Agents with or without Emotions?

Matthias Scheutz
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556, USA
mscheutz@cse.nd.edu
http://www.nd.edu/~mscheutz/

Abstract

This paper presents a critical view of the *status quo* of some of the emerging research efforts on emotional agents. It attempts to isolate (at least roughly) some of the reasons why emotional agents may be desirable and points to the difficulties of making notions of emotion precise. It lists various problems connected to emotional agents and concludes that it is counterproductive to the whole endeavor of understanding and modeling emotions if "emotion labels" are conferred upon states of agents prematurely without justification.

Introduction

Current research in AI shows an increasing interest in "agents with emotions".[1]. From "believable agents" in the entertainment industry, to human-computer or human-robot interaction in educational or instructional domains, "emotional agents" seem to find applications in a wide variety of areas. Yet, it is not clear at all whether these so-called emotional or affective agents deserve the properties that are–often very quickly–attributed to them by their users and creators alike. In short, it is at best an open question whether these agents *really have emotions*. This paper presents a critical view of the *status quo* of some of the emerging research efforts on emotional agents. It attempts to isolate (at least roughly) some of the reasons why emotional agents may be desirable and points to the difficulties of making notions of emotion precise. It then lists various problems connected to "alleged" emotional agents (for example, the difficulties of assessing whether an agent actually implement emotions). The concluding discussion briefly reiterates the word of caution, which is the underlying theme of the overall paper, namely that it is counterproductive to the whole endeavor of understanding and modeling emotions if "emotion labels" are conferred upon states of agents prematurely, without justification.

[1]For an overview of various models, see for example (Pfeiffer 1988; Picard 1997; Hatano, Okada, & Tanabe 2000)

Why emotional agents?

Research in AI has, since its beginning, focused on representational, deliberative mechanisms and processes (such as planning, searching, reasoning, etc.) to control various kinds of agents. These mechanisms have been employed in a variety of areas with great success. So, why should we be interested in emotional agents, where it seems–at least from our human experiential perspective–that emotions do not square well with rationality?

One reason may be that there are application domains where various kinds of constraints limit the application of deliberative methods. For example, certain domains or environments may impose time constraints on the processing time, or constraints on the available memory or computational resources of an agent. Or there could be requirements on reliability, fault tolerance, and others; most generally constraints on energy of some kind. In these resource constrained environments, deliberative methods may not be the best ways of controlling agents. Rather, different, partly non-representational, reactive methods of controlling agents, similar to those employed in many primitive biological organisms, may prove more appropriate and effective. In particular, so-called "affective states" (of which "emotions" are a special kind) can be quite useful as efficient, low-cost control mechanisms in unstructured, resource constrained, competitive environments.

Another reason may be that we need to build and implement models of emotions to verify our theoretical analyses and predictions as part of our efforts to understand the wealth of animal and human emotions. This kind of modeling fits nicely into (an extension of) the classical research loop of empirical discovery and theorizing: starting with a cognitive phenomenon in reality and its description in a theory, experiments are conducted to verify/falsify the given theory (e.g., (Oreskes, Shrader-Frechette, & Belitz 1994) or (Popper 1962)) by deducing a prediction for a particular case. The (empirically constructed) theory is then transformed into a computational model, i.e., the "emotional agent", and the empirical experiment replaced by a "virtual experiment" which is conducted by running a simulation of the agent model on a computer. The result of this vir-

tual and cyclic simulation process is twofold: (1) it creates predictions for "real world dynamics"; (2) if these predictions are not satisfactory, a possible change in the agent model may be required which, in turn, may necessitate changes in the original (empirically based) theory. In this case a rewritten version of the theory acts as the starting point for a new cycle of empirical and/or simulation experiments. Note, however, that it is quite possible to build agents that try to measure and assess human emotional states without the need for these agents to have emotions themselves (e.g., see (Picard 1997)).

Yet another reason may be that certain kinds of emotions, in particular, what (Sloman 1999) calls "tertiary emotions", may be a byproduct of certain processing mechanisms. For example, typical human-like emotions such as "guilt", "infatuation", and others seem to result in interruptions in deliberative and reflective processes (e.g., diverting attention to past episodes or to a restricted class of features). Some emotions can be construed as the loss of control of certain reflective processes that balance and monitor deliberative processes. Such emotions, then, are not part of the architectural specification of an agent, but rather *emerge* as a result of the interactions of various components in the agent's control system.[2]. If this analysis of tertiary emotions is right, a robot, which is capable of reflective processes similar to those of humans, will therefore also be capable of having emotions that can emerge from the interaction of those processes.[3]

The main reasons for emotional agents in AI, however, does not seem to fall into any of the above three categories. Rather, they seem to be based on a somewhat superficial analysis of how human-computer or human-robot interactions could be improved, be it in the entertainment industry or in the realm of eduction or computer-based instruction. The need for emotional agents is typically based on the claim that such agents can interact better, in a more *natural* way with humans and vice versa. While it is certainly true that (1) agents that take human affect into account in their responses will appear more "believable", "realistic", "interesting", etc. to humans (e.g., (Hayes-Roth 1995)), and (2) that such agents can be very helpful in computer-based instruction, it does not automatically follow that these agents have to be capable of having emotions themselves, nor that implementations of agents that appear "emotional" *actually implement emotions*. Yet, some authors seem to imply that the "believability" of an agent is intrinsically tied to its having emotions ((Reilly 1996)).

While the possibility for these agents to have emotions (for a given analysis of "emotion") cannot be excluded *apriori*, any such conclusion, if it is to be drawn, certainly requires an argument supporting it. An additional argument is necessary if such agents are taken to have "human-like emotions", since most likely these agents will be vastly different from humans in many respects (at least today's agents). Hence, a case needs to be made why they should be *similar with respect to emotions*. The next section will point to the intrinsic difficulties connected to such an argument given the very nature of emotions themselves.

Emotions are "cluster concepts"

Modeling and explaining mental phenomena in humans and other animals requires us to use concepts referring to those phenomena. The history of the philosophy of mind, and some of the methodological, terminological and scientific disagreements found in psychology and neuroscience, however, all point to serious problems in defining mental concepts like "belief", "desire", "intention", "emotion", "personality", and many others. This difficulty partly stems from the fact that mental concepts do not seem to be *classical concepts* in the sense that there is a clearly specified and delimited set of necessary and sufficient conditions to determine whether something is an instance of the concept. The problem is that while we can say what is more or less typical of emotions, for example, we cannot provide a set of individually necessary and jointly sufficient conditions, which would define the concept.[4] The difficulty in defining what emotions are within psychology is apparent from the numerous different, partly incompatible characterizations (e.g., see (Griffiths 1997) or (Oatley & Jenkins 1996)). In fact, it is not even clear what "basic emotions" are (Ortony & Turner 1990). These definitions stress various different aspects of emotions such as brain processes, peripheral physiological processes, patterns of behavior, eliciting conditions, functional roles, introspective qualities, etc.

Definitions of emotions typically also differ in scope. Some writer, for example, treat all motives or desires (e.g. hunger or curiosity) as emotions while others do not. Some regard "surprise" as an emotion, whereas others (e.g. (Ortony, Clore, & Collins 1988)) regard it as basically a cognitive state in which a belief or expectation has been found to be violated, which in turn may or may not produce an emotional reaction. While some (e.g., social scientists) tend to define "emotion" so as to focus on social phenomena, such as embarrassment, attachments, guilt or pride, brain scientists, for example, might define it to refer to brain processes and widespread animal behaviors. All of this diversity provides strong evidence that emotions are cluster con-

[2]Similar phenomena of "emergent states" can be observed in computer systems, e.g., when they "lock up" because all involved processes are waiting for resources that are owned by another waiting process

[3]As an aside, there will not be any robot such as "Commander Data" from Star Trek, which has human-like higher cognitive capacities (and beyond), while not being capable of having the emotions connected to that class of (cognitive) processes.

[4]Note that this "cluster property" does not make cluster concepts necessarily *vague* or *ambiguous*, nor do they have to be *degree concepts* (e.g., like "baldness").

cepts, and that it is unlikely that we will be able to find a characterization of "emotion" that equally applies to all different subspecies of emotions.

Another reason for the difficulty of getting a useful, workable definition of emotions is due to the fact that most (if not all) mental concepts seem to be intrinsically *architecture-based concepts* ((Sloman 2000)). When we use these concepts, we tacitly assume and refer to a particular (i.e., a human-like) "architecture". "Forgetting", "dreaming", "imagining", "feeling guilty", and many others are concepts that implicitly depend on a particular architecture (or class of architectures). It is not possible to "forget" something, if there is no sort of "memory" that has stored something in the first place. Hence, "forgetting" is a concept, which is defined relative to architectures that have memory components. Similarly, "disappointment" can be construed as an emotion defined for architectures with certain deliberative capacities (e.g., the ability to construe future scenarios, assess their likelihood and/or desirability, and compare them to current outcomes).

The architecture-based nature of mental concepts in general has consequences for our investigations of concepts like emotions in particular: if we attempt to define emotions simply in terms of familiar examples, such as "anger", "shame", "anxiety", etc. we risk implicitly restricting them to organisms with architectures sufficiently like ours. That would rule out varieties of fear or aggression found in insects, for example. Hence, we need an *architecture-neutral* characterization, which is hard to come by if it is to be applicable across a wide range of architectures (such as insect-like reactive architectures or deliberative architectures with mechanisms that can represent and reason about possible, but non-existent alternatives or future states). Our best hope is to define emotions in terms of *their functional role* which can be specified independent of the specific features of a given architecture. Rather, such functional definitions will have to be given relative to a class of architectures, where architectures are specified as *schemata* with schematic components that are instantiated by all architectures in the class (possibly in different ways). Emotion concepts, then, need to be shown to be supported by these components (directly or via processes supported by them). Once the concepts are analyzed as architecture-based and supported by a particular class of architectures, we will be able to specify exactly when a system "has" a certain emotion, namely if and only if its architecture can be seen to be an instance of the architecture scheme relative to which the emotion was defined in the first place.

Emotional agents in AI–what is advertised?

Given (1) that psychologists (let alone philosophers) are not in agreement about what exactly emotions are and (2) that we are still lacking a sufficiently precise characterization of emotions in terms of classes of architectures that support them, it should not come as a surprise that this terminological and conceptual plurality too is spread throughout the AI literature. Although there are many surveys of research on emotions (e.g.,(Ortony, Clore, & Collins 1988; Goleman 1996; LeDoux 1996; Picard 1997)), finding a balanced overview is very difficult. This difficulty is reflected in the struggles of AI researchers to make sense out of the psychological literature, which sometimes can lead to quite confusing, even idiosyncratic terminology. Some researchers, for instance, see emotions as special kinds of motivations (e.g.,(Breazeal 1998)), whereas others draw a clear distinction between motivations and emotions (e.g., (Canamero 1997)). Such subtle changes in terminology have the unhappy consequence that the research community as a whole can easily lose track of what its various members are talking about. This is not only undesirable for any discipline, but may eventually culminate in a terminology so conflated that claims involving any of its terms are nearly meaningless. And it does not help much that researchers often point out that they are not talking about "human emotions", "human motivations", or "human desires", if they later tacitly assume that the same emotion terms used for humans can be used for and meaningfully applied to their artifacts.

It is not a new phenomenon in AI to borrow labels for mental states from ordinary (human) discourse and apply them to state in agents. While this practice is *per se* not problematic as long as the difference is kept in mind, it does convey the flavor of something human-like actually being implemented in the agent. Yet, it is rarely acknowledged that a case needs to be made as to why these states deserve the same label as mental states in humans (i.e., it needs to be explicated what both have in common and why both belong to the same class despite the fact that the two architectures may differ vastly).

This tendency to present simplistic AI programs and robots as if they justified epithets like "emotional", "sad", "surprised", "afraid", "affective", etc. without any deep theory justifying these labels has been debunked and criticized over and over again in the history of AI (e.g., (McDermott 1981)). Yet, it seems that the same habit is still very much alive in AI as it tends to surface in research concerned with "emotional agents".

It is typical and has become practice in the literature of this field to quote emotion research from psychology and neuroscience–interestingly with a strong bias toward a few resources (e.g.,(Damasio 1994) or (LeDoux 1996)), which gives the appearance of "ultimate authorities on emotions", regardless of their actual status in their fields–very little time, if any time at all, is spent on arguing why the respective systems in fact have the required property. It is not uncommon to find statements like (Breazeal 1998) "The robot has several emotion processes" in the recent AI literature on emotions, where it is claimed that a system has a particular property, or that it implements a particular state, without

actually explicating (1) what kind of state it is, (2) why it bears the particular label, (3) how it differs from other states with the same label, (4) how it was implemented, and (5) why the chosen implementation does indeed implement that kind of state.

Another, quite common move in modeling emotions is to "combine" various emotion theories, as exhibited, for example, in (Velśquez 1999), who is "drawing upon ideas from different theorists" in his "Cathexis" system to identify and create "explicit models for six different emotion families". Yet, interestingly, it is rarely reflected whether these combinations actually makes sense, i.e., whether it is possible to combine the respective parts of different theories without obtaining an incompatible, possibly incoherent picture. In other words, what is missing again is (1) an argument as to why particular parts were chosen, (2) that it is indeed possible to combine these parts in a coherent way, and (3) how the resultant combination differs from the original theories. Such an argument is required in particular, if the so-combined theories are mutually incompatible (which different theories about the same subject matter typically tend to be). This move of combining emotional theories usually takes the textual form of "being inspired by" as exemplified by the following passage from (Breazeal 1998) "The organization and operation of the emotion subsystem is strongly inspired by various theories of emotions in humans", which leaves open exactly to what extent the respective theories were followed. The burden of establishing that the resultant "combined" or "inspired" implementation of whatever the original theories call emotions still has anything to do with emotions is usually evaded by not addressing the problem in the first place.

Another common pattern is to (1) overtly acknowledge the psychological theory from which the notion of emotion is borrowed and (2) subsequently deviate from it covertly. (Reilly & Bates 1992), for example, developed the "Em model", which "is based on the OCC (=Ortony, Clore, and Collins) model, but differs in a number of ways that seemed useful or necessary to implement the model and use it to develop a Tok agent, namely Lyotard the cat."[5]

Part of the confusion underlying the discussion about emotions may also be related to the distinction between "having emotions", "having feelings", and "being aware of these emotions and feelings". What often goes by unnoticed is that different architectural requirements underwrite these three properties of agents. "Having emotions" means that certain processes (e.g., of valuation, appraisal, etc.) are supported by the agent architecture, whereas "having feelings" adds another set of architectural requirements to whatever is required for having emotions: for one, the agent has to be able to represent enough (parts of) "itself" within the architecture, and furthermore be capable of processes that monitor these representations in various ways. Even more such processes are required to "be aware of emotions and feeling", in particular, various sort of attention mechanisms, deliberative processes, reflexive processes, etc. (e.g., see (Sloman 2000).

[5]In his dissertation, (Reilly 1996), for example, after summarizing different psychological and philosophical views on emotions, writes "this may be sound psychology, but, as I will discuss, I have chosen to create an explicit emotion system to give artists more direct control over the emotions of their characters", explicitly ignoring these previously mentioned results, while effectively claiming that the characters have emotions nevertheless.

Discussion

It seems that many of today's attempts to create emotions in agents, regardless of the theories of emotions they are based upon or inspired by, take a very high-level causal description of emotional processes and attempt to implement this description more or less directly in agents. However, capturing a causal pattern at a high level of abstraction usually does not capture relevant causal details at lower levels, and does, therefore, not "recreate" the same causal chains that may be responsible for the phenomenon in the first place (mainly because it fails to capture the "counterfactual" properties of the lower-level causal chain). Rather, it only exhibits–in a very crude way–a relationship between higher level states and some lower level states or processes. The left-out details, however, may make all the difference when it comes to actually "having emotions": while a robot can be described as being in a state of "disappointment", which leads to a state of "sadness", these states will bear no resemblance to human states labeled the same if they are, for example, merely implemented as states of a finite automaton (or as activations of connectionists units that are connected by a positive weight for that matter, e.g., see (Velśquez 1999). It is, therefore, of crucial importance for the credibility of emotional models in AI to stay away from anthropomorphizing labels and to be as explicit as possible about the nature of the implemented states, for otherwise we will have no criterion of distinguishing alleged emotional agents from the real ones.

Acknowledgements

I would like to thank Aaron Sloman for many valuable and helpful discussions about and insights into the nature of emotions and agent architectures.

References

Breazeal, C. 1998. Regulating human-robot interaction using 'emotions', 'drives', and facial expressions. In *Proceedings of Autonomous Agents 98*.

Canamero, D. 1997. Modeling motivations and emotions as a basis for intelligent behavior. In *Proceedings of the First International Symposium on Autonomous Agents (Agents'97)*.

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

Goleman, D. 1996. *Emotional Intelligence: Why It Can Matter More than IQ*. London: Bloomsbury Publishing.

Griffiths, P. 1997. *What Emtions Really Are: The Problem of Psychological Categories*. Chicago: Chicago University Press.

Hatano, G.; Okada, N.; and Tanabe, H., eds. 2000. *Affective Minds*. Amsterdam: Elsevier.

Hayes-Roth, B. 1995. Agents on stage: Advancing the state of the art of AI. In *Proc 14th Int. Joint Conference on AI*, 967–971.

LeDoux, J. 1996. *The Emotional Brain*. New York: Simon & Schuster.

McDermott, D. 1981. Artificial intelligence meets natural stupidity. In Haugeland, J., ed., *Mind Design*. Cambridge, MA: MIT Press.

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

Oreskes, N.; Shrader-Frechette, K.; and Belitz, K. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*.

Ortony, A., and Turner, T. 1990. What's basic about basic emotions? *Psychological Review*.

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

Pfeiffer, R. 1988. Artificial intelligence models of emotion. In *Cognitive Perspectives on Emotion and Motivation*. Kluwer Academic Publishers.

Picard, R. 1997. *Affective Computing*. Cambridge, Mass, London, England: MIT Press.

Popper, K. 1962. *Conjectures and refutations; the growth of scientific knowledge*. New York, NY: Basic Books.

Reilly, W., and Bates, J. 1992. Building emotional agents. Technical report, CMU-CS-92-143.

Reilly, N. 1996. *Believable Social and Emotional Agents*. Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Technical Report CMU-CS-96-138.

Sloman, A. 1999. Beyond shallow models of emotion. In Andre, E., ed., *Behaviour planning for life-like avatars*. 35–42. Proceedings I3 Spring Days Workshop March 9th–10th 1999.

Sloman, A. 2000. Architecture-based conceptions of mind. In *Proceedings 11th International Congress of Logic, Methodology and Philosophy of Science*, 397–421. Dordrecht: Kluwer. (Synthese Library Series).

Velśquez, J. 1999. When robots weep: Emotional memories and decision-making. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*.