Reasoning about Emotional Agents

John-Jules Ch. Meyer

ICS, Utrecht University, Utrecht

The full version of this paper appeared in Proceedings of ECAI 2004, IOS Press, 2004.

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

However, emotions and rationality are known to be more interconnected than one may suspect. There is psychological evidence that having emotions may help one to do reasoning and tasks for which rationality seems to be the only factor [1]. Moreover, work by e.g. Sloman [5] shows that one may think of *designing* agent-based systems where these agents show some kind of emotions, and, even more importantly, display behaviour dependent on their emotional state. It is exactly in this sense that we aim at looking at emotional agents: artificial systems that are designed in such a manner that emotions play a role. Also in psychology emotions are viewed as a structuring mechanism. Emotions are held to help human beings to choose from a myriad of possible actions in response to what happens in our complex world (cf. [3]).

So we advocate the use of emotional states to design an artificial intelligent agent. One has to bear in mind, that this has in itself nothing to do with the philosophical and very difficult question whether these agents really possess true emotions in the sense that we humans do! This is similar to the question whether artificial agents possess true intelligence or consciousness like humans do. One can perfectly well think about the design of intelligent agents without addressing this issue. In this paper we argue that emotions make sense in describing the behaviour of certain intelligent agents, and may help structuring the design of the agent (by means of an architecture that caters for emotional aspects) and consequently it is useful to reason about emotions of an agent, or rather about the emotional states an agent may be in, together with its effects on the agent's actions, as an important aspect of the agent's behaviour.

So our logic is more concerned with the behaviour of such a system than with emotions *per se*. This is a perfectly sensible way to go in line with software and system engineering practice. To specify systems in a rigourous way one may employ certain logical methods by which one can unambigously state how the system should behave. In agent-based systems where the agents are perceived as *rational* or *intelligent* ones, possessing some sort of attitudes pertaining to information and motivation such the well-known BDI (belief–desire–intention) agents, we can describe their behaviour in terms of the evolution of the mental states of the agent

over time (e.g. BDI logic [4], and KARO logic [2]). We now want also perceive emotional agents as systems that evolve over time and can be described by some logic as the one mentioned above for rational agents. So what we aim at is describing behaviours of emotional agents in terms of the way their (emotional) states evolve over time. This means that we are interested in at least two things: how do actions of agents (by definition agents act!) change their emotional states and how do emotional states determine what action is taken and what effect is obtained from this in the given state.

The way the full paper proceeds is as follows. From the psychological literature we get evidence that the way emotions influence behaviour is on a rather high level. Emotions like happiness and fear generally do not result directly in taking concrete actions by agents, but rather in an attitude towards handling their goals and intentions. Emotions moderate the execution and maintenance of the agent's agenda, so to speak. It will turn out that we can model these high-level attitudes adequately in the logical framework that we have devised for rational agents. In essence our approach is thus: to reason about the dynamics of (emotional) states we use the framework of *dynamic logic* and (an extension of) the KARO framework ([2]) in particular. In the full paper we provide KARO-style formulas expressing emergence of the four basic emotions of happiness, sadness, anger and fear, as well as their influence on the agent's deliberative behaviour. A simplified example for illustrating the flavor of these formulas is the following: $\mathbf{I}(\pi, \varphi) \wedge \mathbf{Com}(\pi) \rightarrow [\pi](\mathbf{B}\varphi \rightarrow \mathit{happy}(\epsilon, \varphi, \varphi))$, expressing that an agent that was committed to a plan π which it intended to do with φ as goal, and that believes that its goal φ is realised after having executed/performed its plan π, is indeed happy with this (w.r.t. this goal φ and a (by now) empty plan ϵ).

References

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

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

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

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

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