Towards a Quantitative Model of Emotions for Intelligent Agents

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

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

Abstract. Starting out from a qualitative formalization of a well-known psychological model of emotions that we have developed in previous work, we now turn to the quantitative aspects of emotions, and investigate how these can be incorporated into the qualitative model.

1 Introduction

A popular computational model of emotions is that of Ortony, Clore & Collins [1] (henceforth to be referred to as "the OCC model" or simply "OCC"). Steunebrink, Dastani & Meyer [2] have formalized, in dynamic / doxastic logic, the conditions that elicit each of the emotions of the OCC model for cognitive agents. This formalization amounts to a *qualitative* model of emotions, specifying the conditions for *when* an emotion is experienced. However, a *quantitative* model for these emotions is missing. For example, the qualitative model does not specify how strong an emotion should be, for how long an agent should be aware of experiencing an emotion, in what ways and for how long an emotion should influence the behavior of an agent, according to what function the strength of an emotion should change over time, etc. Such quantitative aspects of emotions should be taken into account in any serious attempt at specifying a formal model of emotions, in particular the OCC model.

In this paper, we will treat the questions of which quantitative aspects of emotions (of the OCC model) should be considered, how these quantitative aspects can be formally modeled, and how a quantitative model of emotions can be combined with the qualitative model of Steunebrink *et al.* [2]. To illustrate the presented ideas, some formulas will be shown for "joy" (which is the shortest emotion label), but any emotion can be substituted for it.

This paper is outlined as follows. First the qualitative and quantitative aspects of emotions in the OCC model will be discussed. Then the way in which the qualitative aspects have been formalized will be reviewed. Finally, a way of formalizing and integrating quantitative aspects will be proposed.

2 Qualitative and Quantitative Aspects of Emotions in the OCC Model

The OCC model describes a hierarchy classifying 22 emotions, of which half are positive and half are negative. They are: joy, distress, hope, fear, satisfaction, disappointment, relief, fears-confirmed, happy-for, resentment, gloating, pity, pride, shame, admiration, reproach, love, hate, gratification, remorse, gratitude, and anger. 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). Each pair of emotions in the listing above (i.e. joy–distress, hope–fear, etc.) are called *opposing*, but this does not rule out the possibility of such a pair being triggered simultaneously. Indeed, agents are allowed to experience 'mixed feelings' with respect to a certain situation. The conditions that *elicit* these 22 emotions are described by OCC very concisely in natural language.

Steunebrink *et al.* [2] have formalized these eliciting conditions in dynamic logic, creating a formal qualitative model of emotions. (Note that in [2] only a subset of the qualitative model is described due to space limitations; a complete formalization exists nonetheless.) This model specifies precisely *when* emotions are triggered, but quantitative aspects are missing.

In the OCC model, quantitative aspects of emotions are described in terms of *potentials*, *thresholds*, and *intensities*. For each of the 22 emotions, OCC provide a list of variables that affect the intensity of that emotion if its eliciting conditions hold. Perhaps somewhat misleadingly, the idea is that the weighted sum of these variables equals the emotion's potential. The intensity of an emotion is defined as its potential minus its threshold, or zero if the threshold is greater than the potential. The workings of thresholds of emotions are not specified by OCC, but they probably depend on global variables indicating an agent's 'mood.' For example, if an agent is in a good mood, the thresholds of the eleven negative emotions are increased, causing a lower (or zero) intensity to be associated with a negative emotion, if one is triggered. Emotions that are assigned a positive intensity may in turn influence the mood of an agent, entangling the dynamics of long-term moods and short-term emotions.

The main idea behind separating emotion potentials and intensities is to be able to reason about why some emotions have no effect on an agent (i.e. the agent is not aware of the emotion and it does not influence its behavior) even though their eliciting conditions hold; namely, when the eliciting conditions of an emotion hold (as defined by a formal qualitative model), but its potential is calculated to be below its threshold, an agent can still recognize that an emotion has been triggered and reason about why it does not affect him (e.g., maybe the mood of the agent was 'too good' for him to be affected by shame, even though the agent is aware it performed a blameworthy action).

3 From a Qualitative to a Quantitative Formalization of Emotions

In the construction of a quantitative model of emotions, we build upon the qualitative model of Steunebrink *et al.* [2]. This qualitative model is an extension of the KARO framework [3, 4], which is a mixture of multi-agent dynamic logic and epistemic / doxastic logic, additionally providing several (modal) BDI operators for dealing with the motivational aspects of artificial agents. In the qualitative model, emotions are represented as *emotional fluents* that can be reasoned with in the object language. There are 22 types of emotional fluents, one for each of OCC's emotion types. Each emotional fluent takes between one and three arguments, namely, the *objects* of the emotion in question. For example, the single object of joy is a goal formula (containing one or more subgoals which the agent experiencing joy has just accomplished), whereas the three objects of gratitude are an agent, the (praiseworthy) action it performed, and a conjunction of subgoals it accomplished for the agent experiencing gratitude. Moreover, for each of the 22 emotion types, there is a corresponding set, defined in the semantics of the language, which contains the appropriate objects for that emotion type. An emotional fluent then holds in the object language if and only if the tuple of its objects is an element of the semantically defined set corresponding to that emotional fluent.

The crux here is that the conditions under which these semantically defined sets contain certain elements (i.e. the objects of an emotion) correspond closely to the eliciting conditions of the emotion type in question as prescribed by the OCC model. For example, because the OCC model defines joy as *being pleased about a desirable event*, the semantic set $Joy(a, s)$ for an agent a in state s only contains an event κ if an action has just been performed that has led the agent to become aware of the accomplishment of one or more previously unachieved (sub)goals. Defining emotions by using syntactic emotional fluents backed by semantic sets (i.e. $s \models \mathbf{joy}_a(\kappa) \;\Leftrightarrow\; \kappa \in Joy(a, s)$) has as advantage that propositions in the object language containing emotional fluents can actually be proven semantically. In other words, there is no need for defining emotions as abbreviations of other formulas; in fact, none but two of the emotions (i.e. hope and fear) can be written as one side of a bi-implication in the object language of the qualitative model.

The quantitative model that we propose is built on top of the described qualitative model as follows. The satisfaction of an emotional fluent in a certain state (e.g., $s \models \mathbf{joy}_a(\kappa)$) is regarded as a *trigger* for associating a potential, threshold, and intensity with the emotional fluent and for calculating their quantities. The potential and threshold of each triggered emotion are only defined for the state in which the corresponding emotional fluent holds. Emotion intensities, on the other hand, endure for some time, usually decreasing over time. In order to model enduring quantities, the dynamic logic underlying the qualitative model will need to be extended with a state history. In a dynamic logic (such as KARO), a history of states visited by the multi-agent system and actions taken in those states can easily be kept in the Kripke structure of actions. Given such a history of the multi-agent system, a (collective) memory of emotions experienced by each agent in each state can be constructed; namely, as a set containing for each state in the history all emotional fluents that hold in that state. Emotion intensities are then defined for all emotional fluents that are contained in the emotional memory of the multi-agent system. So potentials and thresholds are only associated with emotional fluents that hold in the actual state, whereas intensities are associated with each emotional fluent in the emotional memory, making them persist over multiple states of the multi-agent system.

Because functions defining the values of intensities are obviously dependent on a time parameter, the dynamic logic of the qualitative model must be further extended to include temporal information so that an explicit representation of time is available to the intensity functions. The intensity of an emotion can be modeled using any function (this is application-dependent), but an inverse sigmoid function could be a natural default choice. However, as an inverse sigmoid function only reaches zero in the limit, an additional cut-off threshold must be set so that the intensity of an emotion (together with its effect on an agent's behavior) can be discarded within a finite amount of time. Note that in the qualitative model there exist several conditions under which a pair of opposing emotions are triggered simultaneously, but this implies by no means that the intensities associated with these opposing emotions should be the same, or be related at all (with the exception of hope and fear, who's intensities should be complementary and always sum to a constant [1]). For example, it is possible that gloating and pity with respect to the same agent and event are triggered simultaneously, but that the emotional fluent corresponding to gloating gets assigned zero intensity, whereas pity gets assigned some positive intensity, because the other agent is being admired. Here we see that even though the qualitative model prescribes that in a certain situation opposing emotions can be experienced together, the quantitative model does not have to lead to mixed feelings.

In order to actually define the emotional 'characters' of agents for some application, 66 functions must be specified for each agent, i.e. a potential, threshold, and intensity function for each of the 22 emotion types. (Although this may appear to be a lot, many of these functions may be very similar.) For the running example, consider for the emotion "joy" the potential function $\mathcal{P}_{\mathbf{joy}}$, the threshold function $\mathcal{T}_{\mathbf{joy}}$, and the intensity functions $\mathcal{I}_{\mathbf{joy}}$ (these are three of the 66 mentioned functions). In general, the parameters of each of these 66 functions are the current state of the multi-agent system, the time, and the objects of the emotion in question, because from these parameters all necessary information about the state of mind of an agent can be derived. The po*ten*tial functions must be defined as the weighted sum (denoted as $\overrightarrow{w} \cdot \_$) of the variables (denoted as $\overrightarrow{x}$) that according to OCC affect intensity, e.g., $\mathcal{P}_{\mathbf{joy}}(a, s) := \overrightarrow{w} \cdot \overrightarrow{x} = q_p$ (we assign this quantity to q_p for brevity). How these variables themselves (e.g., effort, desirability, likelihood) are calculated and what weights are used are again application-dependent and defining for the emotional 'character' of the agent. The OCC model does not prescribe any particular constraints for the threshold functions, but they are hinted to depend on the 'mood' of the agent, e.g., $\mathcal{T}_{\mathbf{joy}}(a, s) := -w \cdot \mathit{Mood}(a, s) = q_t$ (we assign this quantity to q_t for brevity).

Note that because joy is a positive emotion, the mood must negatively influence the emotion's threshold (denoted as $-w \cdot \_$), i.e., a positive mood must lower the threshold so that the intensity will be greater; indeed, there should be no minus sign in the threshold functions for the negative emotions. As mentioned above, a (collective) emotional memory *Emem* is defined over the state history of the multi-agent system, so if at some later time point t' the multi-agent system is in state s', one can retrieve the potential and threshold quantities (q_p and q_t in this example) of the triggered emotions, e.g., $(\mathbf{joy}_a(\kappa), q_p, q_t, t) \in Emem(s')$. (Note that the framework should be extended with temporal information so that the time of triggering t can also be recorded.) An intensity quantity is then associated with each entry in *Emem*. The intensity functions can by default be inverse sigmoid functions, with the difference between the potential and threshold as the numerator (if this difference is negative, the intensity should be set to zero altogether), and the difference between the current time and triggering time in the exponent, e.g., $\mathcal{I}_{\mathbf{joy}}(a, s', t') := \max(0, \frac{q_p - q_t}{1+e^{(t'-t-\mu)\delta}} - \theta)$. Some variables specifying the inverse sigmoid function also need to be set, i.e. the half-life μ, fall-off speed δ, and cut-off threshold θ, such that $\mu\delta \approx -\ln 0.01$ and $0 < \theta \ll q_p - q_t$. Note that other types of intensity functions are also possible (especially for hope and fear).

4 Related Work, Conclusions, and Future Work

Previous work on specifying and implementing emotions carried out by Meyer [3] and Dastani [5] follows Oatley & Jenkins' model of emotions [6] and comprises only four emotion types (happy, sad, angry, fearful), but quantitative aspects of emotions were not considered. Our work is also similar to other computational models of emotions, such as EMA [7], CogAff [8], and the work of Picard [9].

In this paper, we have provided an overview of how the conditions eliciting the 22 emotions of the OCC model have been formalized in a qualitative model. We have described how this qualitative model can be extended to a quantitative model of emotions. Specifically, we have described how potentials, thresholds, and intensities can be associated with the qualitative emotions, and how and when their values should be calculated. Issues such a computational complexity and the possible need to empirically determine parameter settings have been taken into account when developing this formalization, but we have not yet explicitly dealt with these issues and justified our choices.

In future work, the dynamics of a 'mood' and its influence on threshold functions has to be further taken into account. It also remains to be investigated how functions for calculating e.g. effort, desirability, and likelihood can be defined. Furthermore, it remains to be studied what the actual effects of emotions with positive intensities should be on the deliberation of an agent, i.e., in what way emotions can function as heuristics for focussing the attention of the deliberation process of an agent.

References

1. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, UK (1988)
2. Steunebrink, B.R., Dastani, M., Meyer, J.-J.Ch.: A logic of emotions for intelligent agents. In: Proceedings of AAAI'07, AAAI Press (2007)
3. Meyer, J.-J.Ch.: Reasoning about emotional agents. In de Mántaras, R.L., Saitta, L., eds.: Proceedings of ECAI'04, IOS Press (2004) 129–133
4. 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
5. Dastani, M., Meyer, J.-J.Ch.: Programming agents with emotions. In: Proceedings of ECAI'06. (2006) 215–219
6. Oatley, K., Jenkins, J.M.: Understanding Emotions. Blackwell Publishing, Oxford, UK (1996)
7. Gratch, J., Marsella, S.: A domain-independent framework for modeling emotions. Journal of Cognitive Systems Research **5**(4) (2004) 269–306
8. Sloman, A.: Beyond shallow models of emotion. Cognitive Processing **2**(1) (2001) 177–198
9. Picard, R.W.: Affective Computing. MIT Press (1997)