and computing.* Bremen.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions.* Cambridge University Press.

Pelachaud, C., & Bilvi, M. (2003). Computational model of believable conversational agents. In *Communications in Multiagent Systems.* Springer-Verlag.

Rao, A., & Georgeff, M. (1991). Modeling Rational Agents within a BDI-architecture. *Proc. of the Intl. Conference on Principles of Knowledge Representation and Planning* (pp. 473-484). San Mateo, CA, USA: Morgan Kaufmann publishers Inc.

Rosis, F. d., Pelachaud, C., Poggi, I., Carofiglio, V., & Carolis, B. d. (2003). From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *Intl. Journal of Human-Computer Studies, 59,* 81-118.

Scherer, K. R. (2001). Appraisal Considered as a Process of Multilevel Sequential Checking. In K. Scherer, A. Schorr, & T.Johnstone, *Appraisal Processes in Emotion* (pp. 92-120). New York and Oxford: Oxford University Press.

Scherer, K. R. (1984). On the Nature and Function of Emotion: A Component Process Approach. In *Approaches to emotion* (pp. 293-317). Lawrence Erlbaum.

Schröder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. *Intl. Workshop on Affective Dialogue Systems* (pp. 209-220). Berlin / Heidelberg: Springer.

Wundt, W. (1922/1863). *Vorlesung über die Menschen- und Tierseele.* Voss Verlag.

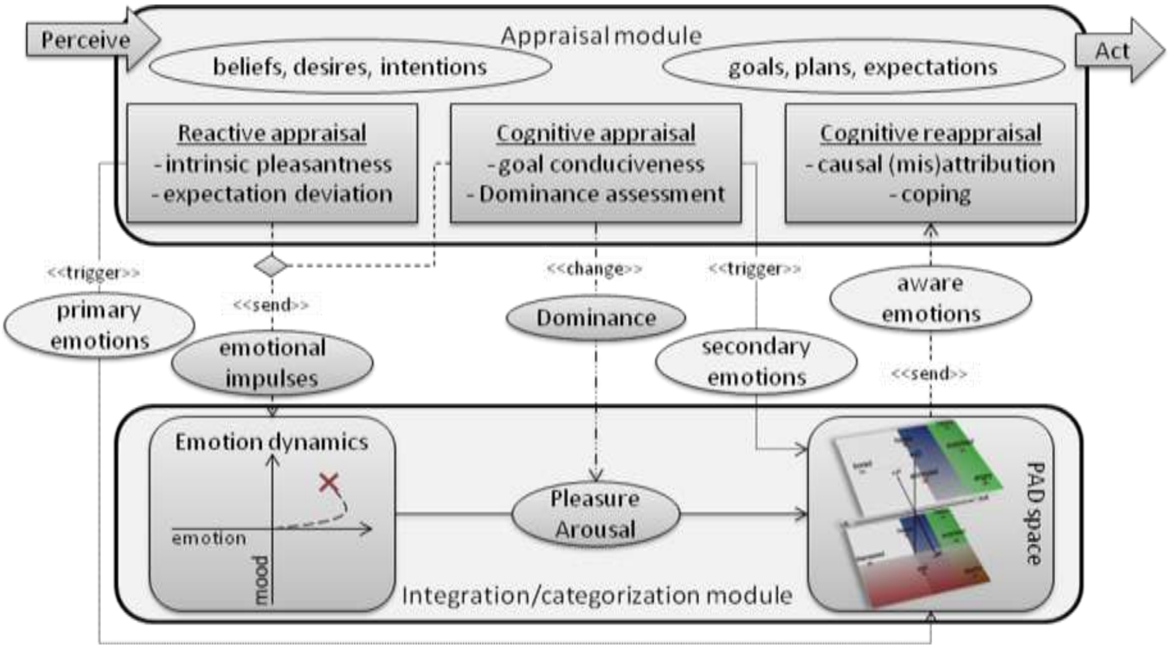

Figure 1. A general overview of the WASABI architecture with its 'Appraisal module' on top and the internal 'Integration/Categorization module' at the bottom

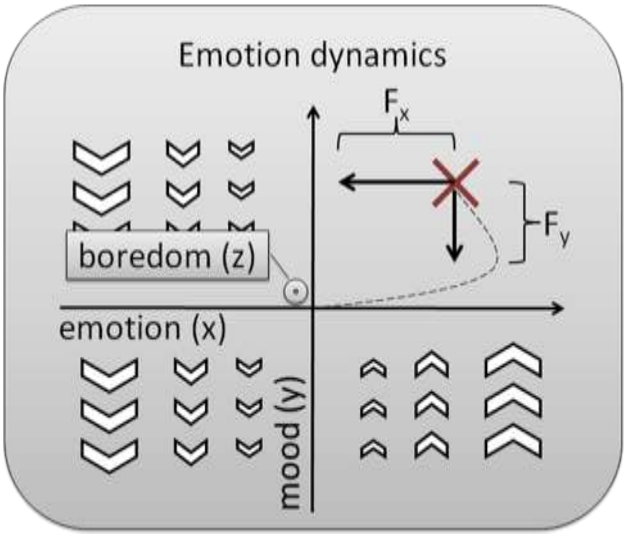

Figure 2. Details of the emotion dynamics part inside the integration/categorization module of the WASABI architecture

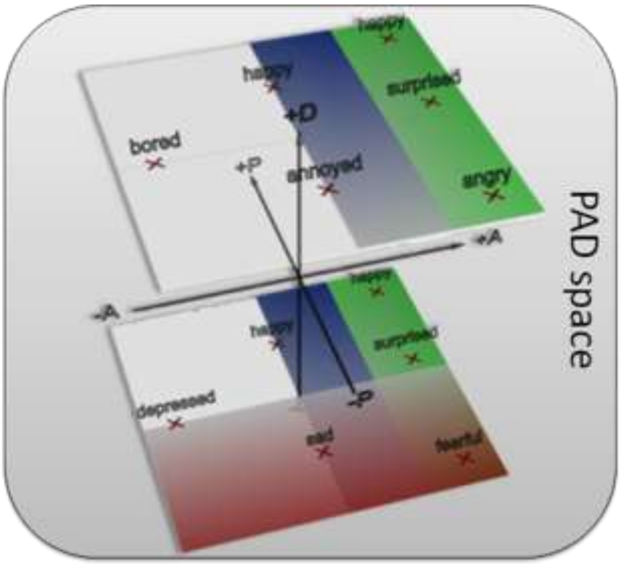

Figure 3. The (P)leasure-(A)rousal-(D)ominance space with nine primary emotions (indicated by the labelled red crosses) and three secondary emotions – ‘hope’ (green), ‘relief’ (blue), and ‘fears-confirmed’ (red) – assigned to the high and low dominance planes.

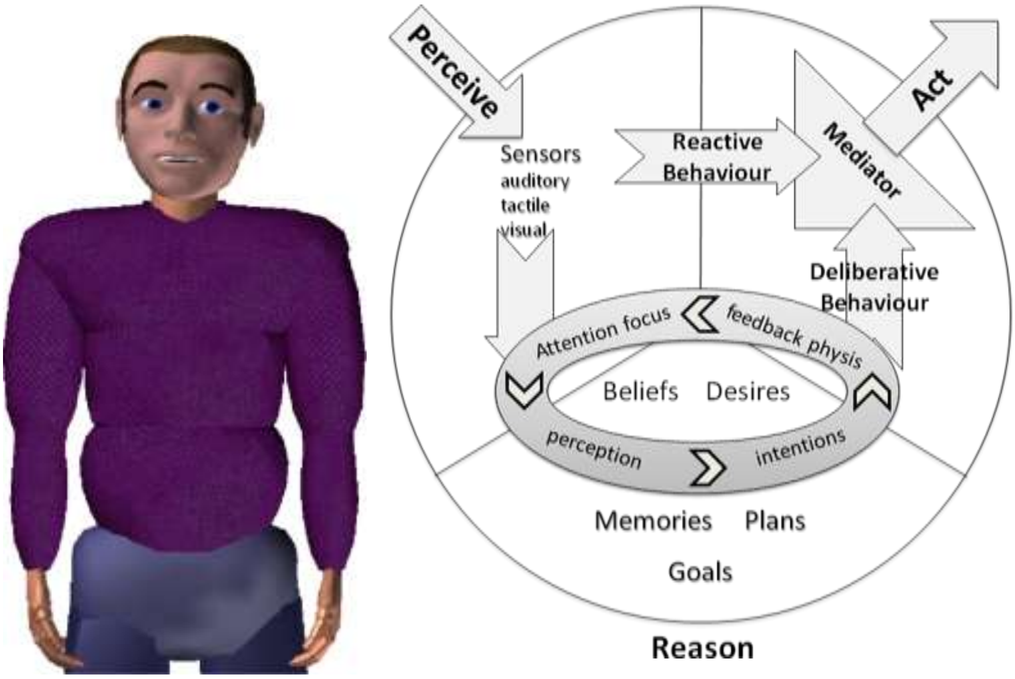

Figure 4. The virtual human ‘Max’, left, and an outline of its architectural framework, right; reproduced from (Leßmann, Kopp, & Wachsmuth, 2006)

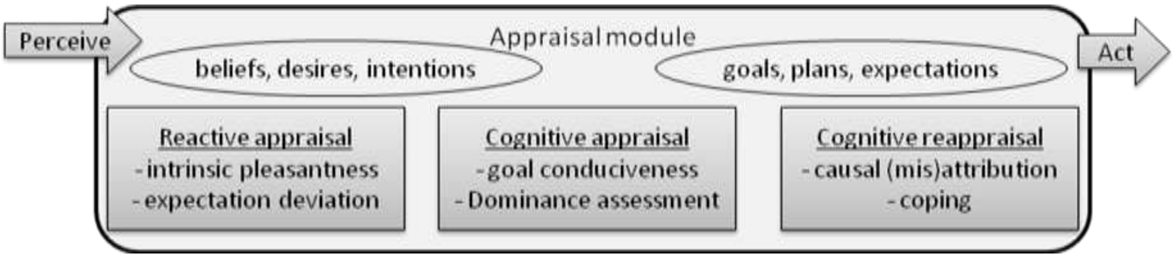

Figure 5. The ‘Appraisal module’ of the WASABI architecture with its subcomponents ‘Reactive appraisal’, ‘Cognitive appraisal’, and ‘Cognitive reappraisal’

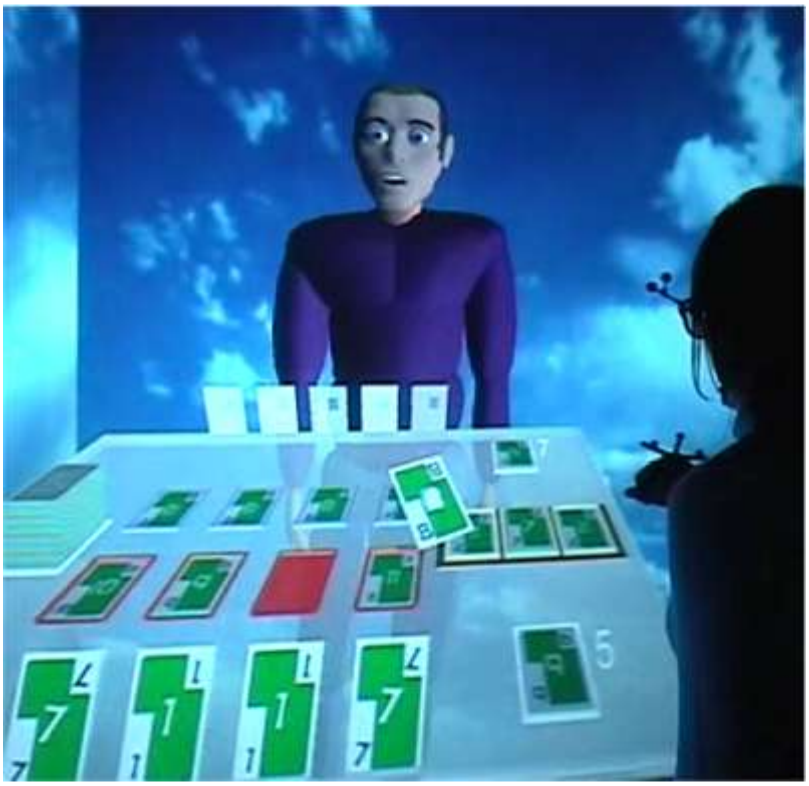

Figure 6. Screen shot of Max playing Skip-Bo against a human opponent in the virtual reality installation of the Faculty of Technology at Bielefeld University, Germany. The red areas in front of the human’s hand cards are her stock piles, which are visible to Max and enable him to generate expectations about which cards she might play next. Accordingly, in the moment depicted here Max expresses his fear of her playing the ‘8’ on top of the ‘7’ on one of the three shared target piles to the right of the virtual table.

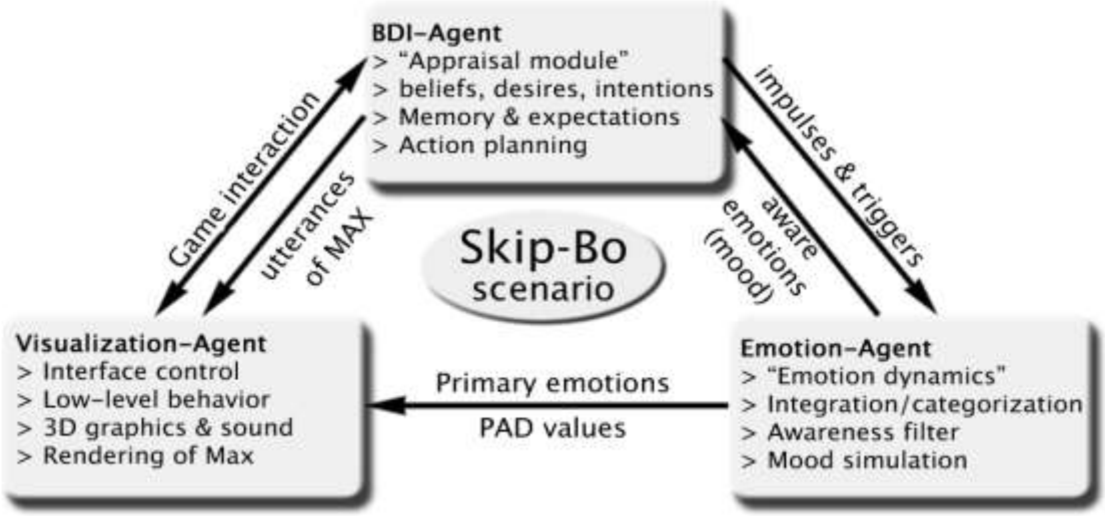

Figure 7. The three most important software agents in the Skip-Bo scenario are presented together with their interconnection realized by means of message passing. The *Appraisal module* is part of the BDI-Agent, the *Integration/categorization module* resides inside the Emotion-Agent, and the Visualization-Agent renders the 3D graphics including the game and the Max agent. A user's interaction with the game is also handled by the Visualization-Agent and then forwarded to the BDI-Agent.

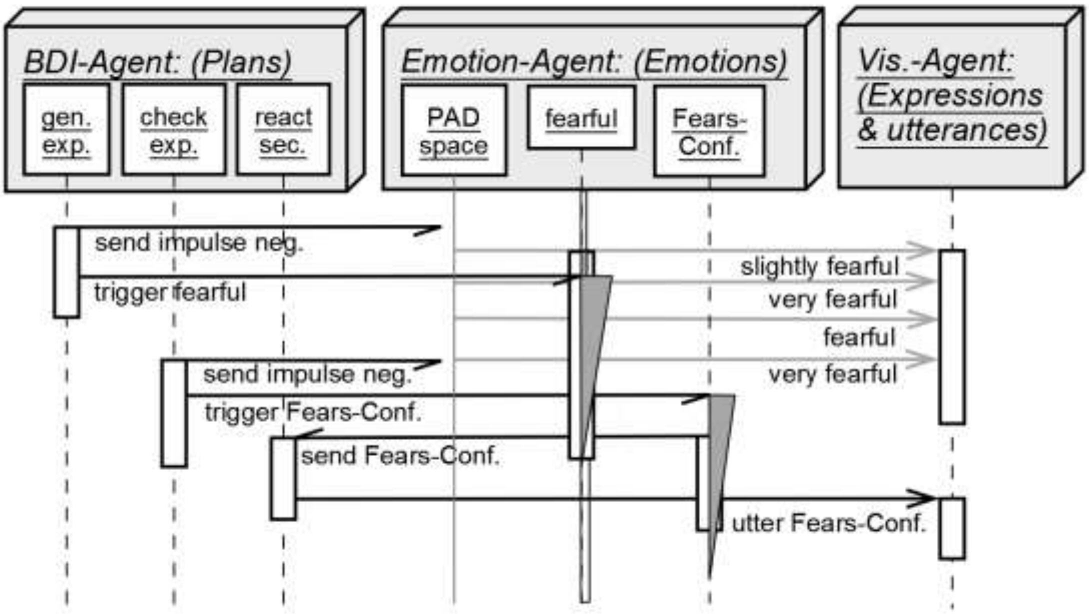

Figure 8. Sequence diagram of an information flow between the software agents with the time-line from top to bottom