Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

12 Artificial emotions and machine consciousness

Matthias Scheutz

12.1 Introduction

Over the last decade, interest in artificial emotions and machine consciousness has noticeably increased in artificial intelligence (AI), as witnessed by a number of specialized conferences and workshops dedicated to these themes. This interest is in part based on the recognition that emotions and consciousness have useful roles in humans and other animals, and that understanding these roles and implementing models of them on computers might help in making artificial agents smarter. But can machines even have emotions and be conscious, and if so, how could we go about designing such machines?

The goal of this chapter is to present an overview of the work in AI on emotions and machine consciousness, with an eye toward answering these questions. Starting with a brief philosophical perspective on emotions and machine consciousness to frame the work, the chapter first focuses on artificial emotions, and then moves on to machine consciousness – reflecting the fact that emotions and consciousness have been treated independently and by different communities in AI. The chapter concludes by discussing philosophical implications of AI research on emotions and consciousness.

12.2 The philosophical perspective

Prima facie, it seems that research on emotions and consciousness in AI would have to start from the assumption that it is actually possible to implement emotions and consciousness in computational artifacts. Why else would one bother attempting this goal if it cannot be reached in principle?

It turns out that AI researchers have typically not been impressed with philosophical arguments about the possibility or impossibility of machines replicating human mental states. Rather, they have always pursued a theoretically unencumbered approach to investigating possible algorithms and mechanisms for achieving intelligent behavior. There are basically two main attitudes in AI toward the question whether machines *actually* can have emotions (e.g., like human emotions) or be conscious (e.g., like a normal human adult in waking states). The first is a pragmatic attitude that underlies much of AI research and connects to related attitudes in psychology: Emotion terms and "consciousness" are used in a pragmatic operational way that allows researchers to make progress without having solved all the conceptual problems that beleaguer these concepts. Researchers in AI who are assuming this attitude will look at results from psychology for the types of processes that psychologists take to underlie or be involved in human mental activity and attempt to formalize aspects of them algorithmically. The goal here is not to replicate or model human mentality in a biologically or psychologically plausible way, but rather to use whatever principles could be taken from emotion processes or theories of consciousness to improve the performance of artificial agents (and possibly surpass human performance).

The other attitude is to seek to refine, revise, or replace emotion concepts or concepts of consciousness as a result of attempting to formally specify processes that can implement emotions or bring about consciousness. This attitude is closely aligned with the endeavor of computational modeling in cognitive science, where the goal of a computational model is to replicate human performance while providing mechanisms that explain how humans perform a given task. Consequently, the way algorithms are generated, implemented, and tested has implications for concepts of emotion and consciousness, which in turn will require a philosophical elaboration.

Clearly, the first attitude is sufficient for research goals in AI (e.g., to build intelligent agents), yet the second attitude will also allow AI researchers to connect to other fields and open up their algorithms and implementations to philosophical and psychological scrutiny. That way, psychologists might be able to derive new experimental designs that can test predictions made by the models, and philosophers might be able to sharpen their intuitions about what these concepts are supposed to refer to.

Historically, the questions of whether machines can have emotions or can be conscious have come up at various times in different fields. Here, we will briefly review the philosophical perspectives of two pioneers in AI and philosophy of mind – Alan Turing and Hilary Putnam, respectively.

Alan Turing, in his famous 1950 paper "Computing machinery and intelligence" (Turing 1950) considers nine objections to his "imitation game," which has subsequently become known as the "Turing Test."[1] The fourth of these, the "Argument from Consciousness," attempts to dismiss machine intelligence by pointing to the lack of emotions and feelings in machines. Here, Turing cites Professor Geoffrey Jefferson as stating that

[1] This an envisioned setup where a human subject has to interact via a chat-like computer interface with two other participants, a human and a machine, without knowing which is which. The subject's goal is to determine which of the participants is human and which the computer within a given time period through natural language interactions.

> Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain – that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants. (Quoted in Turing 1950, pp. 445–6).

Turing diagnoses this line of argument as ultimately promoting a solipsistic perspective where "the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking" (Turing 1950, pp. 446). He points out that the same line of argument would then also hold for people (i.e., one could only be sure that another person has certain mental properties or is in a particular mental state if one *were* that other person), a problem known in philosophy as the "other minds problem." In other words, he reduces the other minds problem for machines to the other minds problem for humans.

Moreover, he points out that a sonnet-writing machine that gives reasonable answers to an interrogator about its own sonnet using *viva voce* (and thus presumably using intonation in a humanly plausible way, including the expression of emotions) would likely not be viewed as an "easy contrivance." The assumption here is that a machine that can interact in natural spoken language in human-like ways would cause people to view it as having pleasure, pain, and so on, in very much the way people infer internal states of other people based on their interactions (e.g., from the tone in a person's voice).

The question of whether machines can have feelings and can be conscious has been revisited in detail by Hilary Putnam. In his 1964 paper "Robots: Machines or artificially created life?" (Putnam 1964), Putnam wants us to imagine the robot Oscar which is *psychologically isomorphic* to a human – that is, which has internal states that play the same causal roles as our mental states do. Suppose Oscar is having the "sensation" of red in this sense, then the question arises whether it is *really* having a sensation of red, that is, whether Oscar is actually *seeing* anything, whether Oscar is feeling, whether Oscar is conscious. Like Turing, Putnam links this question to the other minds problem: "Whether, and under what conditions, a robot could be conscious is a question that cannot be discussed without at once impinging on the topics that have been treated under the headings Mind–Body Problem and Problem of Other Minds" (p. 669). After dispelling several objections to the claim that Oscar is conscious, he concludes that this question

> calls for a decision and not for a discovery. If we are to make a decision, it seems preferable to me to extend our concept so that robots *are* conscious – for "discrimination" based on the "softness" or "hardness" of the body parts of a synthetic "organism" seems as silly as discriminatory treatment of humans on the basis of skin color. (p. 691)

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

Turing and Putnam's view that machines can be conscious in principle has since been echoed by various philosophers (e.g., Lycan 1987). In all cases the assumption is that machines will have to have the right kind of internal structure and cognitive organization – the right type of architecture – for them to be able to have emotions and to be conscious (whether they then will *actually* instantiate emotions and/or be conscious will depend on additional factors, as in the human case). The question about the right kind of architecture that can implement emotions and consciousness, however, is exactly what research in AI has attempted to tackle.

12.3 Emotions in AI

Different forms of emotions have been studied to varying degrees ever since the beginning of AI (e.g., Pfeifer 1988), despite the original focus of AI on deliberative, non-emotional mechanisms. More recently, however, work on emotions and emotional agents has become much more mainstream, not least due to Aaron Sloman's work on emotional architectures (Sloman and Croucher 1981) and Ross Picard's work on "affective computing" (Picard 1997), which stressed the importance of human affect and explored how computers can be made "affect-aware" or emotional. Today, we witness growing numbers of research communities that investigate aspects of emotion and affect, from "emotional" or "affective" user interfaces to "believable" synthetic characters and life-like animated agents with emotions, to emotional or emotion-aware pedagogic and instructional agents, to emotional virtual agents and robots (see Trappl, Petta, and Payr 2002 for an overview).

The motivations for the various research directions and their specific aims are naturally quite different. While for some emotions are about making animated characters more believable (e.g., by endowing them with emotional facial expressions), for others recognizing emotions is crucial for a system to be able to adapt to its user's needs. Yet others take emotions to be an integral part of the control of complex agents, and thus focus on architectural mechanisms that are required for emotion processes. But common to all these different incentives for exploring emotions is the tacit assumption that emotions, in one form or another, may have important applications in artificial agents.

12.3.1 Functional roles of emotions

One major difficulty connected to concepts such as emotion (and consciousness as well) is that they are not clearly specified, and likely not even clearly specifiable in principle. Hence, there is no clear sense in psychology of exactly what an emotion is (Griffiths 1997), and psychological accounts vary greatly

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

as to how emotions are individuated (e.g., based on facial expressions, behavioral patterns, brain regions, etc.). The conceptual difficulties with emotion concepts, however, have not been a deterrent to attempts to implement processes that at least resemble emotion processes, even though researchers in AI often disagree on what they take "emotion" to be and what they believe it means to *implement emotions* in artificial agents (e.g., Scheutz 2002).

Much research on the role of emotions in artificial agents has been motivated by an analysis of possible *functional* roles of emotions in natural systems. The underlying assumptions are that (1) emotions have functional roles in agent architectures, and that (2) having states with the right functional roles is sufficient for having emotions, independent of the particular physical makeup of the agent. While most researchers in the affective sciences will agree on (1) (even though there are many examples of the effects of *dysfunctional affect* as well), their views diverge on (2) – whether having the right kind of functional architecture is *all* there is to having a particular emotion. For example, they might hold that various bodily processes are involved in many affective states: If particular biochemical processes, such as the secretion of particular hormones, or changes in particular neurotransmitters, are taken to be essential to, or constitutive of, affect, then artificial agents will, by definition, not be capable of instantiating affective states so construed. (Compare the views some philosophers have voiced about consciousness or qualitative states, e.g., Searle 1992.) Artificial agents will, however, still be capable of instantiating the same kinds of *control processes* as those implemented in neural activity in animals, since these are, also by definition, independent of the physical makeup of an agent, and this may be sufficient for AI purposes (e.g., for an artificial agent to be able to perform a particular task). If, on the other hand, the exact nature of bodily states and processes does not play a causal role in the functioning of affect processes, so that, for example, simulated hormonal systems could be used to achieve the same effects (e.g., Cañamero 1997), then artificial agents will be able to instantiate affect processes if they have the right architectural prerequisites.

Regardless of what stance one takes on the *qualitative nature* of emotions (i.e., on the question of "what it is like to experience state X"), the functional aspects of emotions in the context of an agent's control system can be independently considered. In particular, there seem to be twelve potential roles of emotions for artificial agents (see also Scheutz 2004):

1. *Alarm mechanisms* – e.g., fast reflex-like reactions in critical situations, such as fear processes, that interrupt current behavior and initiate a retreat response, moving the agent away from the danger zone.
2. *Action selection* – e.g., deciding what to do next based on the current emotional state, such as switching from exploration to foraging behavior based on the agent's needs.

3 *Adaptation* – e.g., short- or long-term changes in behavior due to affective states, such as adapting one's gait to uneven terrain based on negative affect generated by sensors.
4 *Social regulation* – e.g., using emotional signals to achieve social effects, such as aggressive display to deter another agent from interfering with one's activity.
5 *Learning* – e.g., using affective evaluations as utility estimates in reinforcement learning, such as learning the utility of different behaviors to achieve goals in different contexts.
6 *Motivation* – e.g., adopting goals as part of an emotional coping mechanism, such as when a high level of distress and frustration leads to adopting the goal of asking a human supervisor for help)
7 *Goal management* – e.g., the creation of new goals or reprioritization of existing ones, such as using positive and negative affect to modify cost estimates used in the calculation of the expected utility of a goal.
8 *Strategic processing* – e.g., the selection of search strategies based on overall affective state, such as using positive and negative affect to bias search algorithms to top-down versus bottom-up search.
9 *Memory control* – e.g., the strategic use of affective bias on memory access and retrieval as well as decay rate of memory items, such as using current affective state to rank memory items with similar affect as better matches.
10 *Information integration* – e.g., emotional filtering of data from various information channels or blocking of such integration, such as ignoring positively valenced information from vision sensors about a happy face when the acoustic information suggests an angry voice.
11 *Attentional focus* – e.g., selection of data to be processed based on affective evaluation, such as biasing visual search in favor of objects the agent highly desires.
12 *Self model* – e.g., using affect as a representation of "what a situation is like for the agent," such as using the overall affective evaluation of different components of the agent's control system as a measure of the agent's overall mood and how it "feels".

While this list is not intended to be exhaustive, it does point to the varied functional nature of emotions, from architectural roles to roles in social regulation. And it provides a frame within which to locate past accomplishments and future directions in research on architectural aspects of affect.

12.3.2 Communicative vs. architectural aspects of emotion

Work on emotions in AI can be roughly divided into two strands (with a small overlap): *communicative aspects* and *architectural aspects*.

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

Communicative aspects of emotions are mostly concerned with the fourth role (social regulation) and have been explored mainly by the human–computer interaction (HCI) and, more recently, the human–robot interaction (HRI) communities. Efforts focus on emotion recognition, emotional expression, and sometimes on how to connect the two to improve the experience of human users with an interactive system (e.g., via the user interface on a computer or via the sensory and effector repertoire of a robot; Scheutz, Schermerhorn, and Kramer 2006). Both communities have made important advances in understanding the kinds of emotional interactions people engage in (e.g., Brave and Nass 2003) and how to make machines recognize and signal them (e.g., how to explore temporal patterns to detect frustrated vs. delighted human smiles; Hoque, McDuff, and Picard 2012).

The second main strand, the architectural aspect of emotion, has focused on the role and utility of emotions in agent architectures (such as using emotional evaluations as quick heuristics in decision making) and is thus less concerned with the social communicative aspects of emotions. This strand attempts to use emotion mechanisms to improve the agent's capabilities, and most work here has focused on the first five roles. In particular, attention has been given to affective or emotional action selection, both in simulated agents (e.g., Gadanho 2003) and robotic agents (Murphy et al. 2002). Similarly, quite a bit of work has investigated the utility of evaluations that are internally generated and reflect some aspect of the agent's internal state (rather than external environmental states) for reinforcement learning, even though most of these investigations do not call these evaluations "affective" (e.g., Ichise, Shapiro, and Langley 2002). Yet, surprisingly little work has focused on investigating roles (6) through (12), although there are some notable exceptions (e.g., Gratch and Marsella 2004a). Note that especially the last four roles might turn out to be critical for reflective, and thus conscious, systems (e.g., as described in Sloman and Chrisley 2003). For, as we shall see in Section 12.4 on machine consciousness, mechanisms for attentional control, information integration, working memory and its access control, and an agent's self-model are all taken to be essential ingredients for developing conscious machines.

There are several crucial differences between research on the communicative and the architectural aspects of emotions. Most importantly the former does not require the instantiation of emotional states within a system. For example, an agent does not have to be itself emotional (or capable of emotions) to be able to recognize emotional expressions in human faces. The latter, on the other hand, must claim that emotional states of a particular kind are instantiated within the system. Moreover, researchers on the communicative aspects of emotions do not need a satisfactory *theory of emotion* (i.e., a theory of what emotional states are) to be able to produce working systems. Being able to measure changes in a user's skin conductance, breathing frequency, and so on and using this information to change the level of detail in a graphical

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

user interface does not automatically commit one to claiming that what was measured was the user's *level of frustration*, even though this seems to be true in some cases. In fact, a pedagogical agent might learn important facts about its user (e.g., the effectiveness of its instructional strategies) based on such measures without requiring any representation of the user's emotional processes nor any emotional processes itself.

Contrariwise, architectures that claim to use emotional mechanisms (e.g., for the prioritization of goals or for memory retrieval) will have to make a case that the implemented mechanisms indeed give rise to "emotional states" in a clearly specified sense. Otherwise there is no sense, nor any reason, to call them that, even though there is, and always has been, a tendency in AI to present simplistic AI programs and robots as if they justified epithets like "emotional," "sad," "surprised," "afraid," "affective," and so on, without any deep theory justifying these labels (e.g., McDermott 1981). Consequently, the architectural route faces the challenge of saying exactly what it means to "implement emotional states" of the kinds in question.

Researchers pursuing the architectural strand on emotions in AI can be further divided into two main categories: those who attempt to model overt, observable effects of emotion behavior (call these *display models* of emotions), and those who aim to model the internal processes that bring about emotional behavior (call these *process models* of emotion).

Most work on architectural aspects of emotion in AI to date has focused on display models, which are intended to get the "input–output mapping" of a given behavioral description right (e.g., the right kind of emotional response for a given context, such as a fear expression on a robot's face when there is a rapidly approaching object in front of it). In the extreme case, such a mapping could be as simple as that employed in an animated web-based shopping agent which displays a surprised face if the user attempts to delete an item from the shopping basket. Architectures of this kind are found in many so-called "believable agents," where the primary goal is to induce a human observer to think that the agent is in a particular emotional state (see, e.g., Bates, Loyall, and Reilly 1994 for simulated agents, and Murphy et al. 2002 for robots). Whether the agent is indeed in the particular state is irrelevant. In fact, emotions are here often represented as states or values of "emotion variables," either qualitatively, as suggested by emotion terms (e.g., "happy", "afraid", etc.), or quantitatively, using numeric values (e.g., the agent is "0.4 happy," "0.1 afraid," etc.). And while some allow agents to be in only one state at a time, others allow for "emotion blends" (mixtures of simultaneously present emotional states), where individual emotions and their intensities span a multi-dimensional space.

Note that these features should not be taken to imply that the design of the architecture was devoid of biological motivation. Quite the opposite is true: Most (if not all) display models derive their inspiration from research in the affective sciences. However, their goal is not to replicate any particular empirical data from animal or human research, but rather to explore possible mechanisms for yielding the desired observable effects.

The main problem with display models of emotions is that they are ultimately silent about the role of emotions in agent architectures, for they may or may not *actually* implement emotional processes to achieve the desired overt behaviors. And even if they do, they may tell us little about the role of emotions. For although the implemented states are often labeled with familiar terms, they differ significantly from those usually denoted by these terms. A state labeled "surprise," for example, may be functionally defined to be triggered by loud noises and have very little in common with the complex processes underlying notions of "surprise" in humans and various animals, which involve the violation of a predicted outcome. (For a state so defined, "startle" would be the more appropriate label.)

In contrast, process models are intended to model and simulate some aspects of emotional processes as they unfold. As many psychologists and AI researchers have pointed out (e.g., Pfeifer 1988; Cosmides and Tooby 2004), emotion concepts are best characterized as denoting enduring processes of *behavior control*: action and reaction, adjustment and modification, anticipation and compensation of behavior in various (frequently social) situations. Often it is not a single inner state of an agent architecture that determines whether an agent experiences or displays some emotion, but rather a whole sequence of such states in combination with environmental states. "Fear," for example, does not refer to the makeup of an agent at a particular moment in time but to the unfolding of a sequence of events, starting from the perception of a potentially threatening environmental condition, to a reaction of the agent's control system, to a reaction of the agent's body, to a change in perception, and so on. Process models are thus much more complex than display models since they focus on the internal processes (and processing states) involved in emotions, typically drawing on a (psychological, neurological, etc.) theory of emotion (Panksepp 1998; Ortony, Clore, and Collins 1988).

12.3.3 Process models of emotion

Process models are based on the various components that are characteristic of emotion processes: a perceptual component that can trigger the emotion process; a visceral component that affects homeostatic variables of the agent's body; a cognitive component that involves belief-like states as well as various kinds of deliberative processes (e.g., redirection of attentional mechanisms, reallocation of processing resources, recall of past emotionally charged episodes, etc.); a behavioral component that is a reaction to the affect process (e.g., in the form of facial displays, gestures or bodily movements, etc.); and an accompanying qualitative feeling ("what it is like to be in or experience state S"). No single aspect is necessary for emotion, nor is any single aspect sufficient on its own. Yet, most of them are taken to be part of the many forms of human emotions we know from our own experience.

Process models themselves can be categorized into two main classes, based on whether they are aimed at explaining low-level neurological structures and mechanisms of emotion ("low-level process models"), or whether they are intended to model higher-level emotion processes ("high-level process models"). Most research on low-level process models is concerned with Pavlovian conditioning and is targeted at neural structures and processing mechanisms (hence, most low-level models are "neural network" models). Higher-level models of emotions are intended to capture more cognitive aspects involved in affect processes and are typically concerned with a wider range of affect (hence, most higher-level models are "symbolic" models).

The most extensively developed general low-level models are Grossberg's *CogEM* models (e.g., Grossberg and Schmajuk 1987), which are intended to show interactions between emotional and non-emotional areas in the brain (e.g., the amygdala vs. the sensory or prefrontal cortices). *CogEM* models can account for several effects in Pavlovian fear conditioning, but have not been directly applied to empirical data.

Specific low-level affect models, on the other hand, are targeted at modeling the amygdala, which performs several functions in emotion processing (LeDoux 1996). The lateral amygdala, for example, has been shown to be involved in fear conditioning (Blair, et al. 2003), and a preliminary computational model of associative learning in the amygdala has been developed and tested in three associative learning tasks (Balkenius and Morén 2001). Moreover, recent evidence from studies with rats suggests that the amygdala, in particular the frontotemporal amygdala, integrates sensory information and encodes affective evaluations as part of fear memory (Fanselow and Gale 2003). LeDoux and colleagues have hypothesized a dual pathway model of emotional processing in the amygdala, which they tested in auditory fear conditioning studies (LeDoux 1996). These models have been also used in simulated lesion studies and successfully compared to data from actual lesion studies with rats.

While all low-level models are neural network models, higher-level models comprise both connectionist and symbolic approaches. An example of a high-level connectionist approach is the ITERA model (Nerb and Sperba 2001), which is designed to study how media information about environmental problems influences cognition, emotion, and behavior. Facts, input types, emotions, and behavioral intentions are all represented in terms of individual neural units that are connected via excitatory and inhibitory links and compete for activation.

Most attempts to model emotions at higher levels, however, are based on symbolic architectures, for example, Soar (Laird, Newell, and Rosenbloom

1987) or ACT (Anderson 1993). They typically focus on the *OCC model* (Ortony et al. 1988), which provides “update rules” for changes in emotional states that can be directly implemented in rule-based systems. The currently most advanced implementations of high-level affect models are effected in the context of the “virtual humans” (Rickel et al. 2002), where the utility of emotions in artificial agents can be investigated in full immersion interactions with people (Gratch and Marsella 2004b). One particular model, the EMA model (Gratch and Marsella 2009), has also been used to further psychological theories that posit different “emotional appraisal and coping” processes as essential parts of human emotion processes.

Other higher-level architectures attempt to implement different aspects of psychological theories of emotions; examples include the MAMID model, whose emotional components “anger” and “fear” follow Frijda’s definition (Frijda 1994), and the model of “surprise” suggested by Macedo and Cardoso (2001). There are also a few conceptual suggestions for complex human-like architectures that explicitly incorporate human-like emotion and cognition, but without providing particular implementations of the proposed architecture. Examples include Sloman’s *H-CogAff model*, Minsky’s *emotion machine*, and Norman, Ortony, and Revelle’s 3-tier model.

Most emotion models have been implemented and tested in isolation from any *body model*. Consequently, it is difficult, if not impossible, to investigate crucial aspects of emotion processing that need a body to control and thus go beyond functional properties (like the effects of Pavlovian conditioning), which can be tested in stand-alone models (e.g., by applying a stimulus and measuring the output). Various attempts have been made to include bodily processes in simulated and robotic agents. Some have investigated the computational effects of simulated hormones for emotional control (Cañamero 1997), while others have implemented connectionist emotion models on robots, where different emotion types are represented as connectionist units that compete for activation, which in turn cause the robot to exhibit a particular behavior (e.g., Velásquez 1999). The main difference between these approaches and both low-level models of affect and some high-level appraisal-based models (e.g., Gratch and Marsella 2004a, 2009) is that they do not attempt to model any *specific psychological* or *neurobiological* theory of affect (e.g., in an effort to verify or falsify its predictions). Rather, they are concerned with the applicability of a particular control mechanism from an engineering perspective.

The main problem with process models of affect is a direct result of the problems plaguing affect concepts: It is unclear what kind of affective state a particular computational model is a model *of*. In some sense, process models without a functional characterization of the implemented affective states are no more successful from a conceptual point of view than display models which are not intended to implement specific kinds of affective states in the first place. However, even if no conceptual mileage is to be gained from a process model right away, there is an important advantage to the methodological approach of attempting to implement hypothesized affect mechanisms that has borne fruit already in the short term. For the architectural mechanisms intended to allow for the instantiation of affective states can be tested and evaluated as such, regardless of what kinds of functional states they can instantiate (e.g., one could treat them as "quasi-emotions" and investigate their potential for improving an agent's performance; Scheutz 2011). This is analogous to what happened pragmatically within AI with other kinds of architectures, such as belief-desire-intention (BDI) architectures, for example. Here the same kinds of conceptual questions could be raised about the *actual nature* of the instantiated "belief," "desire," and "intention" states, while the architectural mechanisms for problem solving could be evaluated independently in different domains for their technical merit.

Yet, there is an important difference between architectural approaches in the domain of reasoning, problem solving, and so on, and architectural approaches in the domain of affect: The former has often a well-developed theory of the functional potential of the architectural mechanisms, while the latter has currently no such theory. Rather, research on architectural aspects of affect is still in a *pre-theoretic* stage. The current lack of a well-developed theory of the utility of affective states in the control of artificial agents, however, does not take away from the fact that attempts to characterize and implement affective states and processes might yield architectural mechanisms that could prove useful for a variety of domains and applications (e.g., applications that have to deal with severe resource constraints as argued in Scheutz 2001b).

12.4 Machine consciousness in AI

Unlike emotion research, which dates back to the 1960s, investigation of machine consciousness in AI is a much younger endeavor that started in the mid 1990s and is really only beginning to gain momentum (although there were some early attempts at laying out requirements for conscious machines; see, e.g., Angel 1989).

One of the reasons for this later start may be that research on conscious machines must build on research on the various functional components that are required for consciousness, some of which may be emotions (for the suggestion that emotions and consciousness are intrinsically linked, see, e.g., Alexandrov and Sams 2005). Somewhat surprisingly, however, the machine consciousness community is not a subset of the emotion community in AI, nor does it intersect much with it. And while the emotion community in AI has fostered close ties to various psychologists and their theories (e.g., Andrew Ortony and Craig Smith, among others), the machine consciousness

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

community seems to be more connected to philosophers who are interested in giving a functional, implementable account of consciousness.

Similarly to the case of emotions in AI, where researchers working on the communicative and other dimensions of emotion simply ignore questions about what emotions are and how they are implemented, some researchers interested in consciousness are not attempting to give an account of human consciousness. Rather, they are interested in "simulating" processes they take to be essential to consciousness – what (Holland 2003) calls "weak artificial consciousness" – or using principles underwriting human consciousness to design better control systems (Sanz, López, and Hernández 2007). Some, however, are interested in *conscious machines* (Franklin, Kelemen, and McCauley 1998; Aleksander and Dunmall 2003), and thus, like researchers on process models of emotions, have to address the question of what they mean by "consciousness" and, eventually, what it would take to implement it. Clearly, this is a very difficult problem, given that neither philosophers nor psychologists agree on what "consciousness" is supposed to refer to or what it is to *be conscious*. (Theories of consciousness range from neurological theories to cognitive representational theories, such as the various forms of *higher-order thought* theory, which hold that thoughts and perceptions become conscious in virtue of being targeted by further thoughts or perceptions.) As with emotions, AI researchers interested in achieving consciousness in machines have proposed various principles and architectural mechanisms that they take to be necessary for conscious machines.

In general, proposals vary along several dimensions: (1) the extent to which they connect to philosophy, psychology, or neuroscience; (2) the extent to which they lay out a particular architecture that can be conscious, or particular principles for such an architecture; and (3) the extent to which they actually provide implementations of their architectures or models. However, researchers agree that some type of "inner model" is required that is based on representations of the agent's perceptual states and allows the agent to simulate or predict future events and outcomes and what various possible actions would be like for it. Researchers disagree, however, on the exact definition and extension of the internal model and the other components to which it is connected.

12.4.1 Architectural proposals

Most proposals on consciousness in artificial agents are conceptual at present and provide a set of potentially implementable principles (sometimes with preliminary implementations for subsets). Pentti Haikonen, for example, summarizes the architectural requirements for a conscious system as follows:

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

> (1) A suitable method for the representation of information must be devised. (2) Suitable information processing elements that allow the manipulation of information by the chosen representation method must be designed. (3) A machine architecture that can accommodate sensors, effectors, the processes of perception, introspection and the grounding of meaning as well as the flow of inner speech and inner imagery must be designed. (4) The system design must also accommodate the functions of thinking and reasoning, emotions and language. (Haikonen 2003, p. 168)

A more formal approach is taken by Aleksander and colleagues, who list five principles, stated as axioms, that are taken to be sufficient for consciousness. They specify the notion of "conscious of," for an agent and a world, as follows:

> Let *A* be an agent in a sensorily-accessible world *S*. For *A* to be conscious of *S* it is necessary that:
>
> Axiom 1 (Depiction): *A* has *perceptual states* that depict parts of *S*.
> Axiom 2 (Imagination): *A* has internal *imaginational states* that *recall* parts of *S* or *fabricate S*-like sensations.
> Axiom 3 (Attention): *A* is *capable of selecting* which parts of *S* to depict or what to imagine.
> Axiom 4 (Planning): *A* has means of control over imaginational state sequences *to plan actions*.
> Axiom 5 (Emotion): *A* has additional *affective states* that evaluate planned actions and determine the ensuing action.
>
> (Aleksander and Dunmall 2003, p. 9)

The claim is that this combination of sensory, imaginational, attentional and affective depictions is what ultimately leads to a first-person perspective (the "I" in humans). The axioms are motivated, not by a particular theory of consciousness, but by a large collection of individual findings that seem to suggest these principles as abstractions.

Sloman has for quite some time promoted the notion of "virtual machine functionalism" as a way to account for rich internal processes of complex, deliberative and reflective agents that might form the basis of introspection and the development of internal categories and concepts that are not accessible (even via language) to other agents, and thus form the basis of a conscious agent's first-person perspective (e.g., Sloman and Chrisley 2003). There are also several other researchers who are attempting to give functional architectural accounts of the requirements for consciousness. Proposed accounts range from neural (Shanahan 2005), to robotic (Kuipers 2005), to control-theoretic (Sanz et al. 2007), to process-based (Manzotti 2003), and others. Common to all of the above researchers is that they have implemented some rudimentary models

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

that demonstrate parts of the architecture, but not a complete functional, and thus conscious, system.

12.4.2 Conscious agents

A notable exception among researchers in machine consciousness is the work of Franklin and colleagues (Franklin et al. 1998), who have attempted to implement a complete conscious agent, based on Baars' global workspace theory of consciousness. This is a "theater model" of consciousness, which requires a central workspace (the "stage") where "conscious contents emerge when the bright spotlight of attention falls on a player on the stage of working memory" (Baars 1997, p. 44).

The first functional prototype, *"Conscious" Mattie*, was a software agent charged with writing seminar announcements, communicating by email with seminar organizers, and reminding them when were late. A second prototype, IDA for "Intelligent Distribution Agent," was developed for the US Navy to facilitate the process of assigning sailors to new missions. Both architectures include mechanisms for "consciousness," comprising a spotlight controller, a broadcast manager, and a collection of attention codelets which recognize novel or problematic situations, together with modules for perception, action selection, associate memory, emotions, and meta-cognition (see Franklin 2000). The latest model is a complete cognitive architecture called LIDA (Learning Intelligent Distribution Agent), which adds various types of learning to the previous architecture.

12.5 Future perspectives

Emotion research has become an active interdisciplinary subfield in AI, and machine consciousness is on the verge of establishing a research community that pursues the design of conscious machines. Based on the current trajectories, it is likely that both communities will grow together, especially as the emotion community is pursuing more complex emotions, such as regret about one's own behavior or disappointment in someone else's attitude toward one, that require many of the architectural features necessary for conscious machines, as postulated by the consciousness community (representations of one's perceptions, internal focus of attention, memories of past actions, representations of possible futures, etc.).

Research in both areas promises not only to advance the state of the art in AI, but also to shed light, if not directly on the human case, then on the case of possible emotional and conscious beings, which should help us refine our concepts. Moreover, both areas are likely to contribute to a better understanding of the trade-offs between systems that are emotional and conscious

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

compared to systems that lack one or both properties. Given that both endeavors are fairly young, however, it should not be too surprising that the fields have neither worked out satisfactory criteria for success nor reflected on the implications of their work. “Criteria for success” here is intended to refer to ways that would allow us to tell whether a given machine has emotions or is conscious. Presumably, this will involve claims about the machine’s functional architecture and the types of states that it supports. This would also include algorithms to determine whether a given system actually implements the functional architecture, but unfortunately we are currently also missing a good theory of implementation (Scheutz 2001a). Ideally, we would like to have criteria that can establish whether a given machine is in a particular emotional state or is conscious. This could involve procedures analogous to those psychologists use to determine whether a person is in a particular emotional state or is conscious.

While the specific need for such criteria might not arise as much within AI itself, it is likely that there will eventually be strong societal pressure to settle these and other fundamental questions about the nature of artificial minds, especially when claims are made about the emotional and conscious states of machines. This was a point recognized by Puntam over forty years ago:

> Given the ever-accelerating rate of both technological and social change, it is entirely possible that robots will one day exist, and argue “we *are* alive; we *are* conscious!” In that event, what are today only philosophical prejudices of a traditional anthropocentric and mentalistic kind would all too likely develop into conservative political attitudes. But fortunately, we today have the advantage of being able to discuss this problem disinterestedly, and a little more chance, therefore, of arriving at the correct answer. (Putnam 1964, p. 678)

While Putnam was certainly right about the need to clarify questions about machine consciousness, the urgency for working out answers to the problem has clearly changed between when he wrote about discussing it “disinterestedly” and today, with all the recent successes in artificial intelligence and autonomous robotics, and with robots already being disseminated into society. Hence, it is high time for AI researchers and philosophers to reflect together on the potential of emotional and conscious machines. For we do not want to wake up one day to discover that what we treated as emotionless, non-conscious artifacts were really emotional, conscious beings, enslaved and mistreated by us out of ignorance or prejudice.

Further reading

Scherer, K. R., Bänziger, T., and Roesch, E. B. (2010). *Blueprint for Affective Computing: A Sourcebook*. Oxford University Press. A comprehensive collection of research chapters on the various aspects of emotions and current emotion models, ranging from theoretical frameworks to specific algorithms for implementing affectively competent artificial agents.

Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press. A great foray into the problems associated with building intelligent autonomous robots and an appeal to implement moral decision making in artificial agents.

The International Journal of Synthetic Emotions (IGI). A good resource for research papers on different models and implementations of artificial emotions.

The International Journal of Machine Consciousness (World Scientific). A great resource for the latest research papers on the emergent field of machine consciousness.

References

Aleksander, I. and Dunmall, B. (2003). Axioms and tests for the presence of minimal consciousness in agents, in O. Holland (ed.), *Machine Consciousness* (pp. 7–18). New York: Imprint Academic.

Alexandrov, Y. I. and Sams, M. E. (2005). Emotion and consciousness: Ends of a continuum, *Cognitive Brain Research* 25: 387–405.

Anderson, J. R. (1993). *Rules of the Mind*. Mahwah, NJ: Erlbaum.

Angel, L. (1989). *How to Build a Conscious Machine*. Boulder, CO: Westview Press.

Baars, B. J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. New York: Oxford University Press.

Balkenius, C. and Morén, J. (2001). Emotional learning: A computational model of the amygdala, *Cybernetics and Systems* 32: 611–36.

Bates, J., Loyall, A. B., and Reilly, W. S. (1994). An architecture for action, emotion, and social behavior, in C. Castelfranchi and E. Werner (eds.), *Artificial Social Systems: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW '92)* (pp. 55–68). Berlin: Springer.

Blair, H. T., Tinkelman, A., Moita, M. A. P., and LeDoux, J. E. (2003). Associative plasticity in neurons of the lateral amygdala during auditory fear conditioning, *Annals of the New York Academy of Sciences* 985: 485–7.

Brave, S. and Nass, C. (2003). Emotion in human–computer interaction, in J. A. Jacko and A. Sears (eds.), *The Human–Computer interaction Handbook: Fundamentals, Evolving Techbologies, and Emerging Applications* (pp. 81–96). Mahwah, NJ: Erlbaum.

Cañamero, D. (1997). Modeling motivations and emotions as a basis for intelligent behavior, in W. L. Johnson (ed.), *Proceedings of the First International Conference on Autonomous Agents (agents'97)* (pp. 148–55). New York: ACM Press.

Cosmides, L. and Tooby, J. (2004). Evolutionary psychology and the emotions, in M. Lewis and J. M. Haviland-Jones (eds.), *Handbook of Emotions* (2nd edn.) (pp. 91–115). New York: Guilford Press.

Fanselow, M. S. and Gale, G. D. (2003). The amygdala, fear, and memory, *Annals of the New York Academy of Sciences* 985: 125–34.

Franklin, S. (2000). Modeling consciousness and cognition in software agents, in N. Taatgen, J. Aasman (eds.), *Proceedings of the 3rd International Conference on Cognitive Modeling* (pp. 100–9). Veenendal, The Netherlands: Universal Press.

Franklin, S., Kelemen, A., and McCauley, L. (1998). Ida: A cognitive agent architecture, in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3 (pp. 2646–51).

Frijda, N. H. (1994). Varieties of affect: Emotions and episodes, moods, and sentiments, in P. Ekman and R. J. Davidson (eds.), *The Nature of Emotion: Fundamental Questions* (pp. 59–67). New York: Oxford University Press.

Gadanho, S.C. (2003). Learning behavior-selection by emotions and cognition in a multi-goal robot task, *Journal of Machine Learning Research* 4: 385–412.

Gratch, J. and Marsella, S. (2004a). A domain-independent framework for modeling emotion, *Cognitive Systems Research* 5: 269–306.

(2004b). Evaluating the modeling and use of emotion in virtual humans, in *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 1 (pp. 320–7).

(2009). EMA: A process model of appraisal dynamics, *Cognitive Systems Research* 10: 70–90.

Griffiths, P. E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago University Press.

Grossberg, S. and Schmajuk, N. (1987). Neural dynamics of attentionally-modulated Pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing, *Psychobiology* 15: 195–240.

Haikonen, P. O. (2003). *The Cognitive Approach to Conscious Machines*. Exeter: Imprint Academic.

Holland, O. (ed.) (2003). *Machine Consciousness*. New York: Imprint Academic.

Hoque, M., McDuff, D., and Picard, R. (2012). Exploring temporal patterns towards classifying frustrated and delighted smiles, *IEEE Transactions on Affective Computing* 3: 323–34.

Ichise, R., Shapiro, D. G., and Langley, P. (2002). Learning hierarchical skills from observation, in *Proceedings of the 5th International Conference on Discovery Science* (pp. 247–58).

Kuipers, B. (2005). Consciousness: Drinking from the firehose of experience, in *Proceedings of the 20th National Conference on Artificial Intelligence*, vol. 3 (pp. 1298–305).

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence, *Artificial Intelligence* 33: 1–64.

LeDoux, J. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.

Lycan, W. G. (1987). *Consciousness*. Cambridge MA: MIT Press.

Trim: 247mm × 174mm Top: 12.653mm Gutter: 21.089mm
CUUK2567-12 CUUK2567/Frankish ISBN: 978 0 521 87142 6 February 14, 2014 13:39

Macedo, L. and Cardoso, A. (2001). Modeling forms of surprise in an artificial agent, in J. Moore and K. Stenning (eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 588–93). Mahwah, NJ: Erlbaum.

Manzotti, R. (2003). A process-based architecture for an artificial conscious being, in J. Seibt (ed.), *Process Theories: Crossdisciplinary Studies in Dynamic Categories* (pp. 285–312). Dordrecht: Kluwer Academic Press.

McDermott, D. (1981). Artificial intelligence meets natural stupidity, in J. Haugeland (ed.), *Mind Design: Philosophy, Psychology, Artificial Intelligence* (pp. 143–60). Cambridge, MA: MIT Press.

Murphy, R. R., Lisetti, C., Tardif, R., Irish, L., and Gage, A. (2002). Emotion-based control of cooperating heterogeneous mobile robots, *IEEE Transactions on Robotics and Automation* 18: 744–57.

Nerb, J. and Sperba, H. (2001). Evaluation of environmental problems: A coherence model of cognition and emotion, *Cognition and Emotion* 4: 521–51.

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

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions.* Oxford University Press.

Pfeifer, R. (1988). Artificial intelligence models of emotion, in V. Hamilton, G. H. Bower, and N. H. Frijda (eds.), *Cognitive Perspectives on Emotion and Motivation* (pp. 287–320). Dordrecht: Kluwer Academic Publishers.

Picard, R. (1997). *Affective Computing.* Cambridge, MA: MIT Press.

Putnam, H. (1964). Robots: Machines or artificially created life? *The Journal of Philosophy* 61: 668–91.

Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D., and Swartout, W. (2002). Towards a new generation of virtual humans for interactive experiences, *IEEE Intelligent Systems*, 17(4): 32–8.

Sanz, R., López, I., and Hernández, C. (2007). Self-awareness in real-time cognitive control architectures, in A. Chella and R. Manzotti (eds.), *AI and Consciousness: Theoretical Foundations and Current Approaches: Papers from the AAAI Fall Symposium* (pp. 135–40). Menlo Park, CA: AAAI Press.

Scheutz, M. (2001a). Causal versus computational complexity, *Minds and Machines* 11: 534–66.

(2001b). The evolution of simple affective states in multi-agent environments, in D. Cañamero (ed.), *Proceedings of AAAI Fall Symposium* (pp. 123–8). Falmouth, MA: AAAI Press.

(2002). Agents with or without emotions? in R. Weber (ed.), *Proceedings of the 15th International Florida Artifical Intelligence Research Society (FLAIRS) Conference* (pp. 89–94). AAAI Press.

(2004). Useful roles of emotions in artificial agents: A case study from artificial life, in *Proceedings of the 19th National Conference on Artifical Intelligence* (pp. 42–7). AAAI Press.

(2011). Architectural roles of affect and how to evaluate them in artificial agents, *International Journal of Synthetic Emotions* 2(2): 48–65.

Scheutz, M., Schermerhorn, P., and Kramer, J. (2006). The utility of affect expression in natural language interactions in joint human–robot tasks, in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (pp. 226–33).

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge MA: MIT Press.

Shanahan, M. P. (2005). Consciousness, emotion, and imagination: A brain-inspired architecture for cognitive robotics, in *Proceedings aisb 2005 symposium on next generation approaches to machine consciousness* (pp. 26–35).

Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness, *Journal of Consciousness Studies* 10(4–5): 133–72.

Sloman, A. and Croucher, M. (1981). Why robots will have emotions, in *Proceedings of the 7th International Joint Conference on AI* (pp. 197–202).

Trappl, R., Petta, P., and Payr, S. (eds.) (2002). *Emotions in Humans and Artifacts*. Cambridge MA: MIT Press.

Turing, A. (1950). Computing machinery and intelligence, *Mind* 59: 433–60. Reprinted (1963) in E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought* (pp. 11–35). New York: McGraw-Hill.

Velásquez, J. D. (1999). When robots weep: Emotional memories and decision-making, in *Proceedings of the 15th National Conference on Artificial Intelligence* (pp. 70–5). Menlo Park, CA: AAAI Press.