A Logic of Emotions for Intelligent Agents

Bas R. Steunebrink
Department of ICS
Utrecht University
Utrecht, The Netherlands
bass@cs.uu.nl

Mehdi Dastani
Department of ICS
Utrecht University
Utrecht, The Netherlands
mehdi@cs.uu.nl

John-Jules Ch. Meyer
Department of ICS
Utrecht University
Utrecht, The Netherlands
jj@cs.uu.nl

Abstract

This paper formalizes a well-known psychological model of emotions in an agent specification language. This is done by introducing a logical language and its semantics that are used to specify an agent model in terms of mental attitudes including emotions. We show that our formalization renders a number of intuitive and plausible properties of emotions. We also show how this formalization can be used to specify the effect of emotions on an agent's decision making process. Ultimately, the emotions in this model function as heuristics as they constrain an agent's model.

Introduction

In psychological studies, the emotions that influence the deliberation and practical reasoning of an agent are considered as heuristics for preventing excessive deliberation (Damasio 1994). Meyer & Dastani (2004; 2006) propose a functional approach to describe the role of emotions in practical reasoning. According to this functional approach, an agent is assumed to execute domain actions in order to reach its goals. The effects of these domain actions cause and/or influence the appraisal of emotions according to a human-inspired model. These emotions in turn influence the deliberation operations of the agent, functioning as heuristics for determining which domain actions have to be chosen next, which completes the circle.

Although logics for modeling the behavior of intelligent agents are in abundance, the effect of emotions on rational behavior is usually not considered, despite of their (arguably positive) contribution. Philosophical studies describing (idealized) human behavior have previously been formalized using one or more logics (often mixed or extended). For example, Bratman's BDI theory of belief, desire, and intentions (Bratman 1987) has been modeled and studied in e.g. linear time logic (Cohen & Levesque 1990) and dynamic logic (Meyer, Hoek, & Linder 1999).

We propose to model and formalize human emotions in logic. There exist different psychological models of emotions, of which we have chosen to consider the model of Ortony, Clore, & Collins (1988). The "OCC model" is suitable for formalization because it describes a concise hierarchy of emotions and specifies the conditions that elicit each

emotion in terms of objects, actions, and events—concepts that can be captured in a formal language. In this paper, we introduce a logic for studying the appraisal, interactions, and effects of the 22 emotions described in the OCC model. We take a computational approach, building not only a mathematically sound model but also keeping in mind its implementability in a (multi-)agent system. Multi-agent aspects of emotions, however, are not treated in this paper.

It should be noted that previous work on specifying and implementing emotions carried out by Meyer (2004) and Dastani (2006) follows Oatley & Jenkins' model of emotions (Oatley & Jenkins 1996) and comprises only four emotions: *happy*, *sad*, *angry*, and *fearful*. Emotions are represented as *labels* in an agent's cognitive state. Similar to our approach, the deliberation of an agent causes the appraisal of emotions that in turn influence the agent's deliberation. Dastani & Meyer (2006) have defined transition semantics for their emotional model, which we also intend to do for our formalization of OCC. However, we intend to formalize the quantitative aspects of emotions as well, which were not considered in the purely logical model of Dastani & Meyer. Our work is also similar to other computational models of emotions, such as EMA (Gratch & Marsella 2004), CogAff (Sloman 2001), and the work of Picard (1997); however, our goal is not to develop a specific computational model of emotions, but rather to develop a logic for studying emotional models, starting with the OCC model.

Language and Semantics

The OCC model describes a hierarchy that classifies 22 emotions. The hierarchy contains three branches, namely emotions concerning aspects of objects (e.g., love and hate), actions of agents (e.g., pride and admiration), and consequences of events (e.g., joy and pity). Additionally, some branches combine to form a group of compound emotions, namely emotions concerning consequences of events *caused* by actions of agents (e.g., gratitude and anger). Because the objects of all these emotions (i.e. objects, actions, and events) correspond to notions commonly used in agent models (i.e. agents, plans, and goal accomplishments, respectively), this makes the OCC model suitable for use in the deliberation and practical reasoning of artificial agents. It should be emphasized that emotions are not used to describe the entire cognitive state of an agent (as in "the agent is

wholly happy"); rather, emotions are always relative to individual objects, actions, and events, so an agent can be joyous about event X and at the same time distressed about event Y.

The OCC model defines both qualitative and quantitative aspects of emotions. Qualitatively, it defines the conditions that *elicit* each of the emotions; quantitatively, it describes how a potential, threshold, and intensity are associated with each elicited emotion and what are the variables affecting these quantities. 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." The variables affecting its (quantitative) intensity are 1) the judged praiseworthiness of the action, 2) the unexpectedness of the event, and 3) the desirability of the event.

We use KARO (Meyer 2004; Meyer, Hoek, & Linder 1999) as a framework for the formalization of the 22 emotions of the OCC model. The KARO framework is a mixture of dynamic logic, epistemic / doxastic logic, and several additional (modal) BDI operators for dealing with the motivational aspects of artificial agents. We present a modest extension of the KARO framework, so that the eliciting conditions of the emotions of the OCC model can be appropriately translated and modeled. However, a discussion of the exact differences is outside the scope of this paper. We have completed the formalization of a qualitative model of the 22 emotions in the OCC model, but because of space limitations we cannot present this entire formalization here. Instead, we will focus on *hope* and *fear* alone in this paper. We have picked these two emotions because of their pivotal role in reducing nondeterminism in agent implementations.

The KARO framework is designed to specify goal-directed agents; however, in contrast to KARO, we do not allow arbitrary formulas as (declarative) goals and define a goal as a conjunction of literals, where each literal represents a subgoal. This is because we want agents to be able to break up their goals into subgoals to determine which parts of a goal have already been achieved and which subgoals have yet to be pursued. Furthermore, we require goals to be consistent (individually, not mutually) and non-empty (i.e. there must be at least one literal in the conjunction).

Definition 1. *(Consistent conjunctions). Let $\mathcal{P}$ be a set of atomic propositions and $Lits = \mathcal{P} \cup \{ \neg p \mid p \in \mathcal{P} \}$ be the set of literals. With respect to the conjunction and disjunction of the empty set, let $\bigwedge \emptyset = \top$ and $\bigvee \emptyset = \bot$, where $\bot$ stands for falsum and $\top$ for verum. Then $\mathcal{K}$ is the set of all consistent conjunctions of literals, and $\mathcal{K}'$ does not contain the empty conjunction:*

$$\mathcal{K} = \{ \bigwedge \Phi \mid \Phi \subseteq Lits, \ \Phi \nvdash \bot \}, \qquad (1)$$

$$\mathcal{K}' = \mathcal{K} \setminus \{\top\} \qquad (2)$$

Thus the empty conjunction is denoted by $\top$, and note that $\top \in \mathcal{K}$ whereas $\top \notin \mathcal{K}'$.

Below we explain how 'OCC ingredients' are translated into 'KARO ingredients.' When formalizing the branch (of the OCC hierarchy) of emotions concerning consequences of events, we will translate OCC's notion of an event as the accomplishment or undermining of a goal (or part thereof), because these are the kinds of events telling an agent how its goals and plans toward them are progressing. For example, the failure to achieve certain subgoals during the execution of a plan may cause the appraisal of *fear*, which, consequently, might trigger the agent to revise its plan.

When formalizing the branch (of the OCC hierarchy) of emotions concerning actions of agents, we will translate OCC's notion of actions as plans consisting of domain actions and sequential compositions of actions. Note that besides domain actions that can be performed by agents, we also distinguish deliberation operations (e.g., operations for selecting and applying planning rules and for selecting plans to execute) as actions that can be performed by agents.

Definition 2. *(Plans). Let $\mathcal{A}$ be a set of atomic domain actions. The set $Plans$ of plans consists of all actions and sequential compositions of actions. It is the smallest set closed under:*

- *If $\alpha \in \mathcal{A}$ then $\alpha \in Plans$.*
- *If $\alpha \in \mathcal{A}$ and $\pi \in Plans$ then $(\alpha; \pi) \in Plans$.*

When formalizing the branch (of the OCC hierarchy) of emotions concerning objects, we only consider agents as objects, because there are no other notions in our framework that could reasonably be regarded as objects.

We define an *emotional fluent* for each of the 22 emotions of the OCC model. The emotions are outlined below such that each row contains two emotions that are defined by OCC to be each other's opposites, with the left column displaying the positive emotions and the right column displaying the negative emotions (for agent i). It should be noted that it is allowed for an agent to have 'mixed feelings,' i.e. it can experience opposing emotions simultaneously. However, our model will ensure that the objects of opposing emotions are distinct (e.g., an agent can experience both joy and distress in response to some event, but the objects of these two emotions will concern different parts of the event).

Definition 3. *(Emotional fluents). Let $\mathcal{G}$ be a set of agent names. The set $Emotions$ is the set of emotional fluents, which is defined as follows:*

$$
\begin{aligned}
Emotions = \\
\{ \ &\mathbf{joy}_i(\kappa), & &\mathbf{distress}_i(\kappa), \\
&\mathbf{hope}_i(\pi, \kappa), & &\mathbf{fear}_i(\pi, \neg\kappa), \\
&\mathbf{satisfaction}_i(\pi, \kappa), & &\mathbf{disappointment}_i(\pi, \kappa), \\
&\mathbf{relief}_i(\pi, \neg\kappa), & &\mathbf{fears\text{-}confirmed}_i(\pi, \neg\kappa), \\
&\mathbf{happy\text{-}for}_i(j, \kappa), & &\mathbf{resentment}_i(j, \kappa), \\
&\mathbf{gloating}_i(j, \kappa), & &\mathbf{pity}_i(j, \kappa), \\
&\mathbf{pride}_i(\alpha), & &\mathbf{shame}_i(\alpha), \\
&\mathbf{admiration}_i(j, \alpha), & &\mathbf{reproach}_i(j, \alpha), \\
&\mathbf{love}_i(j), & &\mathbf{hate}_i(j), \\
&\mathbf{gratification}_i(\alpha, \kappa), & &\mathbf{remorse}_i(\alpha, \kappa), \\
&\mathbf{gratitude}_i(j, \alpha, \kappa), & &\mathbf{anger}_i(j, \alpha, \kappa) \\
\mid \ &i, j \in \mathcal{G}, \ i \neq j, \ \alpha \in \mathcal{A}, \ \pi \in Plans, \ \kappa \in \mathcal{K}' \}
\end{aligned}
$$

(3)

The informal reading of the emotional fluents used in this paper is as follows: $\mathbf{hope}_i(\pi, \kappa)$ means agent i hopes performing plan π will accomplish goal κ; $\mathbf{fear}_i(\pi, \neg\kappa)$ means agent i fears performing plan π will not accomplish goal κ.

We now have all ingredients necessary to modify the KARO framework and construct an agent specification language. This language contains operators for belief ($\mathbf{B}$), goals ($\mathbf{G}$), (cap)ability ($\mathbf{A}$), commitment ($\mathbf{Com}$), and action ($\mathbf{do}$).

Definition 4. *(Language). Let the sets $\mathcal{P}$, $\mathcal{K}'$, Plans, $\mathcal{G}$, and Emotions be defined as above. The agent specification language $\mathcal{L}$ is the smallest set closed under:*

- *If $p \in \mathcal{P}$ then $p \in \mathcal{L}$.*
- *If $\varphi_1, \varphi_2 \in \mathcal{L}$ then $\neg\varphi_1, (\varphi_1 \wedge \varphi_2) \in \mathcal{L}$.*
- *If $\varphi \in \mathcal{L}$ and $i \in \mathcal{G}$ then $\mathbf{B}_i\varphi \in \mathcal{L}$.*
- *If $\kappa \in \mathcal{K}'$ and $i \in \mathcal{G}$ then $\mathbf{G}_i\kappa \in \mathcal{L}$.*
- *If $\pi \in$ Plans and $i \in \mathcal{G}$ then $\mathbf{A}_i\pi, \mathbf{Com}_i(\pi) \in \mathcal{L}$.*
- *If $\pi \in$ Plans and $\varphi \in \mathcal{L}$ and $i \in \mathcal{G}$ then $[\mathbf{do}_i(\pi)]\varphi \in \mathcal{L}$.*
- *If $\epsilon \in$ Emotions then $\epsilon \in \mathcal{L}$.*

We also use the propositional connectives $\vee$, $\rightarrow$, and $\leftrightarrow$ with their usual interpretation. $\mathbf{B}_i\varphi$ means agent i believes in φ; $\mathbf{G}_i\kappa$ means agent i has the (declarative) goal to accomplish κ; $\mathbf{A}_i\pi$ means agent i has the ability to perform π; $\mathbf{Com}_i(\pi)$ means agent i is committed to performing π; $[\mathbf{do}_i(\pi)]\varphi$ means φ holds after agent i has performed π. For convenience, subscript agent indices (e.g., i and j) are omitted if the formula in question concerns only a single agent. We use $\langle \cdot \rangle$ as the dual of $[\cdot]$ for the $\mathbf{do}$ operator. We denote the execution of the deliberation operations as $[\mathbf{do}(deliberate)]$ (details are given later in this paper).

With respect to the semantics of $\mathcal{L}$, we model the belief and action operators using Kripke semantics, while using sets for ability, commitment, goals, and emotional fluents. The semantics of actions are defined over the Kripke models of belief, as actions may change the mental state of an agent.

Definition 5. *(Semantics). Let the sets $\mathcal{P}$, $\mathcal{K}'$, $\mathcal{A}$, Plans, and $\mathcal{G}$ be defined as above. The semantics of the belief and action operators are given by Kripke structures of the form $\mathrm{M} = \langle S, \vartheta, R_{\mathbf{B}} \rangle$ and $\langle \Sigma, R_{\mathcal{A}} \rangle$, respectively, where*

- *S is a non-empty set of states (or worlds);*
- *$\vartheta : S \rightarrow \wp(\mathcal{P})$ is a truth assignment function per state;*
- *$R_{\mathbf{B}} : \mathcal{G} \times S \rightarrow \wp(S)$ is an accessibility relation on S for the belief modality of an agent. $R_{\mathbf{B}}$ is assumed to be serial, transitive, and euclidean;*
- *Σ is the set of possible model–state pairs. A model–state pair is denoted as (M, s), where $\mathrm{M} = \langle S, \vartheta, R_{\mathbf{B}} \rangle$ as above and $s \in S$;*
- *$R_{\mathcal{A}} : \mathcal{G} \times$ Plans $\times \Sigma \rightarrow \wp(\Sigma)$ is an accessibility relation on Σ, encoding the behavior of actions of an agent. $R_{\mathcal{A}}(i, \pi)$ (for $\pi \in$ Plans) is defined as usual in dynamic logic by induction from a given base case $R_{\mathcal{A}}(i, \alpha)$ (for $\alpha \in \mathcal{A}$), i.e. $R_{\mathcal{A}}(i, \alpha; \pi) = R_{\mathcal{A}}(i, \alpha) \bullet R_{\mathcal{A}}(i, \pi)$.*

The semantics of ability, commitment, goals, and emotions are given by means of structures of type $\langle \mathcal{C}, Ag, \Gamma, E \rangle$, where

- *$\mathcal{C} : \mathcal{G} \times \Sigma \rightarrow \wp(Plans)$ is a function that returns the set of actions that an agent is capable of performing per model–state pair;*

- *$Ag : \mathcal{G} \times \Sigma \rightarrow \wp(Plans)$ is a function that returns the set of plans that an agent is committed to (are on an agent's 'agenda') per model–state pair;*
- *$\Gamma : \mathcal{G} \times \Sigma \rightarrow \wp(\mathcal{K}')$ is a function that returns the set of goals that an agent has per model–state pair;*
- *$E = \langle Joy, Distress, Hope, Fear, \ldots, Anger \rangle$ is a structure of 22 functions indicating per model–state pair which emotions are being experienced by an agent.*

Note that *Hope* and *Fear* are semantic functions designed to define the semantics of the syntactic emotional fluents $\mathbf{hope}$ and $\mathbf{fear}$. It is *crucial* to note that the functions in E are constrained by the emotion axioms that we define according to the OCC model, i.e. formulas (6) and (7) in this paper. Because we will only be treating hope and fear in this paper, we will only define the semantics, interpretation, and emotion axioms of these two emotions. The emotion functions in E have the following types:

$$Hope : \mathcal{G} \times \Sigma \rightarrow \wp(Plans \times \mathcal{K}')$$
$$Fear : \mathcal{G} \times \Sigma \rightarrow \wp(Plans \times \mathcal{K}^{\neg})$$
$$\vdots$$

where $\mathcal{K}^{\neg} = \{ \neg\kappa \mid \kappa \in \mathcal{K}' \}$. They have to be defined per agent ($\mathcal{G}$) and model–state pair (Σ); their mappings can be directly derived from Definition 3. The semantics of the other emotions are easily reconstructed by analogy. Furthermore, it is assumed that an action/plan π is removed from an agent's agenda Ag as soon as the agent has executed π, which is expressed by the following constraint:

$$\pi \in Ag(i)(\mathrm{M}, s) \ \& \ (\mathrm{M}', s') \in R_{\mathcal{A}}(i, \pi)(\mathrm{M}, s) \Rightarrow$$
$$\pi \notin Ag(i)(\mathrm{M}', s') \tag{4}$$

This constraint can be read as follows: if π is on the agenda Ag of agent i in state s of model M and executing π leads to the new state s' of model M', then π will not be on the agenda Ag of agent i in state s'. Of course an agent could have put a new instance of plan π on its agenda after performing the 'old' π, but we assume this does not violate the constraint above, because we treat these plans as different *instantiations* of π. Finally, note that we do *not* assume $\Gamma(i)(\mathrm{M}, s) \nvdash \bot$, so goals may be mutually inconsistent.

Having defined the semantic operators, we can present how formulas in $\mathcal{L}$ are interpreted.

Definition 6. *(Interpretation of formulas). Let $\mathrm{M} = \langle S, \vartheta, R_{\mathbf{B}} \rangle$, $\langle \Sigma, R_{\mathcal{A}} \rangle$, and $\langle \mathcal{C}, Ag, \Gamma, E \rangle$ be structures defined as above. Formulas in language $\mathcal{L}$ are interpreted in model–state pairs as follows:*

$$
\begin{aligned}
\mathrm{M}, s &\models p & &\Leftrightarrow p \in \vartheta(s) \quad for\ p \in \mathcal{P} \\
\mathrm{M}, s &\models \neg\varphi & &\Leftrightarrow \mathrm{M}, s \not\models \varphi \\
\mathrm{M}, s &\models \varphi_1 \wedge \varphi_2 & &\Leftrightarrow \mathrm{M}, s \models \varphi_1 \ \& \ \mathrm{M}, s \models \varphi_2 \\
\mathrm{M}, s &\models \mathbf{B}_i\varphi & &\Leftrightarrow \forall s' \in R_{\mathbf{B}}(i)(s) : \mathrm{M}, s' \models \varphi \\
\mathrm{M}, s &\models \mathbf{G}_i\kappa & &\Leftrightarrow \kappa \in \Gamma(i)(\mathrm{M}, s) \\
\mathrm{M}, s &\models \mathbf{A}_i\pi & &\Leftrightarrow \pi \in \mathcal{C}(i)(\mathrm{M}, s) \\
\mathrm{M}, s &\models \mathbf{Com}_i(\pi) & &\Leftrightarrow \pi \in Ag(i)(\mathrm{M}, s) \\
\mathrm{M}, s &\models [\mathbf{do}_i(\pi)]\varphi & &\Leftrightarrow \\
& & &\quad \forall (\mathrm{M}', s') \in R_{\mathcal{A}}(i, \pi)(\mathrm{M}, s) : \mathrm{M}', s' \models \varphi
\end{aligned}
$$

$$M, s \models \textbf{hope}_i(\pi, \kappa) \iff (\pi, \kappa) \in Hope(i)(M, s)$$
$$M, s \models \textbf{fear}_i(\pi, \neg\kappa) \iff (\pi, \neg\kappa) \in Fear(i)(M, s)$$
$$\vdots$$

Note that we evaluate formulas in state s of model M. The Kripke structure $\langle \Sigma, R_{\mathcal{A}} \rangle$ is then used for the interpretation of $[\textbf{do}_i(\pi)]\varphi$ formulas. In the rest of this paper, we will express that some formula φ is a validity (i.e. $\forall (M, s) \in \Sigma : M, s \models \varphi$) simply as $\models \varphi$.

Finally, we define a notion of possible intention equivalent to the one by Meyer (2004). An agent has the possible intention to perform plan π in order to accomplish κ if and only if it believes that 1) it has the ability to perform π, 2) κ is a goal of the agent, and 3) the execution of π possibly leads to a state where κ holds.

Definition 7. *(Possible intention). The possible intention I to perform π in order to accomplish κ is defined as:*

$$I(\pi, \kappa) \leftrightarrow \textbf{B}(\textbf{A}\pi \wedge \textbf{G}\kappa \wedge \langle \textbf{do}(\pi) \rangle \kappa) \qquad (5)$$

We claim that the framework specified above is suitable for formalizing the eliciting conditions of the emotions from the OCC model. We are also developing a quantitative model capable of modeling the intensities, thresholds, and potentials of emotions and their interactions, as described by the OCC model. However, because of space limitations, we cannot present a full quantitative model incorporating all these aspects here. For the example emotions described in this paper (i.e. hope and fear), we omit the treatment of potentials and thresholds. We restrict intensity values to the non-negative reals. This yields the minimal model required for showing the interplay between hope and fear as described by OCC.

Definition 8. *(Emotion intensity). The partial function* $intensity$ *assigning intensities to emotions is declared as:*

$$intensity : \mathcal{G} \times \Sigma \times Emotions \to \mathbb{R}^+$$

When supplied with an emotion, this function determines its intensity value. So the $intensity$ *function has at least 22 definitions (one for each emotion type), of which we will define the* **hope** *and* **fear** *cases later in this paper. Furthermore,* $intensity(i)(M, s)(\epsilon)$ *is undefined if* $M, s \not\models \epsilon$. *The* $intensity$ *function is defined per agent and model–state pair; however, for convenience we will hence omit these arguments.*

A Formal Model of Emotions

The OCC model provides for each emotion (among others) a concise definition in a single sentence and a list of variables affecting the intensity of the emotion in question. Below we will repeat OCC's definitions of *hope* and *fear* (given in Ortony, Clore, & Collins (1988), page 112) and show how they can be formalized in the language we have just defined. We will deal with the intensity part of these emotions later in this paper.

Hope: According to OCC, *hope is being pleased about the prospect of a desirable event.* From the viewpoint of a goal-directed agent, a *desirable event* can be translated to the accomplishment of a goal or part thereof, whereas the *prospect* can be translated to 'having' a plan for accomplishing that goal. More specifically, we require the agent to intend to perform this plan and to be committed to it. An agent that is *being pleased* about the prospect of a desirable event should act according to this mental state, so here we are hinted at a possible heuristic that can be associated with the emotion hope, namely to keep the intention and commitment while this emotion is strong enough. What exactly it means for hope to be strong enough will be formalized later.

Thus phrased in our language, an agent hopes to achieve some goal using some plan if and only if it intends to perform the plan for the goal and is committed to the plan. The objects of the hope emotion are then the goal that the agent intends to achieve and the plan to which it is committed. We thus arrive at the following formula characterizing hope:

$$\textbf{hope}(\pi, \kappa) \leftrightarrow (I(\pi, \kappa) \wedge \textbf{Com}(\pi)) \qquad (6)$$

It is important to note the use of the bi-implication, because it allows for the derivation of interesting properties (to be discussed later).

Fear: According to OCC, *fear is being displeased about the prospect of an undesirable event.* However, OCC note that if one experiences hope with respect to the prospect of a desirable event, then the absence of that event will be undesirable to the same degree. In other words, hope and fear are complementary emotions. This means that the intensities associated with hope and fear with respect to the same prospect and event have to sum to a constant. Note that this is different from what we called opposing emotions.

Because we have translated a desirable event as the accomplishment of a goal (or part thereof), an *undesirable event* will constitute the failure to achieve that goal (or part thereof). So fear will arise when the complement of an event hoped for becomes probable (this is the *prospect* part). An agent that is *being displeased* about the prospect of an undesirable event should start considering alternatives in order to ensure that it is the desirable event which will be achieved. Again, how exactly this can be done will be formalized later.

Thus phrased in our language, an agent fears the failure to achieve some goal using some plan if and only if it hopes the plan will achieve the goal but it believes that it may not. The objects of the fear emotion are then the plan from the corresponding hope emotion and the negation of the goal that it is hoping to achieve. We thus arrive at the following formula characterizing fear:

$$\textbf{fear}(\pi, \neg\kappa) \leftrightarrow (\textbf{hope}(\pi, \kappa) \wedge \textbf{B}\langle \textbf{do}(\pi) \rangle \neg\kappa) \qquad (7)$$

Because fear is the complement of hope (in the sense described above), hope must be a precondition of fear. The other precondition, namely that the complement of the event hoped for has become probable, is expressed as the belief that the execution of the intended plan *may* fail to achieve the desired event (note the angled brackets). As with the definition of hope, it is also important to note the use of the bi-implication.

It should be emphasized that the two emotion axioms above act as constraints on the functions *Hope* and *Fear* in the semantics of our language. The fact that these two axioms look like mere abbreviations of formulas (as in the case

of Equation (5) for I) is coincidental. In our complete qualitative formalization of OCC, most emotional fluents cannot simply be written as one side of a bi-implication.

Properties of Emotions

Having defined hope and fear in our formal model, we can check whether we can derive interesting properties from these definitions and whether the derivable properties are intuitive. In this section we will discuss several propositions; their proofs are omitted due to space limitations, but they are easy to verify.

$$\models \mathbf{hope}(\pi, \kappa) \rightarrow [\mathbf{do}(\pi)]\neg\mathbf{hope}(\pi, \kappa) \quad (8)$$

Hope only lasts for the duration of the prospect. As soon as the agent has performed plan π with which it hoped to achieve goal κ, the hope disappears, because it is no longer committed to π. This follows almost directly from constraint (4) in combination with definition 6, validating $\models \mathbf{Com}(\pi) \rightarrow [\mathbf{do}(\pi)]\neg\mathbf{Com}(\pi)$, and the fact that commitment is a precondition for hope. Note however, that it is possible for an agent to become committed to a *new instance* of π and experience 'renewed hope.'

$$\models \mathbf{fear}(\pi, \neg\kappa) \rightarrow [\mathbf{do}(\pi)]\neg\mathbf{fear}(\pi, \neg\kappa) \quad (9)$$

Similarly to hope, fear only lasts for the duration of the prospect. Indeed, this follows directly from the corresponding property of **hope** above and the fact that hope is a precondition for fear. Note that this proposition does not say anything about whether or not the agent succeeded in bringing about κ by performing π, only that it will not stay afraid afterwards.

$$\models \mathbf{B}[\mathbf{do}(\pi)]\neg\kappa \rightarrow (\neg\mathbf{hope}(\pi, \kappa) \wedge \neg\mathbf{fear}(\pi, \neg\kappa)) \quad (10)$$

If the agent believes it has no chance of accomplishing goal κ using plan π, then it will not hope for the impossible, nor fear the inevitable. The fact that there is also no fear in the consequent follows from the definition of **fear**, which validates $\models \neg\mathbf{hope}(\pi, \kappa) \rightarrow \neg\mathbf{fear}(\pi, \neg\kappa)$.

$$\models \mathbf{fear}(\pi, \neg\kappa) \rightarrow \mathbf{B}(\langle\mathbf{do}(\pi)\rangle\kappa \wedge \langle\mathbf{do}(\pi)\rangle\neg\kappa) \quad (11)$$

An agent experiencing fear with respect to a plan π and a goal κ believes that both the success and failure to accomplish κ are possible outcomes of performing π. This logical model does not tell anything about the likelihood of any of these outcomes; this is indeed something that will have to be taken into account by the quantitative model.

$$\models \langle\mathbf{do}(\pi)\rangle\varphi \rightarrow [\mathbf{do}(\pi)]\varphi \;\Rightarrow\; \models \neg\mathbf{fear}(\pi, \neg\kappa) \quad (12)$$

An agent that can predict the exact outcome of its actions will never experience fear. So with our definitions of **hope** and **fear**, we can express both the complementary nature of these emotions as described by OCC and the non-occurrence of fear in deterministic environments; in other words, an agent will never experience fear with respect to deterministic plans! This agrees with the intuitive notion that agents should not fear an undesirable outcome of deterministic actions, because they can predict the exact outcome beforehand. On the other hand, agents in nondeterministic or partially observable environments should always hope and fear simultaneously, because they cannot predict with absolute certainty whether or not their plan π for achieving goal κ will succeed.

Effects on Deliberation

Now that we have specified *when* an agent experiences hope or fear, we can try to specify *what to do* with these emotions. Recall that hope and fear are complementary emotions, so the specifications of the effects of these emotions must combine them both. There are two possible combinations of **hope** and **fear**:

1. **hope** but no **fear**: this case is similar to the effect of being *happy* as defined by Meyer & Dastani (2004; 2006). When an agent is hopeful with respect to a plan and a goal, it is said by OCC to *be pleased* about how its plan is progressing, so it should keep its intention and commitment with respect to the plan and the goal. So in this case, the heuristic is to ensure that further deliberation of the agent, denoted as *deliberate*, does not change this:

$$(\mathbf{hope}(\pi, \kappa) \wedge \neg\mathbf{fear}(\pi, \neg\kappa)) \rightarrow$$
$$[\mathbf{do}(deliberate)](I(\pi, \kappa) \wedge \mathbf{Com}(\pi)) \quad (13)$$

2. Simultaneous **hope** and **fear**: this is the more interesting case. The OCC model defines fear as *being displeased* about the prospect of an undesirable event. But what could the effect of being displeased be? Doing something about it! An agent experiencing fear with respect to a plan and a goal should be allowed to replan in order to find a new plan that can accomplish the goal:

$$(\mathbf{hope}(\pi, \kappa) \wedge \mathbf{fear}(\pi, \neg\kappa)) \rightarrow$$
$$[\mathbf{do}(deliberate)]((I(\pi, \kappa) \wedge \mathbf{Com}(\pi)) \vee$$
$$(I(\pi'', \kappa) \wedge \mathbf{Com}(\pi'')))$$

where $\pi'' = (replan(\pi, \kappa, \pi'); \pi')$. We assume an agent has the ability to replan by performing the deliberation operation $replan(\pi, \kappa, \pi')$, which provides an alternative plan π' instead of π to achieve goal κ. Plan π' may depend on the original plan π or even be equal to π if an alternative plan cannot be found. For a proper and complete definition of the $replan$ function, we refer the reader to Dastani *et al.* (2003).

Is the formula above a good heuristic? No, because it is not specific enough. The disjunction $\mathbf{Com}(\pi) \vee \mathbf{Com}(replan(\pi, \kappa, \pi'); \pi')$ does not specify *when* an agent should start replanning, only that it *may* do so. Because hope and fear are complementary emotions (i.e. their intensities always add up to a constant), a reasonable heuristic would be one that states that the agent should start replanning as soon as the intensity of the fear w.r.t. the undesirable event is greater than the intensity of the hope w.r.t. the desirable event. However, this cannot be expressed in a purely logical model. Therefore, a quantitative model of emotions is needed in order to complete the heuristic.

According to the OCC model, the variables affecting the intensity of hope w.r.t. a desirable event are 1) the degree to which the event is desirable and 2) the likelihood of the event. Analogously, the variables affecting the intensity of fear w.r.t. an undesirable event are 1) the degree to which the event is undesirable and 2) the likelihood of the event. We can thus define the *intensity* function for **hope** and **fear** as:

146

$$intensity(\mathbf{hope}(\pi, \kappa)) :=$$
$$\mathcal{I}_{\mathbf{hope}}(desirability(\kappa), likelihood(\pi, \kappa)),$$
$$intensity(\mathbf{fear}(\pi, \neg\kappa)) :=$$
$$\mathcal{I}_{\mathbf{fear}}(undesirability(\neg\kappa), likelihood(\pi, \neg\kappa)).$$

Here the functions $desirability$, $undesirability$, and $likelihood$ return application-dependent and platform-dependent measures of the (un)desirability of (not) achieving κ and the likelihood of the execution of plan π resulting in a state where $(\neg)\kappa$ holds, respectively. The functions $\mathcal{I}_{\mathbf{hope}}$ and $\mathcal{I}_{\mathbf{fear}}$ then combine these measures, according to this greatly simplified quantitative model, to a non-negative real value. It should be noted that the functions $\mathcal{I}_{\mathbf{hope}}$, $\mathcal{I}_{\mathbf{fear}}$, $desirability$, $undesirability$, and $likelihood$ all implicitly depend on the agent and model–state pair passed to the $intensity$ function (see Definition 8). Now we can complete the heuristic by specifying that an agent should keep its commitment with respect to a plan while its hope with respect to that plan is greater than its fear, whereas the agent should start replanning when its fear is greater than its hope:

$$(\mathbf{hope}(\pi, \kappa) \wedge \mathbf{fear}(\pi, \neg\kappa) \wedge$$
$$intensity(\mathbf{hope}(\pi, \kappa)) \geq intensity(\mathbf{fear}(\pi, \neg\kappa))) \rightarrow$$
$$[\mathbf{do}(deliberate)](I(\pi, \kappa) \wedge \mathbf{Com}(\pi)), \quad (14a)$$
$$(\mathbf{hope}(\pi, \kappa) \wedge \mathbf{fear}(\pi, \neg\kappa) \wedge$$
$$intensity(\mathbf{hope}(\pi, \kappa)) < intensity(\mathbf{fear}(\pi, \neg\kappa))) \rightarrow$$
$$[\mathbf{do}(deliberate)](I(\pi'', \kappa) \wedge \mathbf{Com}(\pi'')) \quad (14b)$$

where $\pi'' = (replan(\pi, \kappa, \pi'); \pi')$. Formulas (13), (14a), and (14b) now specify the complete heuristic.

As an example of how emotions affect an agent's deliberation, suppose an agent has a number of plans of which it must select one to execute. Without emotions, the agent should decide at each time point which plan to execute, possibly resulting in erratic behavior. However, with emotions the selection of one of the plans commits the agent resulting in the appraisal of *hope*; that is, the agent is pleased with the prospect of achieving some goal associated with the plan. When the execution of the plan fails (e.g., the execution of some action fails or its effect is not perceived), the agent will also experience *fear* with respect to the plan of not achieving the goal. If the intensity of the fear is greater than the intensity of the hope, the agent may drop the plan or start replanning to achieve the goal. Of course, if given multiple revised plans to choose from, the agent will prefer new plans that have previously caused it to experience *relief*, while avoiding those that have previously resulted in the emotion *fears-confirmed*. In this way, the incorporated model of emotions indicates which plans need to be dropped or adopted and thereby helps to reduce the nondeterminism involved in an agent's decision making process. Furthermore, emotions concerning agents as objects and actions of other agents can be used to influence the social behavior of an agent; however, these emotions will be a topic of future work.

Conclusion and Future Work

We have presented a formal model of emotions to specify the behavior of *intelligent* agents. This model is a formalization of a part of the OCC model of emotions. Our formalization aims at a precise description of concepts used in the OCC model in order to improve the decision-making of agents. In particular, we have shown how the emotions hope and fear as described in the OCC model can influence the deliberation process of an agent. Note that our model should be interpreted in a prescriptive way. One may argue that this interpretation falls short of human emotions, e.g., by attacking formula (12), because people sometimes fear inevitable consequences of their actions. However, we consider this as irrational and undesirable for *intelligent* artificial agents.

At the time of writing, we have completed a qualitative formalization of all 22 emotions in the OCC model. We are currently working on a quantitative model incorporating emotion potentials, thresholds, and intensities, as well as investigating how functions like *desirability* and *likelihood* can be defined. For future work, we intend to use our formalization to assist in generating an agent's behavior. To this end, we will develop transition semantics and an implementation of our formalization in an agent-oriented programming language. We plan to evaluate the emotional model by running it on the Philips iCat, which is a cat-shaped robot with a humanlike face capable of making believable emotional expressions. Current scenarios for the robot are that of a companion robot for elderly and a cooking assistant.

References

Bratman, M. E. 1987. *Intention, Plans, and Practical Reason*. Cambridge, Massachusetts: Harvard University Press.

Cohen, P. R., and Levesque, H. J. 1990. Intention is Choice with Commitment. *Artificial Intelligence* 42:213–261.

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

Dastani, M., and Meyer, J.-J. Ch. 2006. Programming Agents with Emotions. In *Proc. of ECAI'06*, 215–219.

Dastani, M.; Boer, F. S. d.; Dignum, F.; and Meyer, J.-J. Ch. 2003. Programming Agent Deliberation: An Approach Illustrated Using the 3APL Language. In *Proceedings of AAMAS'03*, 97–104.

Gratch, J., and Marsella, S. 2004. A Domain-independent Framework for Modeling Emotions. *Journal of Cognitive Systems Research* 5(4):269–306.

Meyer, J.-J. Ch.; Hoek, W. v. d.; and Linder, B. v. 1999. A Logical Approach to the Dynamics of Commitments. *Artificial Intelligence* 113:1–40.

Meyer, J.-J. Ch. 2004. Reasoning about Emotional Agents. In de Mántaras, R. L., and Saitta, L., eds., *Proceedings of ECAI'04*, 129–133. IOS Press.

Oatley, K., and Jenkins, J. M. 1996. *Understanding Emotions*. Oxford, UK: Blackwell Publishing.

Ortony, A.; Clore, G. L.; and Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge, UK: Cambridge University Press.

Picard, R. W. 1997. *Affective Computing*. MIT Press.

Sloman, A. 2001. Beyond Shallow Models of Emotion. *Cognitive Processing* 2(1):177–198.