A Formal Model of Emotion-based Action Tendency for Intelligent Agents*

Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer

Department of Information and Computing Sciences, Intelligent Systems Group,
Utrecht University, The Netherlands
{bass,mehdi,jj}@cs.uu.nl

Abstract. Although several formal models of emotions for intelligent agents have recently been proposed, such models often do not formally specify how emotions influence the behavior of an agent. In psychological literature, emotions are often viewed as heuristics that give an individual the tendency to perform particular actions. In this paper, we take an existing formalization of how emotions come about in intelligent agents and extend this with a formalization of action tendencies. The resulting model specifies how the emotions of an agent determine a set of actions from which it can select one to perform. We show that the presented model of how emotions influence behavior is intuitive and discuss interesting properties of the model.

1 Introduction

There has recently been increasing interest in bringing emotions to Artificial Intelligence, in particular to model intelligent agents [1–10]. There are (at least) three important reasons for this. First, an obvious application of emotions is to make artificial agents and robots more believable to human users (both in the actions that they select and the affective expressions they show based on emotions) [1, 4, 5]. Second, from a more theoretical perspective, it is investigated what the role of emotions is in models of human decision-making and how they may be employed to make them more accurate and effective [11, 2, 12]. Third, there exists psychological [13–16] and neurological [17] evidence that emotions are not only relevant but even necessary for rational behavior.

Although several formal models of emotions for intelligent agents have recently been proposed, these models often only specify when emotions are triggered, but not how they affect the behavior of an agent. In psychological literature, emotions are often viewed as heuristics that give an individual the tendency to perform particular actions [16, 13]. Viewed from this perspective, emotions can be used to determine a set of actions that an agent tends to perform in a certain situation, which is typically a subset of all possible actions it can perform. Moreover, action tendency can provide a measure to order this subset, thereby indicating which action(s) an agent tends to perform most. Thus, a model of emotion-based action tendency can help limiting and ordering options in an agent's action selection process.

* This work is supported by SenterNovem, Dutch Companion project grant nr: IS053013.

In this paper, we incorporate a notion of action tendency in an existing formalization of emotions which is tailored to (artificial) goal-directed agents [9, 10]. The resulting model allows us to reason about situations in which tendencies arise with respect to particular emotions.

Outline: In section 2 we give an overview of psychological literature on emotions and their behavioral effects. A formal model of emotions is presented in section 3 so that a notion of action tendency can be incorporated in section 4. Section 5 discusses alternatives and extensions of the presented model, while section 6 discusses related work.

2 Emotion and Action Tendency in Psychology

There is little consensus among psychologists as to what exactly constitutes an emotion and how it differs from related affective processes. However, this does not mean that making broad classifications is impossible or useless. Following Gross [18], *emotions* typically have specific objects and give rise to action tendencies relevant to these objects. Moreover, emotions can be both positive and negative. Emotions are often distinguished from *moods*, which are more diffuse and last longer than emotions. Other affective processes include *stress*, which arises in taxing circumstances and produces only negative responses; and *impulses*, which are related to hunger, sex, and pain and give rise to responses with limited flexibility. Of these four types of affective processes, we will focus on *emotions* in this paper.

With respect to emotions, usually three phases are distinguished. First, the perceived situation is *appraised* by an individual based on what it thinks is relevant and important (e.g., gratitude is triggered for Alice towards Bob for giving her a necklace). Second, the appraisal of some situation can cause the triggered emotions, if exceeding some threshold, to create a conscious awareness of emotional feelings, leading to the *experience* of having emotions (e.g., Alice's gratitude towards Bob will have a certain intensity, depending on the desirability of receiving a necklace from Bob). Third, emotional feelings need to be *regulated* (e.g., Alice may tend to be nicer to Bob so that he will give her more presents). In fact, some emotion theories posit that the main purpose of emotions is to function as a heuristical mechanism for selecting behaviors [17, 16, 14].

Following Frijda [13] (see also Table 2.1), *action tendencies* are "states of readiness to execute a given kind of action, [which] is defined by its end result aimed at or achieved" (p70). In the case of negative emotions, reaching the associated end state should mitigate its experience (e.g., fear subsides once one believes the object of one's fear cannot reach oneself anymore), whereas positive emotions generally put an individual in a "mode of relational action readiness" (e.g., joy can put one in a mode of readiness for new interactions). In this paper, we will restrict our investigation to action tendencies related to negative emotions, because these signal that some things are not as they should be, making it possible to identify actions that can fix the situation. Indeed, in our formalization, the experience of negative emotions will be constrained such that for some actions, reaching their end state removes the emotional experience. Action tendencies will then follow naturally in response to emotional experience. It should be noted that action tendencies differ from intentions in that they are not goal-directed, but rather stimulus-driven [13]; e.g., fear does not (necessarily) spawn a goal to flee *towards* safety, but rather gives the urge to flee *away* from the perceived danger.

Emotion	Function	Action tendency	End state
Desire	Consume	Approach	Access
Joy	Readiness	Free activation	—
Anger	Control	Agonistic	Obstruction removed
Fear	Protection	Avoidance	Own inaccessibility
Interest	Orientation	Attending	Identification
Disgust	Protection	Rejecting	Object removed
Anxiety	Caution	Inhibition	Absence of response
Contentment	Recuperation	Inactivity	—

Table 2.1. A classification of several relational action tendencies. Adapted from a table on page 88 of Frijda [13].

3 Language and Semantics

In this paper, we build on an existing formalization by Steunebrink *et al.* [9, 10] of the psychological "OCC model" of emotions [19]. The logic of rational agency that underlies this formalization of the OCC model is KARO [6, 20], which is a mixture of dynamic logic, epistemic logic, and several additional operators for dealing with the motivational aspects of artificial agents.

The OCC model describes a hierarchy classifying 22 emotion types. The hierarchy contains three branches, namely emotions concerning aspects of objects (e.g., Alice loves Bob), actions of agents (e.g., Alice admires Bob for helping with her homework), and consequences of events (e.g., Alice pities Bob having lost his job). Additionally, some branches combine to form a group of compound emotions, namely emotions concerning consequences of events *caused* by actions of agents (e.g., Alice is grateful towards Bob that his help resulted in a good grade). It should be noted that emotions are not used to describe the entire cognitive state of an agent (as in "Alice is happy"); rather, emotions are always relative to individual objects, actions, and events. So Alice can be joyous about receiving her new furniture and at the same time be distressed about the height of the accompanying bill.

The OCC model defines both qualitative and quantitative aspects of emotions. Qualitatively, it defines the conditions that trigger each of the emotions. Quantitatively, it describes how an intensity is associated with each triggered emotion and what are the variables affecting emotional intensity. For example, the compound emotion *gratitude* is qualitatively specified as "approving of someone else's praiseworthy action and being pleased about the related desirable event," whereas the variables affecting its (quantitative) intensity are (1) the degree of judged praiseworthiness, (2) the unexpectedness of the event, and (3) the degree to which the event is desirable. It should be noted that in [9, 10], qualitative and quantitative aspects of emotions (of the OCC model) have been formalized, respectively, corresponding to the first two phases described in section 2. In this paper we build upon their work, focusing on action tendencies, which corresponds to the third phase as described in section 2.

In order to formalize the OCC model in an agent specification language, the concepts of *objects*, *actions*, and *events* must be translated to notions common in agent specification languages, which is done as follows. The agent specification language is based on dynamic logic, so the concept of *actions* is directly available. For goal-directed

agents, important *events* pertain to the accomplishment of (sub)goals and the undermining of subgoals (i.e. the undoing of previously accomplished subgoals). Therefore, an event is defined as the *accomplishment* of a set of subgoals (from a single goal) or the *undermining* of a set of subgoals (from a single goal). The *objects* to which an agent can have an affective attitude include all agent names.

The formalization by Steunebrink *et al.* distinguishes between the conditions that trigger emotions and their actual experience. The satisfaction of the triggering conditions of an emotion is seen as merely a property of one moment in time, whereas emotional experience is something that endures over time. This means that an emotion that has been triggered does not necessarily have to be experienced, either because it was triggered some time in the past and its intensity has since dropped to zero, or because it was assigned zero intensity to begin with.

To cater for this distinction, two sets of emotional fluents are defined: *emotion triggering fluents* and *emotional experience fluents*:

$$ETrig = \{\, \mathbf{emotion}^{\mathbf{T}}\bar{o} \mid \mathbf{emotion}\,\bar{o} \in EExp \,\} \tag{3.1}$$

$$\begin{aligned} EExp = \{\, & \mathbf{gratification}_i(\alpha,\varphi), && \mathbf{remorse}_i(\alpha,\varphi), \\ & \mathbf{gratitude}_i(j,\alpha,\varphi), && \mathbf{anger}_i(j,\alpha,\varphi), \\ & \mathbf{pride}_i(\alpha), && \mathbf{shame}_i(\alpha), \\ & \mathbf{admiration}_i(j,\alpha), && \mathbf{reproach}_i(j,\alpha), \\ & \mathbf{joy}_i(\varphi), && \mathbf{distress}_i(\varphi), \\ & \mathbf{happy\text{-}for}_i(j,\varphi), && \mathbf{resentment}_i(j,\varphi), \\ & \mathbf{gloating}_i(j,\varphi), && \mathbf{pity}_i(j,\varphi), \\ & \mathbf{hope}_i(\pi,\varphi), && \mathbf{fear}_i(\pi,\neg\varphi), \\ & \mathbf{satisfaction}_i(\pi,\varphi), && \mathbf{disappointment}_i(\pi,\varphi), \\ & \mathbf{relief}_i(\pi,\neg\varphi), && \mathbf{fears\text{-}confirmed}_i(\pi,\neg\varphi), \\ & \mathbf{love}_i(x), && \mathbf{hate}_i(x)\,\} \end{aligned} \tag{3.2}$$

where i and j are agent names ($i \neq j$), α is an atomic action, π is a plan (i.e. atomic actions and sequential compositions thereof), φ is a goal formula, and x is an agent name or object name. The definition of $ETrig$ should be read as follows: if $\mathbf{emotion}\,\bar{o}$ refers to, e.g., $\mathbf{joy}_i(\varphi)$, then $\mathbf{emotion}^{\mathbf{T}}\bar{o}$ is $\mathbf{joy}_i^{\mathbf{T}}(\varphi)$. The informal reading of, e.g., $\mathbf{joy}_i^{\mathbf{T}}(\varphi)$ is "joy is triggered for agent i with respect to event φ", whereas $\mathbf{joy}_i(\varphi)$ is read as "agent i experiences joy with respect to event φ." By convention, we write $\epsilon^{\mathbf{T}} \in ETrig$ and $\epsilon \in EExp$. With slight abuse of notation, the 'T' is also used to convert between $ETrig$ and $EExp$, e.g., if $\epsilon = \mathbf{joy}_i(\varphi)$ then $\epsilon^{\mathbf{T}} = \mathbf{joy}_i^{\mathbf{T}}(\varphi)$ in the same context.

Definition 1 (Agent specification language). *Let $\mathcal{P}$ be a set of atomic propositions, $\mathcal{A}$ a set of atomic actions, and $\mathcal{G}$ a set of agent names. Plans is the smallest set such that $\mathcal{A} \subseteq Plans$ and if $\alpha \in \mathcal{A}$ and $\pi \in Plans$ then $\alpha;\pi \in Plans$. The agent specification language $\mathcal{L}_{PAG}$ is the smallest set closed under:*

- $\{\bot, \top\} \cup \mathcal{P} \cup ETrig \cup EExp \subseteq \mathcal{L}_{PAG}$.
- *If* $\varphi_1, \varphi_2 \in \mathcal{L}_{PAG}$ *then* $\neg\varphi_1, (\varphi_1 \wedge \varphi_2) \in \mathcal{L}_{PAG}$.
- *If* $\varphi \in \mathcal{L}_{PAG}$, $i \in \mathcal{G}$ *then* $\mathbf{B}_i\varphi, \mathbf{G}_i\varphi \in \mathcal{L}_{PAG}$.[1]
- *If* $\pi \in Plans$, $\varphi \in \mathcal{L}_{PAG}$, $i \in \mathcal{G}$ *then* $\mathbf{A}_i\pi, \mathbf{Com}_i(\pi), [i:\pi]\varphi \in \mathcal{L}_{PAG}$.

With respect to the semantics of $\mathcal{L}_{PAG}$, the belief ($\mathbf{B}$) and action ($[\,\cdot\,]$) operators are modeled in a standard way using Kripke semantics, while using sets for goals, abilities,

[1] For simplicity, we deviate slightly from [9, 10] by allowing arbitrary goal formulas.

commitments, and emotional fluents. The modal logic KD45 is used for belief models, having the form $\mathrm{M} = \langle S, R, V \rangle$, where S is a set of states (or 'possible worlds'), R is a set of relations on S (one for each agent), and V is a valuation on S. The semantics of actions are defined *over* the Kripke models of belief, as actions may change the mental states of agents. Action models have the form $\mathcal{M} = \langle \mathcal{S}, \mathcal{R}, \mathrm{Emo}, \mathrm{Aux} \rangle$, where $\mathcal{S}$ is the set of possible model–state pairs (where models are of the form M as above and states are from S therein) and $\mathcal{R} = \{ \mathcal{R}_{i:\alpha} \mid i \in \mathcal{G},\ \alpha \in \mathcal{A} \}$ is a set of relations on $\mathcal{S}$ (one for each agent–action combination). $\mathcal{R}$ is required to produce a branching future and a single history, allowing its converse $\mathcal{R}^{-1}$ to be used as a history function. Notation: $(\mathrm{M}', s') \in \mathcal{R}_{i:\alpha}(\mathrm{M}, s)$ iff $\mathcal{R}^{-1}(\mathrm{M}', s') = \langle (\mathrm{M}, s), (i, \alpha) \rangle$.

$\mathrm{Emo} = \{ \mathit{Gratification}, \ldots, \mathit{Hate} \}$ is a set of 22 functions designed to define the semantics of the emotion triggering fluents [10]. $\mathrm{Aux} = \langle \Gamma, \mathcal{C}, \mathit{Agd}, T, \mathit{int} \rangle$ is a structure of auxiliary functions, where Γ is a function returning the set of goals an agent has per model–state pair; $\mathcal{C}$ is a function that returns the set of actions that an agent is capable of performing per model–state pair; Agd is a function that returns the set of actions that an agent is committed to (are on its 'agenda') per model–state pair. An external clock function $T : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is used (only) to calculate actual emotion intensity values. Therefore, it is assumed it assigns the same time to each state s of a belief model M, so we will (sloppily) express the time at a model–state pair (M, s) simply as T_{M}. It is also assumed that all actions take time, i.e., $\forall (\mathrm{M}', s') \in \mathcal{R}_{i:\alpha}(\mathrm{M}, s) : T_{\mathrm{M}'} > T_{\mathrm{M}}$.

We say that an emotion ϵ is *triggered* in state (M, s) iff $\mathrm{M}, s \models \epsilon^{\mathbf{T}}$ (see Definition 2). An emotion's triggering conditions often cease to hold when a transition to a next state is made, but emotional *experience* is supposed to be able to endure. Therefore, an *emotional memory* EMem is kept containing all newly triggered emotions (i.e. $\{ \epsilon \in \mathit{EExp} \mid \mathrm{M}, s \models \epsilon^{\mathbf{T}} \}$) plus all previously triggered ones ($\mathit{EMem}^*(\mathrm{M}', s')$). Thus, EMem makes previously triggered emotions available for assigning current intensities to them. We also define an extended emotional memory EMem_i^* as the union of all EMems in an agent's belief model, because it is desirable that emotional experience is the same in all belief states.

$$
\begin{aligned}
\mathit{EMem}(\mathrm{M}, s) &= \{ \epsilon \in \mathit{EExp} \mid \mathrm{M}, s \models \epsilon^{\mathbf{T}} \} \\
&\quad \cup \{ \epsilon \in \mathit{EMem}^*(\mathrm{M}', s') \mid \mathcal{R}^{-1}(\mathrm{M}, s) = \langle (\mathrm{M}', s'), \_ \rangle \} \\
\mathit{EMem}^*(\mathrm{M}, s) &= \textstyle\bigcup_{i \in \mathcal{G}} \mathit{EMem}_i^*(\mathrm{M}, s) \\
\mathit{EMem}_i^*(\mathrm{M}, s) &= \textstyle\bigcup_{s' \in \mathrm{GS}(R_i \cup R_i^{-1}, s)} \mathit{EMem}(\mathrm{M}, s')
\end{aligned}
$$

where $\mathrm{M} = \langle S, R, V \rangle$ and $\mathrm{GS}(R', s)$ returns all states in the *generated submodel* starting from s and following the relation R'. Thus, with slight abuse of notation, $\mathrm{GS}(R_i \cup R_i^{-1}, s)$ denotes all states reachable from s if R_i were an equivalence relation.

As previously stated, an emotion that is triggered is not necessarily *experienced*. To model emotional experience, the function $\mathit{int} : \mathcal{S} \to \mathit{EExp} \to \mathcal{I}$ is used, which assigns, per model–state pair, an intensity function to an emotional experience fluent. Each intensity function, as returned by $\mathit{int}(\mathrm{M}, s)(\epsilon)$, is a monotonically decreasing function of time. So $\mathcal{I}$ denotes the class of monotonically decreasing functions of type $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ (negative intensities are not allowed). As will be formally specified below, the idea is that an emotion ϵ that has been triggered now or in the past (i.e. $\epsilon \in \mathit{EMem}^*(\mathrm{M}, s)$) is currently experienced (i.e. in (M, s)) iff the intensity function f currently associated

with ϵ (i.e. $int(\mathrm{M}, s)(\epsilon) = f$) returns a value greater than zero for the current time (i.e. $f(T_{\mathrm{M}}) > 0$). It is desirable that emotional experience is the same in all belief states, so the function int is constrained as follows: $\forall(\mathrm{M}, s) \in \mathcal{S}, i \in \mathcal{G}, s' \in \mathrm{GS}(R_i, s) : int(\mathrm{M}, s) = int(\mathrm{M}, s')$. It is also assumed for all $(\mathrm{M}, s) \in \mathcal{S}$ and $\epsilon \in \mathit{EExp}$ that $int(\mathrm{M}, s)(\epsilon) = f_0$ if $\epsilon \notin \mathit{EMem}^*(\mathrm{M}, s)$, where $f_0(x) = 0$ for all x. It is crucial to note that the main idea of assigning intensity *functions* to triggered emotions (as opposed to directly assigning intensity *values*) is that we can then say that 'usually', performing an action does not change the way emotion intensities behave, i.e., for 'most' α, $(\mathrm{M}', s') \in \mathcal{R}_{i:\alpha}(\mathrm{M}, s)$ implies $int(\mathrm{M}', s') = int(\mathrm{M}, s)$. For certain actions we can then put useful constraints on int such that these actions can be said to influence emotions by changing the intensity functions that are assigned to these emotions. Indeed, this is exactly what we will do in section 4.

Definition 2. *(Interpretation of formulas).*
Let $\mathrm{M} = \langle S, R, V\rangle$ *and* $\mathcal{M} = \langle \mathcal{S}, \mathcal{R}, \mathrm{Emo}, \mathrm{Aux}\rangle$ *be structures defined as above. Formulas in language* $\mathcal{L}_{\mathcal{PAG}}$ *are interpreted in model–state pairs as follows:*

$$
\begin{array}{lcl}
\mathrm{M}, s \models p & \Leftrightarrow & p \in V(s) \quad \textit{for } p \in \mathcal{P} \\
\mathrm{M}, s \models \neg\varphi & \Leftrightarrow & \mathrm{M}, s \not\models \varphi \\
\mathrm{M}, s \models \varphi_1 \wedge \varphi_2 & \Leftrightarrow & \mathrm{M}, s \models \varphi_1 \ \&\ \mathrm{M}, s \models \varphi_2 \\
\mathrm{M}, s \models \mathbf{B}_i\varphi & \Leftrightarrow & \forall s' \in R_i(s) : \mathrm{M}, s' \models \varphi \\
\mathrm{M}, s \models \mathbf{G}_i\varphi & \Leftrightarrow & \varphi \in \Gamma(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{A}_i\pi & \Leftrightarrow & \pi \in \mathcal{C}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{Com}_i(\pi) & \Leftrightarrow & \pi \in \mathit{Agd}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models [i\!:\!\pi]\varphi & \Leftrightarrow & \forall(\mathrm{M}', s') \in \mathcal{R}_{i:\pi}(\mathrm{M}, s) : \mathrm{M}', s' \models \varphi \\
\mathrm{M}, s \models \mathbf{emotion}^{\mathbf{T}}_i\bar{o} & \Leftrightarrow & \bar{o} \in \mathit{Emotion}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{emotion}_i\bar{o} & \Leftrightarrow & \mathbf{emotion}_i\bar{o} \in \mathit{EMem}^*_i(\mathrm{M}, s)\ \& \\
 & & \quad int(\mathrm{M}, s)(\mathbf{emotion}_i\bar{o})(T_{\mathrm{M}}) > 0
\end{array}
$$

The last two lines abbreviate 2×22 lines; e.g., $\mathrm{M}, s \models \mathbf{joy}^{\mathbf{T}}_i(\varphi) \Leftrightarrow \varphi \in \mathit{Joy}(i)(\mathrm{M}, s)$ and $\mathrm{M}, s \models \mathbf{joy}_i(\varphi) \Leftrightarrow \mathbf{joy}_i(\varphi) \in \mathit{EMem}^*_i(\mathrm{M}, s)\ \&\ int(\mathrm{M}, s)(\mathbf{joy}_i(\varphi))(T_{\mathrm{M}}) > 0$.

Whenever we focus on a single-agent situation, we omit agent indices to ease notation. As explained before, we will write $\epsilon^{\mathbf{T}}$ for $\mathbf{emotion}^{\mathbf{T}}_i\bar{o}$ and ϵ for $\mathbf{emotion}_i\bar{o}$ henceforth. Finally, we use the following abbreviations that are common in KARO:
$\mathbf{P}(\alpha, \varphi) \equiv \mathbf{A}\alpha \wedge \langle\alpha\rangle\varphi$: An agent has the *practical possibility* to perform an action/plan α to bring about φ iff it has the ability to perform α and doing so can bring about φ.
$\mathbf{Can}(\alpha, \varphi) \equiv \mathbf{BP}(\alpha, \varphi)$: An agent *can* perform α to bring about φ iff it believes it has the practical possibility to do so.
$\mathbf{I}(\alpha, \varphi) \equiv \mathbf{Can}(\alpha, \varphi) \wedge \mathbf{BG}\varphi$: An agent has the *possible intention* to perform α to accomplish φ iff it *can* do so and it believes φ is one of its goals.

4 Emotion-based Action Tendency Formalized

In this section, we add a formal notion of action tendency. We write $\mathbf{T}_i(\alpha, \epsilon)$, meaning that agent i has the tendency to perform action α due to negative emotion ϵ:

$$\mathrm{M}, s \models \mathbf{T}_i(\alpha, \epsilon) \quad \Leftrightarrow \quad (\alpha, \epsilon) \in \mathit{RelTen}(i)(\mathrm{M}, s)$$

where the function $RelTen : \mathcal{G} \times \mathcal{S} \rightarrow \wp(\mathcal{A} \times EExp)$ formalizes (relative) action tendency as follows. We examine each negative emotion in the emotional memory of an agent; if there exists an action of which it believes that it is capable of performing the action and doing so may result in a state in which the emotion has strictly less intensity, then it tends to perform this action. Formally, this is written as:

$$RelTen(i)(\mathrm{M}, s) = \{\, (\alpha, \epsilon) \mid \epsilon \in EMem_i^*(\mathrm{M}, s),\ neg(\epsilon),\ \forall s' \in R_i(s) : [\alpha \in \mathcal{C}(i)(\mathrm{M}, s')\ \&\ \exists(\mathrm{M}', s'') \in \mathcal{R}_{i:\alpha}(\mathrm{M}, s') : int(\mathrm{M}', s'')(\epsilon)(T_{\mathrm{M}'}) < int(\mathrm{M}, s)(\epsilon)(T_{\mathrm{M}})] \,\}$$

where $\mathrm{M} = \langle S, R, V \rangle$ and $neg(\epsilon)$ is true iff ϵ is in the right-hand column of formula (3.2). It should be noted that, although there is no consensual definition of action tendency in the literature, our formalization is general enough to capture the basic concept.

With these definitions, we obtain the following propositions:

$$\not\models \epsilon^{\mathbf{T}} \rightarrow \epsilon \tag{4.1}$$

$$\not\models \epsilon \rightarrow \epsilon^{\mathbf{T}} \tag{4.2}$$

$$\models \epsilon \leftrightarrow \mathbf{B}\epsilon \tag{4.3}$$

$$\models \epsilon \wedge \mathbf{Can}(\alpha, \neg\epsilon) \rightarrow \mathbf{T}(\alpha, \epsilon) \qquad \text{(for negative emotions)} \tag{4.4}$$

$$\models \mathbf{T}(\alpha, \epsilon) \rightarrow \epsilon \wedge \mathbf{Can}(\alpha, \top) \tag{4.5}$$

The first proposition (which abbreviates, e.g., $\not\models \mathbf{joy}_i^{\mathbf{T}}(\varphi) \rightarrow \mathbf{joy}_i(\varphi)$) states that a newly triggered emotion is not necessarily experienced, whereas the second proposition (e.g., $\not\models \mathbf{joy}_i(\varphi) \rightarrow \mathbf{joy}_i^{\mathbf{T}}(\varphi)$) means that an emotion that is currently experienced is not necessarily a newly triggered one (as it may have been triggered in the past). The third proposition states that an agent is 'conscious' of each emotion it experiences (but not of emotion triggering, i.e. $\epsilon^{\mathbf{T}} \rightarrow \mathbf{B}\epsilon^{\mathbf{T}}$ is *not* valid). The fourth proposition states that an agent has the tendency to perform action α for emotion ϵ if ϵ is a currently experienced negative emotion and it 'can' perform α to mitigate the experience of ϵ. Note that the antecedent is slightly stronger (hence no bi-implication) than action tendency as formalized by $RelTen$, as $RelTen$ does not require the intensity of ϵ to become zero, but just less than before. The fifth proposition states that action tendency $\mathbf{T}(\alpha, \epsilon)$ requires that ϵ be currently experienced and that α 'can' be performed.

Next we show how, for certain types of actions, constraints can be put on the framework, such that specific emotions can be shown to lead to action tendencies.

4.1 Idling

Consider the most basic emotion regulation strategy: letting feelings subside by themselves. Since time is supposed to "heal all wounds" (and negative emotions in particular), the presented formalization of action tendency should straightforwardly capture tendency towards idling (e.g., 'count till ten before acting when feeling angry').

Let `idle` denote the action that has no effects other than the passage of time. (Here it does not matter how long an agent will actually be idling.) It is not unreasonable to assume that an agent always has the *practical possibility* of idling, i.e., for all states $(\mathrm{M}, s) \in \mathcal{S}$ and agents $i \in \mathcal{G}$ we have that $\texttt{idle} \in \mathcal{C}(i)(\mathrm{M}, s)$ and $\mathcal{R}_{i:\texttt{idle}}(\mathrm{M}, s) \neq \emptyset$. Furthermore, let int be constrained such that performing `idle` does not cause any changes in intensity functions:

$$\forall(\mathrm{M}', s') \in \mathcal{R}_{i:\mathtt{idle}}(\mathrm{M}, s) : \forall \epsilon \in \mathit{EMem}_i^*(\mathrm{M}, s) : \mathit{int}(\mathrm{M}', s')(\epsilon) = \mathit{int}(\mathrm{M}, s)(\epsilon)$$

Now we have that for all negative emotions ϵ: if for all (M, s), $\mathit{int}(\mathrm{M}, s)(\epsilon)$ is *strictly*[2] monotonically decreasing, then:

$$\models \epsilon \leftrightarrow \mathbf{T}(\mathtt{idle}, \epsilon) \tag{4.6}$$

Obviously, idling is only a valid strategy for negative emotions whose intensities actually decrease over time, hence the added requirement of strict monotonicity for this proposition.

4.2 Social Tendencies

In the following, it should be kept in mind that in the framework of [10], pity or resentment can be triggered when an agent believes that another agent's (sub)goal has been undermined (undone) or accomplished, respectively, and the agent views this as undesirable for itself. The (sub)goal that has been undermined or accomplished is denoted as φ, as in $\mathbf{pity}_i^{\mathbf{T}}(j, \varphi)$. Moreover, gratitude or anger can be triggered when a (sub)goal has been accomplished or undermined, respectively, by an action of another agent.

A reasonable constraint on int would now be to require that the intensity of a pity emotion is decreased if the agent believes that the action it performs re-accomplishes the (sub)goal of the other agent that it pitied, which is exactly the case when gratitude is triggered in the other agent. Formally, we constrain int such that for all states $(\mathrm{M}, s) \in \mathcal{S}$, agents $i \in \mathcal{G}$, actions $\alpha \in \mathcal{A}$, and all $(\mathrm{M}', s') \in \mathcal{R}_{i:\alpha}(\mathrm{M}, s)$:

$$\mathit{int}(\mathrm{M}', s')(\mathbf{pity}_i(j, \varphi)) = \begin{cases} f_0 & \text{if } \mathrm{M}', s' \models \mathbf{B}_i \mathbf{gratitude}_j^{\mathbf{T}}(i, \alpha, \varphi) \\ \mathit{int}(\mathrm{M}, s)(\mathbf{pity}_i(j, \varphi)) & \text{otherwise.} \end{cases}$$

Since the same reasoning applies to resentment and anger, we repeat the above with $\mathbf{pity}_i(j, \varphi)$ / $\mathbf{gratitude}_j^{\mathbf{T}}(i, \alpha, \varphi)$ replaced by $\mathbf{resentment}_i(j, \varphi)$ / $\mathbf{anger}_j^{\mathbf{T}}(i, \alpha, \varphi)$. Furthermore, this constraint for resentment and anger can be made more generic by dropping the φ, so we add a third constraint as above with $\mathbf{hate}_i(j)$ / $\mathbf{reproach}_j^{\mathbf{T}}(i, \alpha)$.

With these constraints the following propositions are valid:

$$\models \mathbf{pity}_i(j, \varphi) \wedge \mathbf{Can}_i(\alpha, \mathbf{B}_i \mathbf{gratitude}_j^{\mathbf{T}}(i, \alpha, \varphi)) \rightarrow \mathbf{T}_i(\alpha, \mathbf{pity}_i(j, \varphi)) \tag{4.7}$$

$$\models \mathbf{resentment}_i(j, \varphi) \wedge \mathbf{Can}_i(\alpha, \mathbf{B}_i \mathbf{anger}_j^{\mathbf{T}}(i, \alpha, \varphi)) \rightarrow \mathbf{T}_i(\alpha, \mathbf{resentment}_i(j, \varphi)) \tag{4.8}$$

$$\models \mathbf{hate}_i(j) \wedge \mathbf{Can}_i(\alpha, \mathbf{B}_i \mathbf{reproach}_j^{\mathbf{T}}(i, \alpha)) \rightarrow \mathbf{T}_i(\alpha, \mathbf{hate}_i(j)) \tag{4.9}$$

The first proposition states that if an agent pities another agent because its (sub)goal φ has been undermined, then it has the tendency to perform any action with which it can trigger, in the other agent, gratitude towards itself with respect to φ. Similar readings apply to the other propositions. It should be noted that the third proposition is like the second proposition but without relating α to a goal φ. Applying the same generalization to

[2] Intensity functions are supposed to reach zero within a finite amount of time but not to return negative values, so "strictly" should in this case be interpreted as: $\mathit{int}(\mathrm{M}, s)(\epsilon) = f$ where $f(x) = \max(g(x), 0)$ and g is (truly) strictly monotonically decreasing.

the first proposition yields the formula $\mathbf{love}_i(j) \wedge \mathbf{Can}_i(\alpha, \mathbf{B}_i\mathbf{admiration}^{\mathbf{T}}_j(i,\alpha)) \rightarrow \mathbf{T}_i(\alpha, \mathbf{love}_i(j))$. The problem with this formula is that it expresses action tendencies resulting from positive emotions (i.e. love), whereas action tendency so far has only been defined with respect to negative emotions. Indeed, with a suitable definition of 'positive' action tendency, this proposition will be worth investigating.

4.3 Reconsideration

In this section, we investigate a specific *class* of actions, namely reconsideration actions. For actions of this type, we formulate specific constraints and study their implications.

According to [9], hope to accomplish a goal by performing a plan is triggered when an agent has the possible intention to perform the plan for the goal and is committed to the plan, i.e., $\mathbf{I}(\pi,\varphi) \wedge \mathbf{Com}(\pi) \rightarrow \mathbf{hope}^{\mathbf{T}}(\pi,\varphi)$. Conversely, fear is triggered when the agent believes the plan may fail to accomplish the goal hoped for, i.e., $\mathbf{hope}^{\mathbf{T}}(\pi,\varphi) \wedge \mathbf{B}\langle\pi\rangle\neg\varphi \rightarrow \mathbf{fear}^{\mathbf{T}}(\pi,\neg\varphi)$. Note that these are propositions that are provable given suitable definitions of the functions $\mathit{Hope}, \mathit{Fear} \in \mathrm{Emo}$ (see section 3).

Positive emotions are often associated with a tendency to proceed as planned, whereas negative emotions can cause reconsideration or trying harder [16]. Viewed from this perspective, an agent should have the tendency to reconsider if it fears a current (possible) intention may fail to accomplish its goal. We will first specify in more detail what we mean by reconsidering, and then show how this tendency can be modeled.

More specifically, reconsideration can be done when an agent has the (possible) intention $\mathbf{I}(\pi,\varphi)$ to perform some plan π for some goal φ, with the result that π or φ is dropped and possibly a new plan towards φ is found. It is assumed there exists a set of special reconsideration actions. An example is *uncommitting* (intention reconsideration), i.e., the special action $\mathtt{uncommit}(\pi)$. (It is special in the sense that it is not supposed to be nested [20].) The result of this action is that π is removed from the agent's agenda. Another example is *replanning* (plan reconsideration), i.e. $\mathtt{replan}(\pi,\varphi,\pi')$, which generates a (new) plan π' to accomplish goal φ and replaces π for π' on the agent's agenda. Alternatively, the agent can *drop* its goal φ or substitute it for another (similar) goal (goal reconsideration). In the following, we will write a reconsider action as $\mathit{reconsider}(\pi,\varphi)$. Each formula containing this expression should actually be viewed as a set of formulas; one for each instance of reconsider action.

As a first constraint we specify that an agent has the *practical possibility* of reconsidering a plan toward a goal if it has the possible intention to perform the plan for the goal:

$$\mathrm{M}, s \models \mathbf{I}(\pi,\varphi) \Rightarrow \mathit{reconsider}(\pi,\varphi) \in \mathcal{C}(\mathrm{M}, s) \;\&\; \mathcal{R}_{\mathit{reconsider}(\pi,\varphi)}(\mathrm{M}, s) \neq \emptyset$$

Next, we constrain int such that reconsidering causes all hope and fear with respect to the old plan to be assigned an intensity function that always returns zero (i.e. f_0). So we specify: if $\epsilon \in \{\mathbf{hope}(\pi,\varphi), \mathbf{fear}(\pi,\neg\varphi)\}$, then $\forall(\mathrm{M}, s) \in \mathcal{S} : \forall(\mathrm{M}', s') \in \mathcal{R}_{\mathit{reconsider}(\pi,\varphi)}(\mathrm{M}, s) : \mathit{int}(\mathrm{M}', s')(\epsilon) = f_0$. The result of this is that any old fear (with respect to π possibly failing to accomplish φ) is guaranteed to be gone after reconsidering π for φ. With these constraints, the following proposition is valid:

$$\models \mathbf{I}(\pi,\varphi) \wedge \mathbf{fear}(\pi,\neg\varphi) \rightarrow \mathbf{T}(\mathit{reconsider}(\pi,\varphi), \mathbf{fear}(\pi,\neg\varphi)) \tag{4.10}$$

for any type of reconsider action. In other words, an agent has the tendency to reconsider a plan towards a goal if it has the possible intention to perform the plan for the goal but fears the plan may fail to accomplish the goal.

5 Discussion

Several noteworthy variations on the presented formalization of action tendency are possible. For example:

'Strict' action tendency The definition of $RelTen$ can be made stricter by replacing the inequality by $int(\mathrm{M}', s'')(\epsilon)(T_{\mathrm{M}'}) < int(\mathrm{M}, s)(\epsilon)(T_{\mathrm{M}'})$. The (subtle) difference is that both intensity functions are evaluated at time $T_{\mathrm{M}'}$, so that the function returned by $int(\mathrm{M}', s'')(\epsilon)$ must really be different *and* better than the old one $int(\mathrm{M}, s)(\epsilon)$.

'Long-term' action tendency The formalization of 'strict' action tendency can be made even stronger by adding: $\ldots \ \& \ \forall t > T_{\mathrm{M}'} : int(\mathrm{M}', s'')(\epsilon)(t) \leq int(\mathrm{M}, s)(\epsilon)(t)$. This means that the new intensity function $int(\mathrm{M}', s'')(\epsilon)$ for emotion ϵ must be better than the old one in all future time points (although it is possible that a future action replaces the intensity function for emotion ϵ again with a worse one).

'Overall' action tendency So far we have only dealt with action tendency *relative* to a single (negative) emotion. However, it is possible to formalize a kind of action tendency that takes into account *all* negative emotions of an agent and specifies that an agent tends to perform some action if the sum of the intensities of its negative emotions in the state after performing the action is less than this sum was before. So the inequality in the definition of $RelTen$ would then be replaced by $\sum_{\epsilon} int(\mathrm{M}', s'')(\epsilon)(T_{\mathrm{M}'}) < \sum_{\epsilon} int(\mathrm{M}, s)(\epsilon)(T_{\mathrm{M}})$ where the summations are over all negative emotions $\epsilon \in EMem_i^*(\mathrm{M}', s'')$ and $\epsilon \in EMem_i^*(\mathrm{M}, s)$, respectively.

It may be interesting to note that these alternative specifications of action tendency can be formally compared. Let the three kinds of action tendency as described above be formalized as $\mathbf{T^s}(\alpha, \epsilon)$, $\mathbf{T^l}(\alpha, \epsilon)$, and $\mathbf{T^o}\alpha$, respectively. Then:

- $\mathbf{T^l}(\alpha, \epsilon)$ implies $\mathbf{T^s}(\alpha, \epsilon)$ implies $\mathbf{T}(\alpha, \epsilon)$;
- $\mathbf{T^o}\alpha$ implies that there exists an ϵ such that $\mathbf{T}(\alpha, \epsilon)$.

It is easy to see that an agent would never 'strictly' tend to perform the action `idle` as above, i.e., $\models \neg\mathbf{T^s}(\texttt{idle}, \epsilon)$ for any ϵ. Moreover, the reconsideration result can be strengthened as follows. If it is assumed that the intensities of corresponding hope and fear emotions always sum to a constant,[3] that reconsidering does not change an agent's beliefs, and that any new plan or goal, if found, does not cause more new fear than new hope, then we would have that $\models \mathbf{I}(\pi, \varphi) \wedge \mathbf{hope}(\pi, \varphi) < \mathbf{fear}(\pi, \neg\varphi) \rightarrow \mathbf{T^o}\mathit{reconsider}(\pi, \varphi)$, where $<$ compares emotions by intensity. Indeed, as is evident from the previous sentence, there are usually a lot of constraints that have to be placed in order to attain an 'overall' tendency towards some (type of) action.

It should be emphasized that the presence of action tendencies still does not specify which action will actually be chosen by an agent. However, an ordering can easily be defined on the subset of actions that an agent tends to perform by comparing their 'gain', i.e., the difference in intensity that the agent believes to obtain by performing an action. Then one can determine which action an agent tends to perform most and is thus most likely to be selected. Moreover, properties such as $\mathbf{T}(\alpha_1, \epsilon) \wedge \neg\mathbf{T^s}(\alpha_1, \epsilon) \wedge \mathbf{T^s}(\alpha_2, \epsilon) \rightarrow \alpha_1 \prec \alpha_2$ can then be investigated, where $\alpha_1 \prec \alpha_2$ means that the agent in question strictly prefers to perform α_2 over α_1 to regulate emotion ϵ.

[3] This is not unreasonable; in fact, it is argued for by OCC [19]. Details on how to model this assumption can be found in [10].

6 Related Work

Meyer [6] formalized four basic emotion types (i.e., happiness, sadness, anger, and fear) inspired by the psychological work of Oatley & Jenkins [16]. In addition to formalizing their triggering conditions, a heuristic is associated with each emotion type, indicating how an agent should act on it. However, lacking a formalization of quantitative aspects, it is left unspecified how executing such a heuristic influences the experience of the associated emotion. Moreover, in our approach, any number of 'heuristics' can be defined; the action tendency operator will pick up on any action that can improve the situation.

Adam [8] proposed a purely qualitative formalization of the OCC model also incorporating emotion regulation. However, only the regulation of negative event-based emotions (i.e., distress, disappointment, fear, fears-confirmed, pity, and resentment) is investigated. To this end, seven coping strategies are defined. Some coping strategies (e.g., denial, resign) change the beliefs or desires of an agent such that the triggering conditions for the negative emotion cease to hold. Other coping strategies (e.g, mental disengagement, venting) lead to the adoption of intentions to bring about new positive emotions that "divert the individual from the current negative one" [8]. However, in contrast to our approach, quantitative aspects of emotions are not taken into account, so it is left unspecified how these coping strategies actually mitigate the experience of negative emotions. Moreover, it is not clear how Adam measures whether the situation after coping is better than before.

Gratch & Marsella [4] have been working on a computational framework for modeling emotions inspired by the OCC model, among others. An implementation, named EMA, is used for social training applications. Like Adam, their framework incorporates a number of coping strategies. However, in EMA, the link from appraisal to coping is rather direct. In contrast, we put a notion of action tendency in between emotions and regulatory actions, on which an agent can decide to act or not. Moreover, few formal details about the logic underlying EMA are provided.

7 Conclusion

In this paper, we have formalized a notion of action tendency and incorporated it into an existing formalization of emotions. This formalization is based on the psychological OCC model of emotions and grounded in the KARO framework of rational agency. However, so far it only incorporated the appraisal and experience phases of emotions. What is lacking in this (and several other) formalizations of emotions is a specification of how emotions influence behavior. In psychology, emotions are often seen as learned and innate heuristics that give an individual the tendency to prefer certain actions over others, based on its current emotional state. Thus by incorporating action tendency in a formal model of agency, we have introduced a mechanism for limiting and ordering options in an agent's action selection process. In line with the general view of action tendency in psychology, our formalized notion of action tendency is based on reducing negative emotion intensity. To this end, a goal may be adopted or reconsidered (as shown in section 4.3), but not necessarily so.

It should be noted that the presented formalization of action tendency is only defined with respect to negative emotions. Moreover, due to space limitations, only action

tendencies with respect to several of the emotion types in the formal emotion model have been presented. For future work, responses to positive emotions as well as the remaining negative emotions have to be investigated. Because there are several ways of formalizing action tendency, the implications of choosing any particular 'flavor' have to be explicated. As shown, it is also possible to incorporate multiple notions of action tendency and study their relations.

References

1. Picard, R.W.: Affective Computing. MIT Press (1997)
2. Johns, M., Silverman, B.G.: How emotion and personality effect the utility of alternative decisions: A terrorist target selection case study. In: 10th Conference On Computer Generated Forces and Behavioral Representation, SISO. (2001)
3. Sloman, A.: Beyond shallow models of emotion. Cognitive Processing **2**(1) (2001) 177–198
4. Gratch, J., Marsella, S.: A domain-independent framework for modeling emotions. Journal of Cognitive Systems Research **5**(4) (2004) 269–306
5. Marinier, R.P., Laird, J.E.: Toward a comprehensive computational model of emotions and feelings. In: Proceedings of the International Conference on Cognitive Modeling (ICCM'04), Pittsburgh, PA (2004) 172–177
6. Meyer, J.-J.Ch.: Reasoning about emotional agents. International Journal of Intelligent Systems **21**(6) (2006) 601–619
7. Dastani, M., Meyer, J.-J.Ch.: Programming agents with emotions. In: Proceedings of the 17th European Conference on Artificial Intelligence (ECAI'06). (2006) 215–219
8. Adam, C.: The Emotions: From Psychological Theories to Logical Formalization and Implementation in a BDI Agent. PhD thesis, Institut National Polytechnique de Toulouse (2007)
9. Steunebrink, B.R., Dastani, M., Meyer, J.-J.Ch.: A logic of emotions for intelligent agents. In: Proceedings of the 22nd Conference on Artificial Intelligence (AAAI'07), AAAI Press (2007)
10. Steunebrink, B.R., Dastani, M., Meyer, J.-J.Ch.: A formal model of emotions: Integrating qualitative and quantitative aspects. In: Proceedings of the 18th European Conference on Artificial Intelligence (ECAI'08), IOS Press (2008) 256–260
11. Elster, J.: Rationality and the emotions. Economic Journal **106**(438) (1996) 1386–1397
12. Coppin, G.: Emotion, personality and decision-making: Relying on the observables. In: Proceedings of the 3rd International Conference in Human Centered Processes (HCP-2008). (2008)
13. Frijda, N.H.: The Emotions. Studies in Emotion and Social Interaction. Cambridge University Press (1987)
14. LeDoux, J.E.: The Emotional Brain: Mysterious Underpinnings of Emotional Life. Simon & Schuster (1996)
15. Ekman, P., Davidson, R.J., eds.: The Nature of Emotion: Fundamental Questions. Series in Affective Science. Oxford University Press (1994)
16. Oatley, K., Jenkins, J.M.: Understanding Emotions. Blackwell Publishing, Oxford (1996)
17. Damasio, A.R.: Descartes' Error: Emotion, Reason and the Human Brain. Grosset/Putnam, New York (1994)
18. Gross, J.J., Thompson, R.A.: Emotion regulation: Conceptual foundations. In Gross, J.J., ed.: Handbook of Emotion Regulation. Guilford Press (2007)
19. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, UK (1988)
20. Meyer, J.-J.Ch., Hoek, W.v.d., Linder, B.v.: A logical approach to the dynamics of commitments. Artificial Intelligence **113** (1999) 1–40