Reasoning about Emotional Agents

John-Jules Ch. Meyer
Utrecht University
Institute of Information and Computing Sciences
Intelligent Systems Group
P.O. Box 80.089
3508 TB Utrecht, The Netherlands

February 17, 2004

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 [7] relates the story of a patient called 'Elliott' having a certain kind of brain damage preventing letting him have (secondary) emotions (cf. [36]). 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 [36, 37, 38, 39] 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. In [32] two scenarios are presented in which agents are equipped with emotional states albeit in a very primitive form which have a definite role in their behaviour. In the first scenario a robot exploring a planet may get into a state of 'fear' in which it is more alert in the sense of giving more attention (i.e. resources like time and energy) to the perception and motor systems while in a 'normal' state it devotes its attention / resources more or less entirely to the exploring task and the data analysis associated with it. The second scenario is somewhat more elaborate and involves a robot serving as a personal assistant with 5 emotional states: 'feeling good' (when it performs its job appropriately for a user that is satisfied with it), 'feeling bad' (when the user is dissatisfied with the performance of the agent), 'no emotion', 'feeling curious' (when it has been feeling good for some period of time, curious to explore new ways of helping the user), 'being puzzled'). As Picard writes: *in both scenarios the computer's emotions are labels for states that may not exactly match the analogous human feelings, but that initiate behavior we would expect someone in that state to display*([32]). Thus, the emotional state the agent is in determines its (re/pro)actions, at least partly! One may imagine describing the agent's behaviour by means of some kind of finite automaton, where in a state (also) emotional aspects are captured (cf. [11]).

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. As Oatley and Jenkins put it in [30], p. 258:

> What evolution has equipped us with [...] is a set of emotional states that *organize ready repertoires of action.* Although not perfect, emotions are better than doing nothing, or than acting randomly, or than becoming lost in thought. *Emotions are heuristics.* (My italics)

Another thing that comes to mind when considering the above quote from [32] is the aspect that emotions serve as *labels*, not exactly corresponding to the 'real' thing in humans. This is an aspect that I want to emphasize. Emotional / affective notions may be used to help us think about the agent that we want to build and may help us to structure its design. This is exactly the same as taking the intentional stance in the design of rational agents [9]. Here we may think of agents and its behaviour in terms of informational and motivational attitudes like the well-known BDI architecture for rational agents. The terms beliefs, desires and intentions are of course inspired by the mental attitudes of natural, mostly human agents, but are then used in a more of less metaphoric way to think how artificial agents (should) work: maintaining beliefs about the world, having desires / goals to achieve, employing plans to achieve them, generating intentions on the basis of these plans determining what action to perform next. Although one may argue about it, I tend to see this use of beliefs, desires and intentions in a rather technical way pertaining to certain properties of behaviour of the agent on the one hand and to such profane things like databases in which these beliefs, desires and plans are stored in some symbolic representation on the other. This is also the way I like to view emotional notions: in a metaphoric manner, using them in a more or less technical sense to guide our design of agents. As I've argued above I believe this technical use makes sense. One may conceive of an intricate architecture, where rational and emotional aspects are combined in some structured manner. For example, one may think of Sloman's CogAff architecture ([36]) which is a 'Triple Tower'-like architecture (cf. [28], Ch. 25, where there are separate 'towers' for perception, modelling/reasoning and action) with some extra monitoring devices which are associated with several kinds of emotions, but one may also think of a layered hybrid architecture for agents with reactive and deliberative layers [40], where an emotional state is added determining the priorities between the various layers. For instance, 'normally' the deliberative layer takes precedence of the computation so that well-conceived plans are generated, while in a 'panic' state the agent gives complete priority to the reactive layers to overcome acute problems or escape immanent disaster. A concrete architecture along similar lines has recently been proposed by [5].

So I 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 issue / 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 I'd like to do the same regarding emotions and put the issue of whether computer-based systems may have true emotions (with all the philosophical / ethical ramifications) aside, and concentrate on the engineering aspects of emotions.

From all this we claim 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. from this it follows that it is useful to reason about emotions in an agents, 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, mostly in some pre-/postcondition form where it is indicated how statements in the programming language change the *state* of the system, often expressed in terms of the values of the programming variables employed (cf. [2]).

In *reactive system* specification typically 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 (typically by exchanging messages) (cf. e.g. [22]).

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 [42], we can describe their behaviour in terms of the evolution of the mental states (pertaining to these informational and motivational attitudes) of the agent over time (e.g. Rao & Georgeff's BDI logic [33, 34], which is a branching-time temporal logic enhanced with modal operators for beliefs, desires and intentions, the logical approach of Cohen & Levesque [6], based on linear-time temporal logic, and our own KARO approach [19, 17], based on dynamic logic). Indeed, Shoham in his seminal paper ([35]) on agent-oriented programming says that agent programs are 'mental state transformers'.

We want to take this a step further and also perceive emotional agents as systems that evolve over time and can be described by some logic as the one mentioned above for rational agents. Indeed, as Michael Wooldridge in his book ([41]) describes a Logic of Rational Agents (LORA) built upon Rao & Georgeff's BDI logic, we like to make a start at considering her sister LEA (Logic of Emotional Agents), where we will take our KARO logic as a starting point.

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, viz. the KARO framework. In some sense the KARO framework is perfectly suited for dealing with goal and commitment maintenance, as this was already needed for an adequate treatment of motivational attitudes as described in [26].

In essence our approach is thus: to reason about the dynamics of (emotional) states we use the framework of *dynamic logic* ([13] analogously to what we did for reasoning about rational agents in the KARO framework ([19, 20, 21, 17, 26, 18, 24, 23, 27]. (KARO is based on a blend of dynamic logic with epistemic logic [25], enriched with modal operators for motivational attitudes such as desires and goals.) In fact we will extend the KARO framework to cater for emotional attitudes.

2 Psychological preliminaries

As noted before, in psychological theory emotions are associated with higher-level mental attitudes [29, 30]. As stated in [31], emotions have many facets: involving feelings and experience, physiology and behaviour, as well as cognitions and conceptualisations. Of course, also the expression of emotions is an important subject of study in cognitive science. However, in this paper, as we are interested in constructing artificial agents using emotions as a 'designing tool', we concentrate on their relation to the agent's behaviour, and in particular the agent's actions. This relation has been studied in cognitive science as well. (For example, the work of Arnold ([1]) and Frijda ([10]) discusses the relation of emotions with action tendencies and readiness.)

In general terms one may distinguish so called *"well-being emotions"* and *"prospect-based emotions"* with respect to actions and events ([31]). Both have positive and negative variants. Well-being emotions include being pleased / displeased about a(n un)desirable event, which one may call *joy* or *happiness* and *distress* or *sadness*. Prospect-based emotions can be centered around:

- the *prospect* of an event: pleased or displeased (hope emotions and fear emotions, respectively)
- the *confirmation* of a prospect: satisfaction and fear-confirmed emotions
- the *disconfirmation* of a prospect: relief and disappointment emotions

(Dis)confirmation emotions always co-occur with the well-being emotions (joy and distress emotions, respectively).

Here we describe some of the basic emotions as discussed in [30], 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*.[1]

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.

[1] Some cognitive scientists prefer to think of these as labels of 'families' of emotions rather than specific emotions [4]. Here I follow [30] and use them as conveniently concise labels of the emotions that we will treat formally in the sequel.

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.

Although these descriptions are naturally aimed at human emotions, we feel that some crucial ingredients of these are also important for constructing intelligent artifacts and can, moreover, be captured formally. We also see that some of the attitudes contain both individual and social aspects. Happiness, for example, involves both the individual attitude to continue with one's plan and the social attitude of cooperating, while sadness involves both searching for a new plan yourself and seeking help. We believe that in a theory of multi-agent systems the social aspects are important as well as formalizable, but for the purpose of this paper we will restrict ourselves to the individual aspects.

3 KARO Logic

In this section we 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. [20, 21, 17, 26]) as well as the thesis of Van Linder ([19]).

The KARO formalism is an amalgam of dynamic logic and epistemic / doxastic logic [25], augmented with several additional (modal) operators in order to deal with the motivational aspects of agents. So, besides operators for knowledge (**K**), belief (**B**) and action ($[\alpha]$, "after performance of α it holds that"), there are additional operators for ability (**A**) and desires (**D**).

Assume a set $\mathcal{A}$ of atomic actions and a set $\mathcal{P}$ of atomic propositions.

Definition 3.1 *(Language.) The language $\mathcal{L}_{KARO}$ of KARO-formulas is given by the BNF grammar:*

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

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

Here the formulas generated by the second (α) part are referred to as actions (or rather action expressions). We use the abbreviations $\mathtt{tt} \equiv p \vee \neg p$ (for some fixed $p \in \mathcal{P}$) and $\mathtt{ff} \equiv \neg\mathtt{tt}$. Conditional and while-action are introduced by the usual abbreviations: $\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi} \equiv (\varphi?;\alpha_1) + (\neg\varphi?;\alpha_2)$ and $\mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od} \equiv (\varphi?;\alpha)^*;\neg\varphi?$.

Remark 3.2 *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 3.3 *(KARO models.)*

1. *The semantics of the knowledge, belief and desires operators is given by means of Kripke structures of the following form:* $\mathcal{M} = \langle W, \vartheta, R_K, R_B, R_D \rangle$*, where*
   - W *is a non-empty set of states (or worlds)*
   - ϑ *is a truth assignment function per state*
   - R_K, R_B, R_D *are accessibility relations for interpreting the modal operators* $\mathbf{K}, \mathbf{B}, \mathbf{D}$. *The relation* R_K *is assumed to be an equivalence relation, while the relation* R_B *is assumed to be euclidean, transitive and serial. Furthermore we assume that* $R_B \subseteq R_K$. *No special constraints are assumed for the relations* 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.*

Remark 3.4 *We have elements in the structures for interpreting the operators for knowledge, belief, and desire. Actions are modelled as model/state pair transformers to emphasize their influence on the mental state (that is, the complex of knowledge, belief and desires) of the agent rather than just the state of the world. Both (cap)abilities and commitments are given by functions that yield the relevant information per model / state pair.*

Definition 3.5 *(Interpretation of formulas.) 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:*

- $\mathcal{M}, w \models p$ *iff* $\vartheta(w)(p) = \text{true}$, *for* $p \in \mathcal{P}$
- *The logical connectives are interpreted as usual.*
- $\mathcal{M}, w \models \mathbf{K}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R_K(w, w')$
- $\mathcal{M}, w \models \mathbf{B}\varphi$ *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* $\alpha \in \mathcal{C}(\mathcal{M}, w)$[2]
- $\mathcal{M}, w \models \mathbf{Com}(\alpha)$ *iff* $\alpha \in Ag(\mathcal{M}, w)$[3]

[2] In [18] we have shown that the ability operator can alternatively defined by means of a second accessibility relation for actions, in a way analogous to the opportunity operator below.

[3] The agenda is assumed to be closed under certain conditions such as taking 'prefixes' of actions (representing initial computations). Details are omitted here, but can be found in [26].

Here R_α is defined as usual in dynamic logic by induction from the basic case R_a (cf. e.g. [12, 19, 17], but now on model/state pairs rather than just states). 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 ([19, 17]). We call an action α *deterministic* if $\#\{w' \mid R_\alpha(w,w')\} \leqslant 1$ for any $w \in W$, and *strongly deterministic* if $\#\{w' \mid R_\alpha(w,w')\} \leqslant= 1$. (Here $\#$ stands for cardinality.)

Remark 3.6 *We have clauses for knowledge, belief and desire. The action modality gets a similar interpretation: something (necessarily) holds after the performance / execution of action α if it holds in all the situations that are accessible from the current one by doing the action α. The only thing which is slightly nonstandard is that, as stated above, a situation is characterised here as a model / state pair. The interpretations of the ability and commitment operators are rather trivial in this setting (but see the footnotes): an action is enabled (or rather: the agent is able to do the action) if it is indicated so by the function C, and, likewise, an agent is committed to an action α if it is recorded so in nthe agent's agenda.*

Furthermore, we will make use of the following syntactic abbreviations serving as auxiliary operators:

Definition 3.7

- *(dual) $\langle\alpha\rangle\varphi = \neg[\alpha]\neg\varphi$, expressing that the agent has the opportunity to perform α resulting in a state where φ holds.*
- *(opportunity) $\mathbf{O}\alpha = \langle\alpha\rangle\mathtt{tt}$, i.e., an agent has the opportunity to do an action iff there is a successor state w.r.t. the R_α-relation;*
- *(practical possibility) $\mathbf{P}(\alpha,\varphi) = \mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$, i.e., 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;*
- *(can) $\mathbf{Can}(\alpha,\varphi) = \mathbf{BP}(\alpha,\varphi)$, i.e., 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]*, i.e., 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) $\mathbf{G}\varphi = \neg\varphi \wedge \mathbf{D}\varphi \wedge \Diamond\varphi$, i.e., a goal is a formula that is not (yet) satisfied, but desired and realisable.*[6]

[4] Here we deviate from our previous work [17, 26], 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 [26] for a proper definition.

[6] In fact, here we simplify matters slightly. In [26] we also stipulate that a goal should be explicitly selected somehow from the desires it has, which is modelled in that paper by means of an additional modal operator. Here we leave this out for simplicity's sake.

- *(possible intend)* $\mathbf{I}(\alpha, \varphi) = \mathbf{Can}(\alpha, \varphi) \wedge \mathbf{BG}\varphi$, *i.e., 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.*[7].

Remark 3.8

- *The dual of the (box-type) action modality expresses that there is at least a resulting state where a formula* φ *holds. It is important to note that in the context of* deterministic *actions, i.e. actions that have at most one successor state, this means that the* only *state satisfies* φ*, and is thus in this particular case a stronger assertion than its dual formula* $[\alpha]\varphi$*, which merely states that if there are any successor states they will (all) statisfy* φ*. Note also that if atomic actions are assumed to be deterministic all actions including the complex ones will be deterministic.*
- *Opportunity to do an action is modelled by having at least one successor state according to the accessibility relation associated with the action.*
- *Practical possibility to to an action with a certain result is modelled as having both ability and opportunity to do the action with the appropriate result. Note that* $\mathbf{O}\alpha$ *in the formula* $\mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$ *is actually redundant since it already follows from* $\langle\alpha\rangle\varphi$*. However, to stress the opportunity aspect it is added.*
- *The Can predicate applied to an action and formula expresses that the agent is 'conscious' of its practical possibility to do the action resulting in a state where the formula holds.*
- *A formula* φ *is realisable if there is a 'plan' consisting of (a sequence of) atomic actions of which the agent has the practical possibility to do them with* φ *as a result.*
- *A formula* φ *is a goal in the KARO framework if it is not true yet, but desired and realisable in the above meaning, that is, there is a plan of which the agent has the practical possibility to realise it with* φ *as a result.*
- *An agent is said to (possibly) intend an action* α *with result* φ *if it 'Can' do this (believes that he has the practical possibility to do so), and, moreover, believes that* φ *is a goal.*

In order to manipulate both knowledge / belief and motivational matters special actions `revise`, `commit` and `uncommit` are added to the language. (We assume that we cannot nest these operators. So, e.g., `commit` (`uncommit`α) is not a well-formed action expression. For a proper definition of the language the reader is referred to [26].) The semantics of these are again given as model/state transformers (We only do this here in a very abstract manner, viewing the accessibility relations associated with these actions as functions. For further details we refer to e.g. [19, 17, 26]):

Definition 3.9 *(Accessibility of revise, commit and uncommit actions.)*

[7] Again we deviate from our previous work, and use a belief operator rather than a knowledge operator in this definition

1. *$R_{\texttt{revise}\varphi}(\mathcal{M}, w) = update\_belief(\varphi, (\mathcal{M}, w))$.*
2. *$R_{\texttt{commit}\alpha}(\mathcal{M}, w) = update\_agenda^{+}(\alpha, (\mathcal{M}, w))$, if $\mathcal{M}, w \models \mathbf{I}(\alpha, \varphi)$ for some φ, otherwise $R_{\texttt{commit}\alpha}(\mathcal{M}, w) = \emptyset$ (indicating failure of the commit action).*
3. *$R_{\texttt{uncommit}\alpha}(\mathcal{M}, w) = update\_agenda^{-}(\alpha, (\mathcal{M}, w))$, if $\mathcal{M}, w \models \mathbf{Com}(\alpha)$, otherwise $R_{\texttt{uncommit}\alpha}(\mathcal{M}, w) = \emptyset$ (indicating failure of the uncommit action);*
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.*

Remark 3.10 *Here $update\_belief$, $update\_agenda^{+}$ and $update\_agenda^{-}$ are functions that update the agent's belief and agenda (by adding or removing an action), respectively. Details are omitted here, but essentially these actions are model/state transformers again, representing a change of the mental state of the agent (regarding beliefs and commitements, respectively). The $update\_belief(\varphi, , (\mathcal{M}, w))$ function changes the model $\mathcal{M}$ in such a way that the agent's belief is updated with the formula φ, while $update\_agenda^{+}(\alpha, (\mathcal{M}, w))$ changes the model $\mathcal{M}$ such that α is added to the agenda, and like wise for the $update\_agenda^{-}$ function, but now with respect to removing an action from the agenda. The formal definitions can be found in [20, 21] and [26]. The `revise` operator can be used to cater for revisions due to observations and communication with other agents, which we will not go into further here (see [21]).*

The interpretation of formulas containing revise and (un)commit actions is now done using the accessibility relations above. One can now define validity as usual with respect to the KARO-models. One then obtains the following validities (of course, in order to be able to verify these one should use the proper model and not the abstraction / simplification we have presented here.) Typical properties of this framework, called the KARO logic, include (cf. [20, 26]):

Proposition 3.11

1. $\models \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$, *for* $\Box \in \{\mathbf{K}, \mathbf{B}, \mathbf{D}, [\alpha]\}$
2. $\models \langle\alpha\rangle\varphi \to [\alpha]\varphi$, *for deterministic* α
3. $\models \Box\varphi \to \Box\Box\varphi$, *for* $\Box \in \{\mathbf{K}, \mathbf{B}\}$
4. $\models \neg\Box\varphi \to \Box\neg\Box\varphi$, *for* $\Box \in \{\mathbf{K}, \mathbf{B}\}$
5. $\models \mathbf{K}\varphi \to \varphi$
6. $\models \neg\mathbf{B}\texttt{ff}$
7. $\models \mathbf{O}(\alpha; \beta) \leftrightarrow \langle\alpha\rangle\mathbf{O}\beta$
8. $\models \mathbf{Can}(\alpha; \beta, \varphi) \leftrightarrow \mathbf{Can}(\alpha, \mathbf{P}(\beta, \varphi))$
9. $\models \mathbf{I}(\alpha, \varphi) \to \mathbf{B}\langle\alpha\rangle\varphi$
10. $\models \mathbf{I}(\alpha, \varphi) \to \langle\texttt{commit}\alpha\rangle\mathbf{Com}(\alpha)$
11. $\models \mathbf{I}(\alpha, \varphi) \to \neg\mathbf{A}\texttt{uncommit}(\alpha)$
12. $\models \mathbf{Com}(\alpha) \to \langle\texttt{uncommit}(\alpha)\rangle\neg\mathbf{Com}(\alpha)$

13. $\models \mathbf{Com}(\alpha) \wedge \neg\mathbf{Can}(\alpha, \top) \rightarrow \mathbf{Can}(\mathtt{uncommit}(\alpha), \neg\mathbf{Com}(\alpha))$

14. $\models \mathbf{Com}(\alpha) \rightarrow \mathbf{K}\mathbf{Com}(\alpha)$

15. $\models \mathbf{Com}(\alpha_1; \alpha_2) \rightarrow \mathbf{Com}(\alpha_1) \wedge \mathbf{K}[\alpha_1]\mathbf{Com}(\alpha_2)$

16. $\models \mathbf{Com}(\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}) \wedge \mathbf{K}\varphi \rightarrow \mathbf{Com}(\varphi?; \alpha_1)$

17. $\models \mathbf{Com}(\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}) \wedge \mathbf{K}\neg\varphi \rightarrow \mathbf{Com}(\neg\varphi?; \alpha_2)$

18. $\models \mathbf{Com}(\mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}) \wedge \mathbf{K}\varphi \rightarrow \mathbf{Com}((\varphi?; \alpha); \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od})$

Remark 3.12 *The first of these properties says that all the modalities mentioned are 'normal' in the sense that they are closed under implication. The second states that the dual operator $\langle\alpha\rangle$ is stronger than the operator $[\alpha]$ in case the action α is deterministic: if there is at least one successor state after performing α and we know that there is at least one successor state satisfying φ then* all *successor states satisfy φ. The third and fourth properties are the so-called introspection properties for knowledge and belief. The fifth property says that knowledge is true, while the sixth states that belief (may not be true but) is not inconsistent. The seventh property states that having the opportunity to do a sequential composition of two actions amounts to having the opportunity of doing the first action first and then having the opportunity to do the second. The eighth states that an agent that* can *do a sequential composition of two actions with result φ iff the agent can do the first actions resulting in a state where it has the practical possibility to do the second with φ as result. The ninth states that if one possibly intends to do α with result φ then one believes that there is a possibility of performing α resulting in a state where φ holds. The tenth asserts that if an agent possibly intends to do α with some result φ, it has the opportunity to commit to α with result that it is committed to α (i.e. α is put into its agenda). The eleventh says that if an agent intends to do α with a certain purpose, then it is unable to uncommit to it (so, if it is committed to α it has to persevere in it). The twelfth property says that if an agent is committed to an action and it has the opportunity to uncommit to it with as result that indeed the commitment is removed. The thirteenth says that whenever an agent is committed to an action that is no longer known to be practically possible, it knows that it can undo this impossible commitment. The fourteenth property states that commitments are known to the agent. The last four properties have to do with commitments to complex actions. For instance, the eighth says that if an agent is committed to a sequential composition of two actions then it is committed to the first one, and it knows that after doing the first action it will be committed to the second action.*

4 The Dynamics of Emotion

We are now ready to deal with the logic of emotional agents, where we especially focus on the influence of emotions on agenda maintenance. Instead of 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, or fluents, in the jargon of reasoning about action and change, to indicate that these predicates change over time. 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. [40]. In the KARO framework this looks like a program $(\textit{deliberate}; \textit{execute})^*$, where the actions *deliberate* and *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 ([8]). 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. In other words, our axioms to follow are 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. [8]

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'.[9] We use ϵ to denote the empty sequence (of atomic actions). When $\pi = \alpha; \pi'$, we denote π' by $\pi \backslash \alpha$[10], the remainder of π if its initial part α has already been executed.

happiness An agent that is happy observes that its subgoals (towards certain goals) are being achieved, and is 'happy' with it.[11]. This means that in a happy state the agent does not need to be alarmed or extraordinary cautious with respect to these subgoals as in some other emotional states (see below), and can just keep its agenda and goal. We first describe the situation in which happiness comes about. This is a little intricate. We want to express that an agent that is striving for a particular goal by working on a subgoal by means of

[8] In general, one should consider computation sequences $Comp(\alpha)$ of (semi-)atomic actions α, as defined in e.g. [26].

[9] In the general case $\preceq$ should be defined by means of computation sequences: $\alpha \preceq \pi$ iff $Comp(\alpha)$ is a prefix of $Comp(\pi)$.

[10] In the general case $\pi \backslash \alpha$ is defined as some action such that $Comp(\pi \backslash \alpha) = Comp(\pi) \backslash Comp(\alpha)$.

[11] Castelfranchi [4] 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[30] for the sake of having an appealingly concise name of the operator.

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.[12] Note that the agent initially (when it has just the intention $\mathbf{I}(\pi,\varphi)$, but has not done anything yet to achieve it, is happy with that situation ($\mathit{happy}(\pi,\varphi)$), since it follows from the above (taking $\alpha = \epsilon$) that, for any relevant ψ,

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

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

that is:

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

$$((\mathbf{B}\psi \wedge \mathbf{Com}(\pi\backslash\epsilon)) \rightarrow \mathit{happy}(\pi\backslash\epsilon,\varphi,\psi))$$

so

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

As said before happiness causes a kind of persistence with respect to possible intention (including goal and plan) and agenda:

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

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.

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. [13]

[12] 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.

[13] In particular this is the emotion of a *disheartened, discouraged* agent ([4]). Again we use the more general label 'sad' in line with [30].

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 neither reason for 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. [8]). 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 **I**) or by a new plan.

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, which it can do by Proposition 3.11, item 13.) 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 [3]. Bearing the definition of **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. One might also consider a more refined notion of anger, where one also records the goal one had in mind and the action that frustrated the fulfillment of this goal and the plan associated with it. Here one has to bear in mind that the one-time (possible) intention $\mathbf{I}(\pi, \varphi)$ was the reason for committing to the plan π, and that if at some point in time one cannot execute the plan anymore ($\neg\mathbf{Can}(\pi, \varphi)$), it means that by the execution of some action $\alpha \preceq \pi$[14] the possible intention $\mathbf{I}(\pi, \varphi)$ ceases to hold! So, formally we can specify this type of anger by: for $\alpha \preceq \pi$,

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

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 then it is angry with respect to this action α and its plan π and goal φ. In particular, we have, equating $\mathbf{P}(\epsilon, \varphi)$ with φ:

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

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

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

(In the special case $\alpha = \epsilon$, since in this case there is no real plan left, this amounts to the agent committing to seeing to it that it knows that φ holds — under the reasonable assumption that $\mathbf{A}\epsilon = \mathtt{tt}$.)
Interestingly, this 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 $\mathbf{B}\langle\pi\rangle\varphi$ (by Proposition 3.11, item 9), so $\mathbf{B}[\pi]\varphi$ (since π is deterministic), and thus $\mathbf{B}[\alpha]([\pi\backslash\alpha]\varphi)$.

[14] For simplicity we assume here that the agent is the only actor in the environment that may influence the truth of $\mathbf{Can}(\pi, \varphi)$.

Finally, by Proposition 3.11, item 15, we have that $\mathbf{Com}(\pi) \to [\alpha]\mathbf{Com}(\pi \backslash \alpha)$ 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 $[\alpha]sad(\pi\backslash\alpha, \varphi)$. 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 readily checked that the two conditions are consistent. (Moreover, if one only admits one plan on the agenda, the constraint associated with anger involving a commitment to seeing to it that the agent can perform the rest of its plan eliminates the possibility in the repertoire of dealing with sadness of committing to a replanning to achieve φ).

fear Fear comes about if some crucial self-preservation goal[15] 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 (predicate).[16]

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:

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

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

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

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

[16] 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 $\models \varphi \to \neg\psi \;\Rightarrow\; \models (Goal_m(\psi) \wedge \mathbf{G}(\varphi)) \to fearful(\psi)$.

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.

As a note on the psychological aspect of this paper: Ortony, Clore and Collins [31] had as one of their goals to lay the foundations for a *computationally tractable model of emotion.* Although what we have done in this paper is much more intended towards using emotions as a 'tool' for constructing artificial agents in a better way, in retrospect our work may perhaps also be considered as pushing the goal of Ortony, Clore and Collins one step further into a formal theory, at the cost of large simplification... (As stated before, we only looked at some particular aspects of emotions.)

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 [15, 14] 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 [8] in order to obtain control over the kind of higher-level attitudes that correspond to emotions.

Acknowledgements The author likes to thank Aaron Sloman and Paolo Petta for introducing him to the subject of emotions, and Cristiano Castelfranchi, Marco Wiering, Patrick Pepels, Stefan Leijnen, Cilia Witteman, Wiebe van der Hoek, Peter-Paul van Maanen, Frank Dignum, Mehdi Dastani, and Ismail Ibrahim for various discussions on the subject of emotions, convincing him that emotions are worthwhile studying, also from the perspective of computer science and artificial intelligence... Also my thanks to the attendants of the ISIS colloquium who served as a first audience for the material in this paper, for their feedback.

References

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

[2] J.W. de Bakker, *Mathematical Theory of Program Correctness*, Prentice-Hall International, Englewood Cliffs NJ, 1980.

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

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

[5] L. Chen, K. Bechkoum & G. Clapworthy, Equipping a Lifelike Animated Agent with a Mind, in: *Intelligent Virtual Agents, Proc.*

IVA 2001 (A. de Antonio, R. Aylett & D. Ballin, eds.). LNAI 2190, Springer, Berlin, 2001, pp. 72–85.

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

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

[8] 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.

[9] D.C. Dennett, *The Intentional Stance*, MIT Press, Cambridge, Mass., 1987.

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

[11] 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.

[12] D. Harel, Dynamic Logic, in: D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic, Vol. II*, Reidel, Dordrecht/Boston, 1984, pp. 497–604.

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

[14] K.V. Hindriks, Agent Programming Languages Programming with Mental Models), PhD. Thesis, Utrecht University, Utrecht, 2001.

[15] 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.

[16] W. van der Hoek, Systems for Knowledge and Belief, *Journal of Logic and Computation* 3(2), 1993, pp. 173–195.

[17] 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.

[18] W. van der Hoek, J.-J. Ch. Meyer & J.W. van Schagen, Formalizing Potential of Agents: The KARO Framework Revisited, in: *Formalizing the Dynamics of Information* (M. Faller, S. Kaufmann & M. Pauly, eds.), CSLI Publications, (CSLI Lect. Notes 91), Stanford, 2000, pp. 51–67.

[19] B. van Linder, Modal Logics for Rational agents, PhD. Thesis, Utrecht University, 1996.

[20] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Actions that Make You Change Your Mind: Belief Revision in an Agent-Oriented

Setting, in: *Knowledge and Belief in Philosophy and Artificial Intelligence* (A. Laux & H. Wansing, eds.), Akademie Verlag, Berlin, 1995, pp. 103–146.

[21] 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.

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

[23] J.-J. Ch. Meyer, Dynamic Logic for Reasoning about Actions and Agents, in: *Logic-Based Artificial Intelligence* (J. Minker, ed.), Kluwer, Boston/Dordrecht, 2000, pp. 281–311.

[24] J.-J. Ch. Meyer, F.S. de Boer, R.M. van Eijk, K.V. Hindriks & W. van der Hoek, On Programming KARO Agents, in: Proc. Int. Conf. on Formal and Applied Practical Reasoning (FAPR2000) (J. Cunningham & D. Gabbay, eds.), Imperial College, London, 2000, pp. 93–103 (ISSN1469-4166).

[25] 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.

[26] 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.

[27] J.-J. Ch. Meyer & J. Treur (eds.), Agent-Based Defeasible Control in Dynamic Environments, Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 7, Kluwer, Dordrecht/Boston/London, 2002.

[28] N.J. Nilsson, *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann, San Francisco, 1998.

[29] 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.

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

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

[32] R.W. Picard, Does HAL cry digital tears? Emotion and Computers, Chapter 13 of: *HAL's Legacy* (D.G. Stork, ed.), MIT Press, Cambridge MA, 1997.

[33] A.S. Rao & M.P. Georgeff, Modeling rational agents within a BDI-architecture, in *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)* (J. Allen, R. Fikes & E. Sandewall, eds.), Morgan Kaufmann, 1991, pp. 473–484.

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

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

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

[37] A. Sloman, What kinds of machine can have emotions?, paper presented at the British Association Annual Festival, 1996.

[38] A. Sloman, What sort of architecture is required for a human-like agent?, Techn. Report CSRP-96-12, School of Computer Science and Cognitive Science Research Centre, Birmingham, 1996. Invited talk for Cognitive Modelling Workshop at AAA—I'96.

[39] 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.

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

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

[42] M.J. Wooldridge & N.R. Jennings (eds.), *Intelligent Agents*, Springer, Berlin, 1995.