Reasoning about Emotional Agents

John-Jules Ch. Meyer[1]

Abstract. In this paper we discuss the role of emotions in artificial agent design, and the use of logic in reasoning about the emotional or affective states an agent can reside in. We do so by extending the KARO framework for reasoning about rational agents appropriately. In particular we formalize in this framework how emotions are related to the action monitoring capabilities of an agent.

1 INTRODUCTION

In this paper we are concerned with reasoning about agents with emotions. To be more precise: we aim at a logical account of emotional agents. The very topic may already raise some eyebrows. Reasoning / rationality and emotions seem opposites, and reasoning about emotions or a logic of emotional agents seems a contradiction in terms.

However, emotions and rationality are known to be more interconnected than one may suspect. Damasio [5] relates the story of a patient called 'Elliott' having a certain kind of brain damage preventing him have (secondary) emotions (cf. [21]). Although this would seem to make the patient 'superrational' in the sense of performing extremely well at rational tasks like decision-making (not being disturbed by emotions), this turns out to be completely the opposite: by not being able to employ emotions to stop endless deliberations, he performs really poorly at these tasks. So there seems to be psychological evidence that having emotions may help one to do reasoning and tasks for which rationality seems to be the only factor.

Moreover, the ground-breaking work by e.g. Sloman [21, 22] shows that one may think of *designing* agent-based systems where these agents show some kind of emotions, and, even more importantly, display behaviour dependent on their emotional state. It is exactly in this sense that we aim at looking at emotional agents: artificial systems that are designed in such a manner that emotions play a role (cf. [8]). Interestingly also in psychology emotions are viewed as a structuring mechanism. Emotions are held to help human beings to choose from a myriad of possible actions in response to what happens in our complex world (cf. [17]).

So we advocate the use of emotional states to design an artificial intelligent agent. One has to bear in mind, that this has in itself nothing to do with the philosophical and very difficult question whether these agents really possess true emotions in the sense that we humans do! This is similar to the question whether artificial agents possess true intelligence or consciousness like humans do. One can perfectly well think about the design of intelligent agents without addressing this issue.

In this paper we argue that

1. emotions make sense in describing the behaviour of certain intelligent agents, and may help structuring the design of the agent (by means of an architecture that caters for emotional aspects) and
2. consequently it is useful to reason about emotions of an agent, or rather about the emotional states an agent may be in, together with its effects on the agent's actions, as an important aspect of the agent's behaviour.

So our logic will be more concerned with the behaviour of such a system than with emotions *per se*. This is a perfectly sensible way to go in line with software and system engineering practice. To specify systems in a rigourous way one may employ certain logical methods by which one can unambigously state how the system should behave. In classical *imperative programming* this involves the specification of input-output relations. In *reactive system* specification one specifies how the state of the system evolves over time in possibly never ending computations arising from interactions with the environment of the system (cf. e.g. [13]). In agent-based systems where the agents are perceived as *rational* or *intelligent* ones, possessing some sort of attitudes pertaining to information and motivation such the well-known BDI (belief–desire–intention) agents, we can describe their behaviour in terms of the evolution of the mental states of the agent over time (e.g. BDI logic [19, 24], Cohen & Levesque' approach [4], and KARO logic [11]). Indeed, Shoham in his seminal paper ([20]) on agent-oriented programming says that agent programs are 'mental state transformers'. We now also want to perceive emotional agents as systems that evolve over time and can be described by some logic as the one mentioned above for rational agents.

So what we aim at is describing behaviours of emotional agents in terms of the way their (emotional) states evolve over time. This means that we are interested in at least two things: how do actions of agents (by definition agents act!) change their emotional states and how do emotional states determine what action is taken and what effect is obtained from this in the given state.

The way we will go about is as follows. From the psychological literature we get evidence that the way emotions influence behaviour is on a rather high level. Emotions like happiness and fear generally do not result directly in taking concrete actions by agents, but rather in an attitude towards handling their goals and intentions. Emotions moderate the execution and maintenance of the agent's agenda, so to speak. It will turn out that we can model these high-level attitudes adequately in the logical framework that we have devised for rational agents. In essence our approach is thus: to reason about the dynamics of (emotional) states we use the framework of *dynamic logic* ([9] and (an extension of) the KARO framework ([12, 11, 15]) in particular.

2 PSYCHOLOGICAL PRELIMINARIES

As noted before, in psychological theory emotions are associated with higher-level mental attitudes [16, 17] and have many facets ([18]). In this paper, as we are interested in constructing artificial agents using emotions as a 'designing tool', we concentrate on their

[1] ICS, Utrecht University, Utrecht, The Netherlands, email: jj@cs.uu.nl

relation to the agent's behaviour, and in particular the agent's actions. This relation has been studied in cognitive science as well ([1, 7]).

In general terms one may distinguish so called *"well-being emotions"* and *"prospect-based emotions"* with respect to actions and events ([18]). Both have positive and negative variants. In this paper we describe some of the basic emotions as discussed in [17]: in particular those that can occasionally be so-called *free-floating*, i.e. not having a particular object towards which the emotion is directed. These emotions are *happiness*, *sadness*, *anger* and *fear*.[2]

Happiness Happiness is taken to be the emotion or mood of achieving (sub)goals, of being engaged in what one is doing. It is triggered by the fact that (sub)goals are being achieved. The attitude(s) associated with happiness is/are: continue with plan, modifying if necessary; cooperate; show affection.

Sadness Sadness is the emotion of losing a goal or social role, and knowing it cannot be reinstated. Sadness is triggered by the failure of a major plan or the loss of an active goal. Associated attitudes: do nothing; search for a new plan; ask help.

Anger Anger is the emotion of asserting oneself in dominance. Triggered by an active plan being frustrated. Associated attitudes: try harder; aggress.

Fear Fear is the emotion of anticipated danger. Fear is triggered by a self-preservation goal being threatened or a goal conflict. Associated attitudes: stop current plan; attend vigilantly to the environment; freeze and/or escape.

Observe that some attitudes contain both individual and social aspects. In this paper we restrict ourselves to the individual aspects.

3 KARO LOGIC

In this section we briefly review the KARO formalism, in which *action*, together with knowledge / belief, is the primary concept, on which other agent notions are built. The KARO framework has been developed in a number of papers (e.g. [12, 11, 15]). Here we employ a version with a belief rather than knowledge operator.

The KARO formalism is an amalgam of dynamic logic and epistemic / doxastic logic [14], augmented with several additional (modal) operators in order to deal with the motivational aspects of agents. So, besides operators for belief ($\mathbf{B}$) and action ($[\alpha]$, "after performance of α it holds that"), there are additional operators for ability ($\mathbf{A}$) and desires ($\mathbf{D}$). We assume a set $\mathcal{A}$ of atomic actions and a set $\mathcal{P}$ of atomic propositions.

Definition 1 *The language $\mathcal{L}_{KARO}$ of KARO-formulas is given by the BNF grammar:*

$$\begin{array}{rcl} \varphi & ::= & p(\in \mathcal{P}) \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \ \ldots \\ & & \mathbf{B}\varphi \mid \mathbf{D}\varphi \mid [\alpha]\varphi \mid \mathbf{A}\alpha \end{array}$$

$$\alpha \quad ::= \quad a(\in \mathcal{A}) \mid \varphi? \mid \alpha_1;\alpha_2 \mid \alpha_1 + \alpha_2 \mid \alpha^*$$

We refer to the α-type formulas as actions. We use the abbreviations $\mathtt{tt} \equiv p \vee \neg p$ (for some fixed $p \in \mathcal{P}$) and $\mathtt{ff} \equiv \neg\mathtt{tt}$. The conditional action is introduced by the usual abbreviation: $\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi} \equiv (\varphi?;\alpha_1) + (\neg\varphi?;\alpha_2)$.

Thus formulas are built by means of the familiar propositional connectives and the modal operators for knowledge, belief, desire, action and ability. Actions are the familiar ones from imperative programming: atomic ones, tests, sequential composition, (nondeterministic) choice and repetition.

Definition 2 *1. We consider Kripke structures of the following form: $\mathcal{M} = \langle W, \vartheta, R_B, R_D\rangle$, where*

- *W is a non-empty set of states (or worlds)*
- *ϑ is a truth assignment function per state*
- *R_B, R_D are accessibility relations for interpreting the modal operators $\mathbf{B}, \mathbf{D}$. The relation R_B is assumed to be euclidean, transitive and serial. Nothing special is assumed for the relation R_D.*

2. The semantics of actions is given by means of structures of type $\langle \Sigma, \{R_a \mid a \in \mathcal{A}\}, \mathcal{C}, Ag\rangle$, where

- *Σ is the set of possible model/state pairs (i.e. models of the above form, together with a state appearing in that model)*
- *R_a ($a \in \mathcal{A}$) are relations on Σ encoding the behaviour of atomic actions*
- *$\mathcal{C}$ is a function that gives the set of actions that the agent is able to do per model/state pair*
- *Ag is a function that yields the set of actions that the agent is committed to (the agent's 'agenda') per model/state pair.*

Definition 3 *In order to determine whether a formula $\varphi \in \mathcal{L}$ is true in a model/state pair (M, w) (if so, we write $(M, w) \models \varphi$), we stipulate (omitting the purely propositional cases):*

- *$\mathcal{M}, w \models \mathbf{B}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_B(w, w')$*
- *$\mathcal{M}, w \models \mathbf{D}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_D(w, w')$*
- *$\mathcal{M}, w \models [\alpha]\varphi$ iff $\mathcal{M}', w' \models \varphi$ for all M', w' with $R_\alpha((\mathcal{M}, w), (\mathcal{M}', w'))$*
- *$\mathcal{M}, w \models \mathbf{A}\alpha$ iff $\alpha \in \mathcal{C}(\mathcal{M}, w)$*
- *$\mathcal{M}, w \models \mathbf{Com}(\alpha)$ iff $\alpha \in Ag(\mathcal{M}, w)$*[3]

Here R_α is defined as usual in dynamic logic by induction from the basic case R_a (cf. e.g. [9, 11]). So, e.g. $R_{\alpha_1+\alpha_2} = R_{\alpha_1} \cup R_{\alpha_2}$, $R_{\alpha^*} = R_\alpha^*$, the reflective transitive closure of R_α, and $R_{\alpha_1;\alpha_2}$ is the relational product of R_{α_1} and R_{α_2}. Likewise the function $\mathcal{C}$ is lifted to complex actions ([11]). We call an action α *deterministic* if $card\{w' \mid R_\alpha(w, w')\} \leq 1$ for any $w \in W$. and *strongly deterministic* if $card\{w' \mid R_\alpha(w, w')\} = 1$.

So we use a standard modal semantics for knowledge, belief, desire and action. The agent is able to do the action if it is indicated so by the function C, and an agent is committed to an action α if it is recorded so in the agent's agenda. Furthermore, we will make use of the following syntactic abbreviations serving as auxiliary operators:

Definition 4

- *(dual) $\langle\alpha\rangle\varphi = \neg[\alpha]\neg\varphi$: the agent has the opportunity to perform α resulting in a state where φ holds.*
- *(opportunity) $\mathbf{O}\alpha = \langle\alpha\rangle\mathtt{tt}$: an agent has the opportunity to do an action iff there is a successor state w.r.t. the R_α-relation;*
- *(practical possibility) $\mathbf{P}(\alpha, \varphi) = \mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$: an agent has the practical possibility to do an action with result φ iff it is both able and has the opportunity to do that action and the result of actually doing that action leads to a state where φ holds;*

[2] Some cognitive scientists prefer to think of these as labels of 'families' of emotions rather than specific emotions [3]. Here I follow [17] and use them as conveniently concise labels of the emotions that we will treat formally in the sequel.

[3] The agenda is assumed to be closed under certain conditions such as taking 'prefixes' of actions. Details can be found in [15].

- *(can)* $\mathbf{Can}(\alpha, \varphi) = \mathbf{BP}(\alpha, \varphi)$*: an agent can do an action with a certain result iff it believes it has the practical possibilty to do so;*[4]
- *(realisability)* $\Diamond\varphi = \exists a_1, \ldots, a_n \mathbf{P}(a_1; \ldots; a_n, \varphi)$[5]*: a state property* φ *is realisable iff there is a finite sequence of atomic actions of which the agent has the practical possibility to perform it with the result* φ*;*
- *(goal)* $\mathbf{G}\varphi = \neg\varphi \wedge \mathbf{D}\varphi \wedge \Diamond\varphi$*: a goal is a formula that is not (yet) satisfied, but desired and realisable.*
- *(possible intend)* $\mathbf{I}(\alpha, \varphi) = \mathbf{Can}(\alpha, \varphi) \wedge \mathbf{BG}\varphi$*: an agent (possibly) intends an action with a certain result iff the agent can do the action with that result and it moreover believes that this result is one of its goals.*

In order to manipulate both knowledge / belief and motivational matters special actions `revise`, `commit` and `uncommit` are added to the language, the semantics of which is given here very abstractly by means of functions. (cf. [11, 15]):

Definition 5

1. $R_{\texttt{revise}\varphi}(\mathcal{M}, w) = \mathit{update\_belief}(\varphi, (\mathcal{M}, w))$.
2. $R_{\texttt{commit}\alpha}(\mathcal{M}, w) = \mathit{update\_agenda}^{+}(\alpha, (\mathcal{M}, w))$*, if* $\mathcal{M}, w \models \mathbf{I}(\alpha, \varphi)$ *for some* φ*, otherwise* $R_{\texttt{commit}\alpha}(\mathcal{M}, w) = \emptyset$ *(failure).*
3. $R_{\texttt{uncommit}\alpha}(\mathcal{M}, w) = \mathit{update\_agenda}^{-}(\alpha, (\mathcal{M}, w))$*, if* $\mathcal{M}, w \models \mathbf{Com}(\alpha)$*, otherwise* $R_{\texttt{uncommit}\alpha}(\mathcal{M}, w) = \emptyset$ *(failure);*
4. $\texttt{uncommit}\alpha \in \mathcal{C}(\mathcal{M}, w)$ *iff* $\mathcal{M}, w \models \neg\mathbf{I}(\alpha, \varphi)$ *for all formulas* φ*, that is, an agent is able to uncommit to an action if it is not intended to do it (any longer) for any purpose.*

Here $\mathit{update\_belief}$, $\mathit{update\_agenda}^{+}$ and $\mathit{update\_agenda}^{-}$ are functions that update the agent's belief and agenda, respectively. The $\mathit{update\_belief}(\varphi, (\mathcal{M}, w))$ function changes the model $\mathcal{M}$ in such a way that the agent's belief is updated with the formula φ, while $\mathit{update\_agenda}^{+}(\alpha, (\mathcal{M}, w))$ changes the model $\mathcal{M}$ such that α is added to the agenda, and likewise for the $\mathit{update\_agenda}^{-}$ function, but now with respect to removing an action from the agenda. The formal definitions can be found in [12, 15].

4 THE DYNAMICS OF EMOTION

We are now ready to deal with the logic of emotional agents, where we especially focus on the dynamics of emotions ('emotions in flux') and the influence of emotions on agenda maintenance, in particular. Rather than trying to capture the informal psychological descriptions exactly (or as exact as possible), we primarily look here at a description that makes sense for artificial agents.

Emotions are high-level attitudes in the sense that they determine how an agent deals with its goals and plans to reach them. An emotional state thus represents a certain attitude towards goal keeping and execution. In the KARO framework we represent emotions with special predicates, ('fluents'). In general, we must specify how the truth of these 'emotional fluents' arises. But we must also represent what the effect of emotions is on the agent's goal/plan keeping strategy. To this end in the sequel we will have (mostly) two axioms per emotional fluent, describing the conditions under which its truth comes about and the effects of the emotion on the agent's behaviour in the above sense. Furthermore, to be able to describe the latter effects, we assume a 'classical' deliberation cycle as in e.g. [23]. In the KARO framework this looks like a program $(\mathit{deliberate}; \mathit{execute})^*$, where the actions $\mathit{deliberate}$ and $\mathit{execute}$ denote actions that select goals and plans to be put on the agenda, and the execution of (part of) the plan on the agenda, respectively. For particular agent systems these actions are such that they adopt a particular strategy when choosing actions and plans ([6]). Here, we will keep this as general as possible, abstracting from particular strategies, but in the sequel we will focus on parts of those strategies that involve emotions. Our axioms to follow are to be seen as constraints on the general deliberation and execution strategies that agents may use.

To keep things relatively simple, in this section we assume that plans have a simple form: just a sequence of deterministic atomic actions (and thus not containing choice and repetition constructs). This enables us to speak about initial parts and remainders of plans in a succinct and comprehensible way.

First we fix some notation. On sequences (of atomic actions) we denote the prefix (or initial part) relation by $\preceq$: for plans α and π it holds that $\alpha \preceq \pi$ if $\pi = \alpha; \pi'$ for some (possibly empty) sequence of actions π'. We use ϵ to denote the empty sequence (of atomic actions). When $\pi = \alpha; \pi'$, we denote π' by $\pi\backslash\alpha$, the remainder of π if its initial part α has already been executed.

4.1 Happiness

An agent that is happy observes that its subgoals (towards certain goals) are being achieved, and is 'happy' with it.[6]. We first describe the situation in which happiness comes about: we want to express that an agent that is striving for a particular goal by working on a subgoal by means of a (sub)plan observes that everything is going according to plan (as it expects). More precisely, we first describe that an agent that has the intention to do π for achieving goal φ, and is committed to it, and that believes that by performing the initial part α the subgoal ψ should be achieved, is happy (with respect to the remainder $\pi\backslash\alpha$ of the plan—to which it is still committed, the goal φ and subgoal ψ) if after the performance of α it believes that indeed the subgoal ψ has been achieved. Formally, we may put this as:

$$\mathbf{I}(\pi, \varphi) \wedge \mathbf{Com}(\pi) \wedge \alpha \preceq \pi \wedge \mathbf{B}([\alpha]\psi) \rightarrow$$
$$[\alpha]((\mathbf{B}\psi \wedge \mathbf{Com}(\pi\backslash\alpha)) \rightarrow \mathit{happy}(\pi\backslash\alpha, \varphi, \psi))$$

Note that in particular it holds under the reasonable condition that the goal φ itself is deemed important by the agent (since $\pi \preceq \pi$, $\mathbf{Com}(\epsilon)$ is true, and $\mathbf{I}(\pi, \varphi)$ implies $\mathbf{B}[\pi]\varphi$) that

$$\mathbf{I}(\pi, \varphi) \wedge \mathbf{Com}(\pi) \rightarrow [\pi](\mathbf{B}\varphi \rightarrow \mathit{happy}(\epsilon, \varphi, \varphi))$$

which expresses that the agent that believes that its goal is realised after having executed/performed its plan, is happy, as to be expected. Now we define: $\mathit{happy}(\pi, \varphi) \Leftrightarrow \mathit{happy}(\pi, \varphi, \psi)$ for all formulas ('subgoals') ψ that are considered important/crucial by the agent.[7] As said before happiness causes a kind of persistence with respect to possible intention (including goal and plan) and agenda:

$$\mathbf{I}(\pi, \varphi) \wedge \mathbf{Com}(\pi) \wedge \mathit{happy}(\pi, \varphi) \rightarrow [\mathit{deliberate}](\mathbf{I}(\pi, \varphi) \wedge \mathbf{Com}(\pi))$$

[4] Here we deviate from our previous work [11, 15], where we use a knowledge operator rather than a belief one. We feel that in the present context belief is more appropriate, since in the next section we will reason about the deliberation of an agent, which may be wrong in its assessment of action results.

[5] We abuse our language here slightly, since strictly speaking we do not have quantification in our object language. See [15] for a proper definition.

[6] Castelfranchi [3] calls this the emotion of a *confirmed, encouraged, enhanced* agent, a particular form of happiness. Here we stick to the term 'happy' in line with [17] for the sake of having an appealingly concise name of the operator

[7] The set of subgoals (mile stones) that are considered important by the agent, is a parameter of this notion of happiness, which is clearly application-dependent.

This is to be regarded as a requirement / condition on the deliberation process, which should be such that $\mathbf{I}(\pi,\varphi)$ and $\mathbf{Com}(\pi)$ persist.

4.2 Sadness

A sad agent is disappointed about the way its plans are progressing, and will look for ways of revising its plans, or perhaps even adjust the goals to be achieved) and make them more realistic. [8] The way sadness comes about is similar to that of happiness. Formally.

$$\mathbf{I}(\pi,\varphi) \wedge \mathbf{Com}(\pi) \wedge \alpha \preceq \pi \wedge \mathbf{B}([\alpha]\psi) \rightarrow$$

$$[\alpha]((\mathbf{B}\neg\psi \wedge \mathbf{Com}(\pi\backslash\alpha)) \rightarrow \mathit{sad}(\pi\backslash\alpha,\varphi))$$

(Since sadness is induced when *any* anticipated subgoal is not believed to be realized, this axiom can be phrased in a simpler form than that for happiness where a ternary fluent had to be used.) In particular we have as a consequence:

$$\mathbf{I}(\pi,\varphi) \wedge \mathbf{Com}(\pi) \rightarrow [\pi](\mathbf{B}\neg\varphi \rightarrow \mathit{sad}(\epsilon,\varphi))$$

Note that we did not postulate a direct relation between sadness and happiness, such as $\mathit{sad}(\pi,\varphi) \leftrightarrow \neg\mathit{happy}(\pi,\varphi)$. In fact, the postulates / constraints for happiness and sadness that we have given so far do suggest (but this depends on the other possible constraints that might be around) that e.g. both $\mathit{sad}(\epsilon,\varphi)$ and $\mathit{happy}(\epsilon,\varphi)$ do not occur at the same time (since in our logic $\mathbf{B}\varphi \wedge \mathbf{B}\neg\varphi$ is inconsistent). However, it might be that neither $\mathbf{B}\varphi$ nor $\mathbf{B}\neg\varphi$ holds after the performance of π so that there is reason for neither happiness nor sadness...

Sadness results in a revision of intention/plan or goal:

$$\mathbf{I}(\pi,\varphi) \wedge \mathbf{Com}(\pi) \wedge \mathit{sad}(\pi,\varphi) \rightarrow [\mathit{deliberate}](\neg\mathbf{I}(\pi,\varphi) \vee \neg\mathbf{Com}(\pi)$$

$$\vee\ \mathbf{Com}(\mathit{if}\ \mathbf{I}(\pi,\varphi)\ \mathit{then}\ \pi\ \mathit{else}\ \mathit{replan}(\pi',\varphi)))$$

Here $\mathit{replan}(\pi',\varphi)$ is an action that constructs a new plan π' for achieving φ for which it should be assumed that $\mathbf{I}(\pi',\varphi)$ holds (cf. [6]). The formula expresses that sadness causes the agent either to drop its (possible) intention (i.e. it does not believe that it can achieve its goal any longer or it has dropped its goal altogether) or uncommit to the plan or try to achieve the goal again by the old plan if that is now possible for him (see the definition of $\mathbf{I}$) or by a new plan.

4.3 Anger

An agent gets angry if its active plan is frustrated. We can coin this frustration in our setting as not being able to perform the plan:

$$\mathbf{Com}(\pi) \wedge \neg\mathbf{Can}(\pi,\mathtt{tt}) \rightarrow \mathit{angry}(\pi)$$

Of course, it depends on the type of agent whether this situation makes him angry. (One might also imagine an agent which is much more 'cool' and just will drop a current commited plan that is frustrated.) So the above formula is to be viewed as a possible characterisation of a particular agent type. An angry agent will try to see to it that he *will* be able to achieve his plan and goal:

$$\mathit{angry}(\pi) \rightarrow [\mathit{deliberate}]\mathbf{Com}(\mathit{stit}(\mathbf{Can}(\pi,\mathtt{tt})))$$

Here $\mathit{stit}(\varphi)$ stands for a basic action that (somehow) sees to it that φ [2]. Bearing the definition of $\mathbf{Can}$ in mind, this means that the agent will try to improve its (believed) capabilities and/or place itself in (a) situation(s) where (it believes) it has the opportunity to perform its plan successfully. We may also consider a more refined notion of anger, where one records the goal one had in mind and the action that frustrated the fulfillment of this goal and the plan associated with it. This notion is characterized by: for $\alpha \preceq \pi$,

$$\mathbf{I}(\pi,\varphi) \wedge \mathbf{Com}(\pi) \rightarrow [\alpha](\mathbf{B}(\neg\varphi \wedge \neg\mathbf{P}(\pi\backslash\alpha,\varphi)) \rightarrow \mathit{angry}(\pi\backslash\alpha,\pi,\varphi))$$

In words, if an agent has the possible intention to do π with goal φ to which it has committed and the performance of the action α results in a state where the agent believes it has not succeeded yet in achieving φ while it also believes that it has not the practical possibility to achieve φ by persuing the rest of its plan $\pi\backslash\alpha$ then it is angry with respect to $\pi\backslash\alpha$ and its plan π and goal φ. In particular, we have, equating $\mathbf{P}(\epsilon,\varphi)$ with φ:

$$\mathbf{I}(\pi,\varphi) \wedge \mathbf{Com}(\pi) \rightarrow [\pi](\mathbf{B}\neg\varphi \rightarrow \mathit{angry}(\epsilon,\pi,\varphi))$$

Now we can, for $\alpha \preceq \pi$, put a constraint on the deliberation process when angry in this sense like the following:

$$\mathit{angry}(\alpha,\pi,\varphi) \rightarrow [\mathit{deliberate}]\mathbf{Com}(\mathit{stit}(\mathbf{Can}(\alpha,\varphi)))$$

As an example of reasoning with these formal notions of emotion (and also to illustrate that we can now also investigate logical relations between these notions!), we show that the above notion of anger is related with sadness under certain circumstances. Suppose $\mathbf{I}(\pi,\varphi)) \wedge \mathbf{Com}(\pi)$. Furthermore, we assume that it holds that $[\alpha]\mathbf{B}(\neg\varphi \wedge \neg\mathbf{P}(\pi\backslash\alpha,\varphi))$, and that the plan $\pi\backslash\alpha$ is believed to be strongly deterministic by the agent. Then, besides $[\alpha]\mathit{angry}(\pi\backslash\alpha,\pi,\varphi)$, we have: $[\alpha]\mathbf{B}(\neg\mathbf{P}(\pi\backslash\alpha,\varphi))$ and so

$$[\alpha]\mathbf{B}(\neg\langle\pi\backslash\alpha\rangle\varphi \vee \neg\mathbf{A}(\pi\backslash\alpha) \vee \neg\mathbf{O}(\pi\backslash\alpha))$$

Since $\pi\backslash\alpha$ is strongly deterministic, we have that $\langle\pi\backslash\alpha\rangle\varphi \leftrightarrow [\pi\backslash\alpha]\varphi$. If we furthermore suppose that $[\alpha]\mathbf{B}(\mathbf{A}(\pi\backslash\alpha) \wedge \mathbf{O}(\pi\backslash\alpha))$ i.e. after doing α the agent believes that it is able to do the rest of its plan and that it has the opportunity to do so), we obtain $[\alpha]\mathbf{B}\neg[\pi\backslash\alpha]\varphi$. Furthermore, we have $\mathbf{I}(\pi,\varphi)$, so $\mathbf{B}\langle\pi\rangle\varphi$, so $\mathbf{B}[\pi]\varphi$, and thus $\mathbf{B}[\alpha]([\pi\backslash\alpha]\varphi)$. Finally, we have that $\mathbf{Com}(\pi) \rightarrow [\alpha]\mathbf{Com}(\pi\backslash\alpha)$ for $\alpha \preceq \pi$. So altogether we now have:

$$\mathbf{I}(\pi,\varphi) \wedge \mathbf{Com}(\pi) \wedge \alpha \preceq \pi \wedge \mathbf{B}[\alpha]([\pi\backslash\alpha]\varphi) \wedge [\alpha](\mathbf{B}\neg[\pi\backslash\alpha]\varphi \wedge \mathbf{Com}(\pi\backslash\alpha))$$

and thus also $[\alpha]\mathit{sad}(\pi\backslash\alpha,\varphi)$. So, in the given circumstances, after the performance of α sadness and anger co-occur. Perhaps one wonders whether this gives rise to impossible constraints on the deliberation process. This is not the case since it is readily checked that the two conditions are consistent.

4.4 Fear

Fear comes about if some crucial self-preservation goal[9] ψ is threatened. Since it is hard to uniquely specify how fear comes about, we will not give an axiom for this, and just treat it as an atomic fluent

[8] In particular this is the emotion of a *disheartened, discouraged* agent ([3]). Again we use the more general label 'sad' in line with [17].

[9] Note that a self-preservation goal should be considered as a kind of *maintenance goal*, for which obviously it does not hold that $\mathit{Goal}_m(\psi) \rightarrow \neg\psi$, as is the case with our regular notion of (achievement) goal. For this reason we denote such a goal with Goal_m rather than $\mathbf{G}$.

(predicate).[10] Fear is similar to sadness, in the sense that a fearful agent will interrupt current plans. But whereas in the case of sadness the current plan is fundamentally revised to obtain the original goal (or perhaps a completely different one, for that matter), here the agent is (overly) cautious. It will constantly observe and check its environment. In particular, it will constantly check whether some crucial maintenance goal ψ is still valid: so, a fearful agent will constantly put a check for ψ on top of its agenda:

$$Goal_m(\psi) \wedge \mathbf{Com}(\pi) \wedge fearful(\psi) \rightarrow$$

$$[deliberate]\mathbf{Com}(if\ \psi\ then\ \pi\ else\ stit(\psi); \pi)$$

5 CONCLUSION AND FUTURE WORK

In this paper we have made a case for the usefulness of the concept of emotion in devising artificial agent-based systems. The notion of emotion can be used as a further structuring element in the line of taking an intentional stance and employing BDI-like cognitive notions to organise agent architectures and programming. We have also indicated how a formal description of emotional agents may look like, building on top of our KARO theory for rational agents, where an emphasis lies on the dynamics of mental (including emotional) states of agents and the effects on their actions and behaviours. As a disclaimer we would like to stress: our paper is certainly not meant to be the ultimate logical theory of emotion, but rather a promising first step, showing that certain aspects of emotion are amenable to logical analysis and representation, which in turn can be employed for the specification of artificial agent-based systems. One aspect, for example, that we have discarded completely is that of intensities of emotions. We have only looked at a very abstract qualitative analysis. Since we have related emotions to the deliberation cycle of (rational) agents, we do feel we have operated in the spirit of Damasio showing how emotions may play a (supporting) role in the agent's decision-making and its BDI-like mental attitudes more in general!

The next step would be to really put this formal theory to work in a concrete architecture or agent programming language. We believe that this is not too hard to do in principle, since agent programming languages like our own 3APL [10] are especially devised to implement mental state changes in terms of beliefs, goals and commitments. Also, 3APL seems to be suited for dealing with the higher level of attitudes that are associated with emotions as we have described in this paper, since it has, besides beliefs and goals, also practical reasoning rules that enable one to program goal / commitment changes under specified conditions on the belief base. These rules are handled by the deliberation cycle (main loop) of the interpreter of 3APL, which decides which rules to pick and apply to which goals. We are now attempting to make this deliberation cycle programmable itself [6] in order to obtain control over the kind of higher-level attitudes that correspond to emotions. If we succeed in this we will be able to experiment and see whether emotions can really improve the efficacy of agents.

Finally, we remark that for future research it is very interesting to incorporate also *multi*-agent aspects. This includes the aspects of the four basic emotions as discussed in section 2, but also forms of emotions that are particularly directed at other agents such as being happy-for, sorry-for, angry-with,... Here is a rich area still to be explored in a more logical way.

ACKNOWLEDGEMENTS

The author likes to thank Aaron Sloman and Paolo Petta for introducing him to the subject of emotions, and Cristiano Castelfranchi and the anonymous referees for their comments on this paper.

REFERENCES

[1] M.B. Arnold, *Emotion and Personality*, Columbia University Press, New York, 1960.

[2] N. Belnap & M. Perloff: Seeing To It That: A Canonical Form for Agentives. *Theoria* 54, 1988, pp. 175-199.

[3] C. Castelfranchi, personal communication, 2003.

[4] P.R. Cohen & H.J. Levesque, Intention is Choice with Commitment, *Artificial Intelligence* 42(3), 1990, pp. 213–261.

[5] A.R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Grosset / Putnam Press, New York, 1994.

[6] M. Dastani, F. de Boer, F. Dignum, J.-J. Ch. Meyer, Programming Agent Deliberation: An Approach Illustrated Using the 3APL Language, in Proc. 2nd Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS03), (J.S. Rosenschein, T. Sandholm, M. Wooldridge, M. Yokoo, eds.), Melbourne Australia, ACM Press, New York, 2003, pp. 97-104.

[7] N. Frijda, *The Emotions*, Cambridge University Press, New York, 1987.

[8] P.J. Gmytrasiewicz & C.L. Lisetti, Emotions and Personality in Agent Design and Modeling, in: *Intelligent Agents VIII* (J.-J. Ch. Meyer & M. Tambe, eds.), LNAI 2333, Spinger, 2002, pp. 21–31.

[9] D. Harel, D. Kozen & J. Tiuryn, *Dynamic Logic*, The MIT Press, Cambridge MA, 2000.

[10] K.V. Hindriks, F.S. de Boer, W. van der Hoek & J.-J. Ch. Meyer, Agent Programming in 3APL, *Int. J. of Autonomous Agents and Multi-Agent Systems* 2(4), 1999, pp.357–401.

[11] W. van der Hoek, B. van Linder & J.-J. Ch. Meyer, An Integrated Modal Approach to Rational Agents, in: M. Wooldridge & A. Rao (eds.), *Foundations of Rational Agency*, Applied Logic Series 14, Kluwer, Dordrecht, 1998, pp. 133–168.

[12] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Seeing is Believing (And So Are Hearing and Jumping), *Journal of Logic, Language and Information* 6, 1997, pp. 33–61.

[13] Z. Manna & A. Pnueli, Temporal Verification of Reactive Systems, Springer, New York/Berlin, 1995.

[14] J.-J. Ch. Meyer & W. van der Hoek, *Epistemic Logic for AI and Computer Science*, Cambridge Tracts in Theoretical Computer Science 41, Cambridge University Press, 1995.

[15] J.-J. Ch. Meyer, W. van der Hoek & B. van Linder, A Logical Approach to the Dynamics of Commitments, *Artificial Intelligence* 113, 1999, 1–40.

[16] K. Oatley & P.N. Johnson-Laird, The Communicative Theory of Emotions: Empirical Tests, Mental Models, and Implications for Social Interaction, in: L.L. Martin & A. Tesser (eds.), *Goals and Affect*, Erlbaum, Hillsdale, NJ, 1995.

[17] K. Oatley & J.M. Jenkins, *Understanding Emotions*, Blackwell Publishing, Malden/Oxford, 1996.

[18] A. Ortony, G.L. Clore & A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, 1988.

[19] A.S. Rao & M.P. Georgeff, Decision Procedures for BDI Logics, *J. of Logic and Computation* 8(3), 1998, pp. 293–344.

[20] Y. Shoham, Agent-Oriented Programming, *Artificial Intelligence* 60(1), 1993, pp. 51–92.

[21] A. Sloman, Motives, Mechanisms, and Emotions, in: *The Philosophy of Artificial Intelligence* (M. Boden, ed.), Oxford University Press, Oxford, 1990, pp. 231–247.

[22] A. Sloman, 'Damasio, Descartes, Alarms, and Meta-Management', in: *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC'98)*, IEEE Computer Society Press, Los Alamitos CA, 1998, pp. 2652–2657.

[23] M.J. Wooldridge, Intelligent Agents, in: *Multiagent Systems* (G. Weiss, ed.), The MIT Press, Cambridge, MA, 1999, pp. 27–77.

[24] M.J. Wooldridge, *Reasoning about Rational Agents*, The MIT Press, Cambridge, MA, 2000.

[10] One reason for the occurrence of fear with respect to a self-preservation goal ψ might be that it is in conflict with another (achievement) goal or with the execution of a plan for a certain achievement goal. The former case could be formalised by $\varphi \rightarrow \neg\psi \models (Goal_m(\psi) \wedge \mathbf{G}(\varphi)) \rightarrow fearful(\psi)$.