The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents?

Matthias Scheutz
Human-Robot Interaction Laboratory
Department of Computer Science
Tufts University, Medford, MA 02155, USA
mscheutz@cs.tufts.edu

Abstract—Humans are deeply affective beings that expect other human-like agents to be sensitive to and express their own affect. Hence, complex artificial agents that are not capable of affective communication will inevitably cause humans harm, which suggests that affective artificial agents should be developed. Yet, affective artificial agents with genuine affect will then themselves have the potential for suffering, which leads to the "Affect Dilemma for Artificial Agents", and more generally, artificial systems.
In this paper, we discuss this dilemma in detail and argue that we should nevertheless develop affective artificial agents; in fact, we might be morally obligated to do so if they end up being the lesser evil compared to (complex) artificial agents without affect. Specifically, we propose five independent reasons for the utility of developing artificial affective agents and also discuss some of the challenges that we have to address as part of this endeavor.

Index Terms—affect processing, intelligent artificial agent, affect dilemma, ethics

1 Introduction

Recent evidence from research in human-machine interaction suggests that appearance and behavior of artificial agents will cause humans to form beliefs about their human-likeness: the more human-like the appearance and behavior, the more people will be inclinded to believe that the agent has human-like capacities (e.g., [32], [7]). In particular, natural language interactions – a prototypical human form of communication – with embodied autonomous agents such as robots will cause people to automatically form expectations about the agent's behavior and internal states, similar to the ones they would form about other humans (e.g., see [38]). For example, an elderly human being helped out of her bed by a medical assistant robot will likely expect the robot to show empathy when she displays pain. And while there is no need for the robot to show empathy to get the job done, the lack of displayed empathy in a robot incapable of emotional expressions might give the elder the impression that the robot does not care (lack of expression is often taken as a meaningful expression itself). And while the robot does not necessarily have to be capable of empathy to produce behavior that expresses empathy in this case, it has been argued that, in general, it will difficult for such a robot to display and express emotions appropriately (i.e., at times when humans would) without actually having them. Moreover, given that affect processes are critical in humans for managing computational resources, complex artificial agents might benefit from such mechanisms as well. And some have even argued that complex robots with human-like cognitive capabilities will automatically be capable of instantiating some human-like emotions (e.g., [45]).

We will return to this point about the need for artificial agents to be affect-sensitive in more detail later; for now, let it suffice to say that being able to take the human affective makeup into account will likely improve human-machine interactions and thus the machine's task performance (if it depends on human collaboration and compliance as in the elder-care scenario). Hence, all else being equal, we should endow artificial agents with affect mechanisms as long as we manage to develop sufficiently sophisticated computational models of human affect that capture the critical functional roles of affect for the type of agent and interaction under consideration (e.g., in social interactions, say). Note that this challenge of developing appropriate human affect models might well require us to develop truly cognitive human-like agents, which is far beyond our current abilities. However, *if* we had such models, then in principle, at least, the proposal is that we *should* use them to improve interactions between humans and artificial agents.

However, as we will see shortly, this normative implication quickly runs into problems of its own. For endowing artificial agents with *genuine affect processes* entails that these agents will *be capable of having affective states of their own*.[1] And since these affective states will comprise positive and negative affect, artificial agents will instantiate, at least in part, negative affective states, something that is clearly *bad for the system*. And while this does not imply that all agents will be aware of their affective state or that they will be aware that they are in a particular affective state (i.e., that they have self-awareness with respect to their affective make-up), complex artificial agents with human-

1. Note that we take "affect" to be a more general concept subsuming all kinds of related concepts such as emotions, moods, etc. [48].

like cognitive mechanisms that include the capability for awareness and self-awareness will inevitably also be aware of their affective state (for otherwise affect would not be integrated in a way that models human-like affect processes).[2] Thus, by developing affect models and endowing artificial agents with them, we will have generated artificial agents that have the potential for *suffering*.[3] Note that this does not necessarily mean that they actually *will* suffer (e.g., their operation might be too short or the circumstance might not arise for them to instantiate negative affective states). Nor does this necessarily mean that these agents will suffer in exactly all the same ways humans can suffer (e.g., if some of the concomitant cognitive states are different from human-like cognitive states). Yet, it does not either remove the possibility for genuine suffering just because the computational architectures of those agents are realized in non-biological substrates (e.g., imagine a human brain whose neurons are replaced one-by-one by functionally equivalent artificial ones over many years a la Pylyshyn [31]) or because the control processes in the architecture are not quite human-like (e.g., most people would concede that some animals can suffer and there is mounting evidence for it, e.g., see [20]). And it seems difficult to argue (although this point is still open for debate) that for complex artificial agents capable of genuine affect (even if it is not human-like), the moral value associated with those kinds of affective states and the moral implications for those agents to be in or to be put into those kinds of affective states ought to be different from moral values associated with similar types of states in animals and humans. Hence, the very fact that we could create artificial agents with the potential to suffer seems to imply, on moral grounds, that *if* such agents *could suffer*, that we *should not* engineer artificial agents capable of affect.

Thus, in sum, we are faced with the "Affect Dilemma for Artificial Agents", namely that there are moral reasons both for endowing and not endowing artificial agents with mechanisms that enable them to instantiate affect processes.[4]

2 The Affect Dilemma for Artificial Agents

The Affect Dilemma is particularly worrisome because, on the one hand, there is increasing urgency to resolve it given that we are already witnessing the deployment of intelligent autonomous artificial agents into society, while, on the other hand, it is not immediately clear *how* to resolve it. The obvious way of avoiding it – to simply eliminate affect processes in humans and make humans operate at a purely rational level as embodied, for example, by the fictional Vulcans in the Star Trek science fiction series – is a non-starter. For it seems reasonable to assume that either human affect cannot be fundamentally altered (for evolutionary reasons) or that even if it could, we would not want to fundamentally alter it (possibly again for moral reasons or because it would essentially change us, as affect is a critical part of what makes humans human).

Another solution, to stop building and deploying artificial systems, seems unrealistic from a practical perspective: the market dynamics of scientific research and development are not easily impressed by moral arguments and typically find a way to develop potentially dangerous technology despite intellectual worries about its consequences. And even if policy makers should step in and explicitly prevent certain applications or products, the basic research itself is not subject to these constraints and cannot be stopped.[5] Moreover, there is already too much interest and momentum in the international cognitive systems community, fueled by ambitious funding programs in Japan, Korea, the EU, and most recently the US as well, for us to put a halt on it (who would do it and why?).[6]

Hence, if affect cannot be eliminated in humans and the development of intelligent artificial systems, in particular, artificial cognitive systems, cannot be stopped, we have to find a resolution of the Affect Dilemma that involves affective processes, either in humans alone or in humans and artificial agents, or more generally, artificial systems. The trouble is that the same causes both obligate us to pursue affect mechanisms in artificial systems – to prevent human suffering – and prevent us from building artificial systems with affect – to prevent their suffering. More specifically, if we do not design artificial systems that are affect-sensitive (i.e., that are capable of recognizing human affect and responding appropriately to it, which will ultimately require them to instantiate affect processes), we will face the problem that humans expect artifacts (especially artifacts that project agency, due to autonomous mobile behavior, see [38]) to be sensitive to their affect and likewise to display affect of their own. For systems that do not recognize and display affective states will cause harm to humans for the very fact that "no affective display" is also a display (that of indifference, not caring, etc.) and that "not recognizing affect in humans" is also a "recognition" (that of ignoring the expressed affect). Hence, for moral reasons, we should avoid harm to humans and thus not design (complex) artificial agents that interact with humans

2. Note that we are assuming that complex artificial agents with a human-like computational architecture will be capable of human-like psychological states, (e.g., see the discussion about "psychological equivalence" in [30]).

3. Some have argued that we should stop designing agents *now* because current agents are already capable of and, in fact, *are* suffering [24].

4. Compare this to the three paradoxes described in [8].

5. Even amateurs will soon be able to build quite complex cognitive systems based on widely available software components and technology, very much like Libyan rebels are already able to build improvised military robots combining existing technologies, e.g., see the video at http://english.aljazeera.net/video/africa/2011/06/201161411201323416.html (accessed on 08/06/2011).

6. Note that artificial intelligence has made steady progress ever since its in inception over fifty years ago, even though the ambitious goals of the early AI pioneers to reach human-level intelligence have not yet come to fruition as they imagined. The various success stories such as IBM's Deep Blue and Watson programs, Google search, Nasa's Mars rovers, and many other sophisticated autonomous machines hint at the enormous potential of future truly intelligent artificial systems for all aspects of society. Assuming there is nothing in principle that prevents non-biological machines from reaching human-level intelligence and beyond, it seems very likely that truly intelligent artificial systems will be developed, sooner or later.

without being at least "affect-sensitive".

On the other hand, if we do design such affect-sensitive artificial systems (that are capable of recognizing human affect, responding to it appropriately, and likely having genuine affective processes of their own), then we might still cause harm to humans and possibly to those systems themselves. To see this, consider two ways of being "affect-sensitive": the affect is "fake" (i.e., the system only displays it without having or experiencing it), in which case, the system may or may not get the social aspects of the affect behavior right (in interactions). Either way, however, this can result human suffering; for if the affective social interactions do not work properly, it will cause harm to humans. If, however, the affective interactions with humans work out, then this opens up the possibility for the system to exploit human affect without being subjected to any adverse affective effects itself. Specifically, the system might engage in affective behavior that to the human is suggestive of there being particular affective states realized in the system, while, in fact, the system is in no such inner states (we will return to this point later). As a result, the human might, for example, develop unidirectional emotional bonds with a system that itself is not capable of having emotions [38]. And this very fact, that humans will automatically develop certain kinds of affective reactions and dependencies, could be exploited by the system, thus causing humans harm. And even if it is not intentionally exploited, the creation of accidental human affective dependencies on the system can still cause humans harm. Hence, for moral reasons, we should avoid harm and thus not design artificial systems that are affect-sensitive in the first sense.

The second way to be affect sensitive is for the system's affect to be genuine, i.e., the system displays and responds to human affective states because of its own affective states. And again, the system may or may not get the affective interaction with humans right: if it does not get it right (e.g., because its affective states and their causal potential are somewhat different from human states due to the system's different cognitive make-up), the disconnect in the affective communication has the potential to cause both the human and the system harm, because their affective perceptions and responses do not match up. If the system gets the affective behaviors right, however, then we managed to design a system with genuine human-like affect that is as much subject to negative affect as humans are, and we have thus increased the potential for the realization of artificial systems that suffer. Hence, for moral reasons, we should avoid harm and thus not design artificial systems that are affect-sensitive in the second sense. And in sum, we should not design artificial systems that are affective-sensitive, for moral reasons.[7]

As the attribute "dilemma" suggests, we cannot have it both ways, one direction has to give. Which one, then, depends on the associated advantages or disadvantages, and possible additional arguments that we might bring to bear in order to argue for one or the other side. For example, one might weigh the social benefits and social costs of having affective systems as part of our society, such as the benefit of better human-robot interaction (with affective robots), as compared to the potential for abusive human-robot relationships. Or one may weigh the individual benefits (for the artificial system) and the social cost of having affective systems as part of our society, such as the benefit of being able to make reasonable (affect-based) decisions under time pressure in many situations (where deliberative mechanisms are not applicable due to time or resource constraints) or the costs of robots instantiating negative affective states or potentially developing affective disorders (very much like in the human case).

One might argue that, given that we have to make a decision on which direction to pursue and that we do not yet have affective systems capable of human-like affect, it might be best to stop designing affective artificial systems and accept that human interactions with artificial systems will cause humans harm of varying degrees. Even though this line of reasoning is clearly the easier path from a technological perspective, we will argue that the disadvantages of this position by far outweigh the disadvantages of designing systems with affect, and that the advantages of artificial systems with affect outweigh the disadvantages of such systems. Hence, the proposed resolution of the Affect Dilemma will be to recognize that the moral force with which the two incompatible actions are recommended is not equal, and that we should go for the lesser evil, which is to explore affect mechanisms for artificial systems. Thus, we will next make the case for why we should pursue affective artificial systems.

3 Five Reasons for Affective Artificial Systems

First note that it is not clear whether there are any good reasons in favor of or against research on affect (or any topic or subfield in artificial intelligence, for that matter), based on *apriori grounds* alone (e.g., based on some notion of "apriori utility" of a particular formalism, tool, or mechanism or based on some apriori ethical considerations). From the perspective of artificial intelligence research, for example, the question is an empirical one (and thus aposteriori): how well affect mechanisms will work for different tasks in different environments. In other words, the performance comparison of different system architectures is what ultimately determines whether one mechanism should be preferred over another. And one could also take a wider perspective and include the impact of artificial systems on humans, in particular, and society, in general, into account as part of the performance measure, thus including ethical reasons in the evaluation. In the following, we will propose five independent reasons for investigating affect processes in artificial systems, all of which have value in their own

7. Note that there are parallels here to ethical questions that arise in the context of animal agriculture, where biological affective agents are suffering to alleviate human suffering caused by malnutrition. Similarly, the situation of building and deploying affective artificial systems that have the potential to suffer is not unlike giving birth to children, a process that also creates new agents with the potential to suffer.

right, but together make in our view a strong case for the utility of affective artificial systems vis-a-vis the potential dangers connected with and resulting from them.

3.1 Computational Models of Affect

Computational models are playing an increasingly important role in the sciences and have been successfully employed in cognitive science to study various mental phenomena. Hence, there is intrinsic utility to building and implementing models of human affect as part of the classical research loop of empirical discovery and theorizing: starting with a cognitive/mental phenomenon in reality and its description in a theory, experiments are conducted to verify/falsify the given theory (e.g., [26] or [29]) by deducing a prediction for a particular case. The (empirically constructed) theory is then transformed into a computational model, i.e., the "affective system", and the empirical experiment replaced by a "virtual experiment" which is conducted by running a simulation of the system model on a computer. The result of this virtual and cyclic simulation process is twofold: first, it creates predictions for "real world dynamics" and second, 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 that case, a rewritten version of the theory acts as the starting point for a new cycle of empirical and/or simulation experiments. Note that because some computational models will require social situations and real-time interactions with humans who are embodied and situated in an environment, these models will have to be implemented on robots.

3.2 The Utility of Affective Control Mechanisms

Artificial agents, much like real agents, are often subject to real-world constraints such as constraints on processing time and computational resources. Hence, while it is often in principle possible to develop an algorithm that computes the theoretically optimal solution to a given problem, the algorithm practically fails because its resource requirements exceed the available resources. In those cases, affective control mechanisms might help [36] as they, in part, serve the functional role of providing evaluations of situations that allow an agent to make "good enough" decisions within the given resource constraints (rather than optimal ones). Among the functional roles are simple control mechanisms such as fast reflex-like reactions in critical situations that interrupt other processes and help to decide what to do next, but also more complex internal processes related to the focus of attention, the retrieval of relevant memories, the integration of information from multiple sources, the allocation of processing resources and the control strategy for deliberative processes [36]. In humans, affective processes complement deliberative processes in ways that lead to better performance at a great variety of tasks [4]. Similarly, in artificial agents, affective control mechanisms, supplementing non-affective control strategies, can lead to better agent performance [40]. It thus seems that whether affect is useful or better than other methods for artificial agents, is a question that should be investigated and determined in systematic analytic and experimental evaluations of affective mechanisms, e.g., comparing affective to non-affective mechanisms [36] (we will expand on this point later).

3.3 Ethical Decision-Making in Artificial Systems

There is currently increasing interest in ethical decision-making and some have proposed architectural mechanisms that would allow machines to make some ethical decisions [2], [1]. However current algorithms are very limited in scope and the question arises whether there is a general purpose algorithm for allowing machines to make ethical decisions. Aside from the fact that it is not even clear from a philosophical perspective which ethical theory these decision algorithms should follow (e.g., deontological, utilitarian, virtue- or rights-based, etc.), it is an open practical question whether these algorithms would be feasible and produce good results, even if the theoretical preference issue were settled. This is because many ethical decision processes would require vast amounts of data for the outcome to be justifiable (e.g., a utilitarian approach would require that an artificial system be able to compute all possible action outcomes together with an assessments of their costs and benefits, aside from questions about the scope of the assessment). Especially under time pressure, such general approaches seem doomed to fail. And while it might be possible to encode parts of the decision procedure in rules that can be applied quickly (e.g., as in Arkin's Ethical Governor [2]), these rules will likely be insufficient to capture cases in previously inexperienced situations that the designers did not foresee at design time. Hence, both the lack of computational power (and knowledge) as well as the lack of preference for an agreed-upon ethical theory together suggest that we might have to contemplate other ways of reaching ethical decisions, in addition to the current approaches. This is exactly where affective states might help. Since affect is at the core a measure of how good or bad a situation is for an organism [36] and some affective evaluations are social (e.g., the ones connected to social emotions like empathy), the right kinds of affective behavior might then also turn out to be ethically justifiable as a byproduct of the system's affective evaluations in social contexts. But note that affect is neither necessary nor sufficient for artificial systems to be able to make ethical decisions. For the former, observe that even an affective artificial system could still reach ethical decisions on purely rational grounds alone; for the latter, note that not all affective evaluations are necessarily good (e.g., there might be a tension between individual and social emotions in the system or the system might be subject to affective disorders in very much the way humans are). Nevertheless, negative affect could prevent an artificial system from acting in ethically unjustifiable ways if the system were endowed with mechanisms that would allow it to "experience" negative affective states (in a clear sense of

"experience" that we would have to specify, of course) and if those experiences were tied to the appropriate corrective actions (to prevent those experiences or to attempt to reduce them).

3.4 Complex Agents will have Affective States

There is another line of reasoning for the investigation of affect in artificial systems that does not even involve arguments in favor of or against designing artificial systems with particular affect mechanisms: we will eventually, by necessity, have to deal with affect in artificial systems once these systems become *complex enough*; not so much for its utility, however, but to be able to avoid its *negative effects*. For systems of a particular complexity (with deliberative and reflective mechanisms) might be capable of instantiating what Sloman [45] calls "tertiary emotions" as a *byproduct of their processing mechanisms*. For example, typical human-like emotions such as "guilt", "infatuation", and others seem to result in frequent interruptions of 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 a system, but rather *emerge* as a result of the interactions of various components in the control system. This is similar in principle to other non-affective "emergent states" such as "thrashing" or "deadlocks" that occur in many computer systems where in neither case the system has been *designed* to enter or instantiate these states, but where the presence of architectural mechanisms (e.g., paged, virtual memory in the first, multiple parallel processes with process synchronization mechanisms in the second case) makes them possible.

One consequence of the nature of these emergent states is that robots like "Commander Data" from Star Trek with at least human-like cognitive capacities and thus all relevant architectural components that give rise to human-like cognitive processes, but *without architectural mechanisms* that allow for the instantiation of (tertiary) emotions, are impossible (e.g., see [49]). This is because if all relevant architectural components for human cognition are part of an architecture, the architecture has the potential to instantiate the same kinds of emergent states as the human cognitive architecture. Yet, it might still be possible to build systems in such a way as to prevent the instantiation of negative affective states in most circumstances (e.g., similar to an operating system that might be able to prevent "thrashing" by monitoring the time spent on swapping memory pages, detecting that it is entering an "unstable state", and suspending processes in response until it reaches a save state again).

3.5 Preventing Human Harm and Suffering

Finally, there is yet another reason for advancing our understanding of affect and exploring affect in artificial systems in the long term, both at a conceptual as well as an implementational level, that has not been addressed yet, but will eventually be of critical importance for its impact on society: this is to prevent human harm resulting from unidirectional emotional bonds with social artificial systems such as social robots. We have previously argued at length [38] that in particular social robots such as entertainment robots, service robots, and robot companions have the potential to automatically and unintentionally exploit human innate emotional mechanisms that have evolved in the context of mutual reciprocity which robots do not have to meet. Specifically, social machines with (limited) autonomy and adaptivity (to be able to change their behaviors) have far-reaching consequences, for their mobility enables them to affect humans in very much the same way that humans are affected by animals or other people. In particular, it allows for and ultimately prompts humans to ascribe intentions to social robots in order to be able to make sense of their behaviors. And there is growing evidence that the exhibited degree of autonomy of an artificial system is an important factor in determining the extent to which it will be viewed as human-like [19], [42]. Hence, designers are faced with the challenge to find ways to counter this process of automatic impression forming in humans that takes place when humans interact with these artifacts (in [38], for example, we show that this even happens with surprisingly simple autonomous robots like the Roomba vacuum cleaner). It is clear that robots are already causing humans harm either out of incompetence by being neglectful of human affective states (e.g., failing to notice and counter human anger directed at the system because of problems with its behavior) or by wrongly indicating states that are not present in the system (e.g., suggesting through facial expressions that the system is interested, happy, sad, etc.).

We believe that rather than trying to make robots behave in ways that do not trigger "Darwinian buttons" in humans (which will be difficult if not impossible), we might have to accept this human propensity and rather put mechanisms into place on the artificial system's side that will prevent, to the extent possible, unidirectional emotional bonds. Endowing artificial systems with genuine affect of their own is one solution that would, at the very least, enable the possibility of bidirectional emotional exchanges between humans and artificial systems. The idea here is to trade the potential suffering of artificial systems with human suffering, and to allow for the possibility of new suffering (by virtue of creating machines capable of it) to reduce or prevent human suffering. And while it will also open up the possibility for artificial systems to suffer, it is exactly the potential for suffering (e.g., the potential for unrequited love) that will be able to prevent, by and large, what could otherwise be large-scale abusive relationships between humans and non-affective robots in the future.

4 Affect in Artificial Systems: Some of the Challenges Ahead

Having made the case for investigating and developing affective artificial systems, it is also important to point out the challenges that have to be addressed as part of this process, which range from conceptual challenges, to architectural and implementation challenges, to challenges regarding the comprehensive evaluation of positive and negative effects of affect mechanisms in artificial systems.

4.1 Conceptual Challenges

So far, we have used "affect" as a pre-scientific, unanalyzed term, relying on our folk intuitions about the types of concepts that "affect" might denote. It is, however, critical to appreciate the range of concepts that fall under the general notion of affect: from mere *sensations* or *feelings* (the purest form of qualitative experience, e.g., pains), to simple *homeostatic control and drives*, to various kinds of *motivations*, to *basic emotions* and *moods*, to all kinds of *complex emotions*, *attitudes*, *tastes*, and many more. Most of these categories are themselves composed of concepts that comprise instances of varying complexities and require elaboration.

The conceptual difficulties in defining affect concepts is widely acknowledged among researchers in the "affective sciences" [6]. First and foremost, the extension of "affect" is unclear: for some affect includes feelings and moods, but not emotions, whereas for others affect is a general notion subsuming feelings.

Second, the differences among the various subclasses of affects (moods, emotions, feelings, etc.) are unclear. For example, consider how *moods* differ from *emotions*: for some, moods are just like emotions except that they evolve at a larger time scale (i.e., they last for a much longer time span and are gradually modified by several events, e.g., [11]). For others, moods are distinctly different from emotions in that moods modulate or bias cognition (e.g., memory, see [9]), whereas emotions are taken to modulate or bias action: they "will accentuate the accessibility of some and attenuate the accessibility of other cognitive contents and semantic networks" [5]. Yet, others draw a distinction between moods and emotions with respect to their *intentionality*: moods, as opposed to emotions, do not have an object, toward which they are directed. They are "non-intentional affective states" [13] in that *they are not about anything*. One can be anxiety-ridden without there being a particular object causing the anxiety, while fear is typically triggered by a perception of a particular kind (i.e., the perception of an object eliciting fear, see also [25]).[8]

Third, the extensions of the various subclasses of affect are themselves unclear. For example, there is no consensus among scholars working on emotions about how to construe *basic emotions* or whether the concept is even coherent (for more details and references see [27] or [16]): while Ekman [10] individuates basic emotions based on universal facial expressions (i.e., expressions of anger, disgust, fear, joy, sadness, and surprise), James [18] takes fear, grief, love, and rage to be basic emotions. For Izard [17], anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, and surprise are basic emotions, while yet others count rage, terror, anxiety, and joy as basic, or expectancy, fear, rage, and panic. Even less agreement can be found regarding higher, or more complex emotions such as infatuation or embarrassment (e.g., see [47]).

Yet, even though there is no formally precise characterization of affect concepts, it is still possible to start with operational definitions of some forms of affect in order to be able to investigate affect phenomena. As many psychologists and artificial intelligence researcher have pointed out (e.g., [28], [44], [14], [3]), affect 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 has, experiences, or displays some form of affect, but rather a whole sequence of such states in combination with environmental states. "Fear", for example, does not refer to the make-up 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, etc. Consequently, we construe affective phenomena as intrinsically *process-like*, which may or may not include *state-like* phenomena (e.g., disappointment includes a surprise state). In accordance with the process view of affect as a means of behavior control, affective processes are instances of *control processes*. As such they serve the primary purpose of initiating, ending, interrupting, modulating, regulating, and adjusting agent behavior, from simple motivational control (like a food drive caused by hunger), to moods (like elation), to emotions (like fear), to complex forms of affect (like embarrassment or passionate love).

To be able to distinguish control processes that involve or instantiate affect concepts from those that do not involve or instantiate them, we use a rough distinction between *belief-like* and *desire-like* states [35]: a state S (of an agents' control system) is *belief-like* if the agent is disposed to change S so as to maintain a correspondence or correlation between S and some other environmental state E, whereas S is *desire-like* if the agent is disposed to change some environmental state E so as to maintain a correspondence or correlation between S and E. Note that we construe the environmental state E in a "wide sense" so as to include states of the agent as well.

Affective control processes, then, comprise more or less *positively or negatively valenced* desire-like states as part of their control process, which the agent does or does not desire to a varying degree, similar to "belief-like" states which can vary according to the degree to which the agent is committed to taking them as true or false (see [48] for

8. It is interesting in this context that what used to be called "affective disorders" in clinical psychology has come to be called "mood disorders".

a much more detailed account and definition of "affect" and "emotion"). Furthermore, affective control processes are *teleological processes* in that they deal with the control of tasks that *matter* to the *well-being* of the agent—for organisms produced by evolution this is related to survival, procreation, etc., while for artifacts *well-being* may have to be determined and specified by the designer. For an "affective multi-user operating system", for example, well-being may be defined as being able to serve all user demands with no latency. Such a system may get "stressed out" and eventually "overwhelmed" when too many users start demanding jobs, which then may cause it to engage "coping mechanisms" (e.g., stopping low-priority jobs, refusing new network connections, etc.) that will bring it back into a desirable, "healthy" state (or it may get "depressed" and never recover).

The last example also raises an important point that may express the sentiments of someone inclined to keep affect out of artificial intelligence research: "All of this seems to be true of standard (non-affective) operating systems, why should such an OS be called 'affective', i.e., why introduce terminology that has all kinds of connotations? It looks like calling these mechanisms 'affective' has no added benefit, but will cause a lot of conceptual problems and lead to wrong conclusions about the nature of the system". This is clearly a valid concern that ultimately has to do with the difference between the inscription of states versus the presence of states in an architecture. A more mundane example to illustrate this difference is that of "wall following behavior" in two kinds of robots: in the first, there is a particular component in the robot's architecture that is activated when a wall is detected and leads to the observable wall following behavior by monitoring the distance of the robot to the wall and adjusting its movements to keep that distance constant. In the other robot, "wall following" does not have an architectural representation, but rather "emerges" from the interplay of two components, one which controls forward movement and the other which controls turning away from something that is taken to be an obstacle (in this case the wall). Both systems exhibit the same observable "wall following" behavior, but only the first implements an explicit control component for it. Note that the difference between the two systems does not show up *factually*, but rather *counterfactually*: what the two robots *would be* capable of doing if different parameters in the architecture *were* to be changed. In particular, suppose we wanted to add an LED to the robot that should be turned on whenever it is performing "wall following" behavior—this would be trivial in the first (just turn it on dependent on the activation of the "wall following" component), but more difficult in the second (as what constitutes "wall following" would have to be determined from the activations of forward movement and turning). It is this kind of difference that distinguishes the "affective" from the "non-affective" operating system in terms of the non-observable behavior, and the motivation for explicit representations of affective control states to be able to utilize them (e.g., ideally to improve task performance, but in the simplest case just to signal as with the robot that the system is entering a particular, possibly undesirable state, which would allow the operator to take actions).

In sum, affective control processes in animals (and consequently in artifacts) may or may not have any of the following characteristics: a perceptual component that can trigger the affect 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 mechanism, 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"). None of these aspects are necessary for affect, nor are they sufficient. Yet, most of them are taken to be part of the many forms of human affect we know from our own experience. Getting clear on what they are and how to define them is part of an ongoing debate in artificial intelligence, robotics, psychology and philosophy.

4.2 Architectural and Implementation Challenges

While much work on affect has been pursued in the context of user interface design, where efforts focused on affect recognition and expression, and sometimes how to connect the two to improve the experience of human users with an interactive system, we really need to worry about "architectural aspect of affect". For the ultimate goal is to produce artificial systems that can actually instantiate affective processes rather than simply detecting and expressing affect without having the internal structure required to possess or have affect. There are several crucial differences between user interface and architectural aspects of affect, most importantly that the former does not require the instantiation of affective states within a system that deals with affective interactions with the user (e.g., 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 be intrinsically committed to claiming that affective states (of a particular kind) can be instantiated within the system. Moreover, the former does not need a satisfactory *theory of affect* (i.e., what affective states are) to be able to produce working systems. Being able to measure changes in a user's skin conductance, breathing frequency, etc. and using this information to change the level of detail in a graphical 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 system might learn the user's preferences based on such measures (if there is a correlation) without requiring any representation of the user's affective processes nor any affective processes itself.

Contrariwise, architectures that claim to use affective mechanisms (e.g., for the prioritization of goals or for

memory retrieval) will have to make a case – by necessity – that the implemented mechanisms (can) indeed give rise to "affective states" (in a clearly specified sense), otherwise there is no sense, nor any reason, to call them that. Nor would it make any sense to insist that a robot is capable of having affect if it there are no criteria for the possibility of affective states being present in the robot that can be objectively assessed and evaluated. Hence, we effectively need to put our cards down on what it means to "implement affect mechanisms", at the very least relative to the kinds of affective states we claim a system can instantiate.

In addition to having a theory of implementation of affective states, we need to ensure that the right kinds of affective states are implemented and are implemented properly (e.g., that the transitions between different affective states are right, that the timing is within human expectations and that expressed affective states are recognizable as what they are supposed to be). Otherwise the results could be particularly problematic and ultimately cause human suffering, because people will not know how to react to the system. In the simplest case, they will sense a feeling of estrangement or eeriness [22], which in the worst case can cause emotional pain if they are not able to interpret the behaviors of their favorite robotic companion. Such failures will almost inevitably occur with shallow models of affect because such models recognize and react to perceived affect not based on a deep model of affect that gets the internal mechanisms and processes right, but on surface rules that directly connect perceptions to actions which, as a result, do not cover all possible cases and affective interactions, and do not have the same causal potential and relations among other states as the deep models.

4.3 Evaluation Challenges

The most difficult challenge in the process of developing affective artificial systems is probably to evaluate their performance and their effects on humans, as there are several parts to the evaluation. For one, we need a thorough comparison of the tradeoffs between affective and non-affective control mechanisms for various tasks in various environments [36], where tasks may be individual or social tasks. And then we ultimately need to evaluate those systems in carefully controlled short-term and long-term interaction experiments with humans in order to assess their effects and impact on humans. Based on our experience from performing both types of evaluation studies, we predict that the picture will be quite complex, i.e., there will be clear cases where particular classes of affective mechanisms are not advantageous and there will be cases where depending on the particular values of task parameters, affective mechanisms may turn out to be better than other non-affective mechanisms. Hence, it is important to acknowledge that general statements about the utility of affect may not be possible in many cases (if at all). Yet, as a community, researchers working on architectural aspects of affect have not yet been able to propose a good, agreed-upon evaluation methodology (different from those working on user interface aspects), even though there some recent promising proposals for simulated interactive agents [15]. We have proposed a comparison of affective and non-affective agents with respect to "relative performance-cost tradeoffs" as we believe that one case where affective mechanisms might turn out to be useful is when the cost of using them is taken into account [36]. Specifically, we proposed the following four-step methodology: (1) (affect) concepts are analyzed and defined in terms of architectural capacities of agent architectures [46], (2) agent architectures with particular mechanisms that allow for the instantiation of affective states as defined in (1) are defined for a given task together with a performance measure, (3) experiments with agents implementing these architectures are carried out (either in simulations or on actual robots[9]), and (4) the performance of the agents is measured for a predetermined set of architectural and environmental parameters. The results can then be used to relate agent performance to architectural mechanisms. Moreover, by analyzing the causes for possible performance differences, it may be possible to generalize the results beyond the given task and set of environments. In the best case, general statements of the form "Mechanism X is better than mechanism Y" can be derived for whole classes of tasks and environments, where "better" is spelled out in terms of the performance ordering obtained from the experiments.

While some evaluations might be carried out in simulated environments without humans (e.g., [37], [36]), it is the human social interaction context with artificial systems that is ultimately of critical importance to the future social structure of our society. Specifically, we need to devise and carry out user studies to determine how humans would react to artificial systems with affect. The results of these study can contribute to the development of both computational models of human behavior as well as better control architectures for interactive artificial systems. Yet, even these user studies might have their own ethical problems associated with them. For example, we have conducted extensive human-robot interaction studies to determine the human reaction to autonomous robots that can make their own decisions as part of a mixed human-robot team [33], [41]. In these experiments, robots could opt to refuse to carry out human commands if those commands were not advancing the joint goals, and we found that humans were accepting of robot autonomy, in part, because robots gave reasons for their refusal, but in part also because of their ability to express affect in their voice. However, it is not clear whether robots should be ever allowed to make decisions that contradict human commands, even in cases where those decisions would be in the interest of some larger goal. Hence, the question arises whether we should even investigate how people might be affected by robot autonomy. The institutional IRB approval required to conduct such user studies might not be enough to justify them [39]. On the other hand, one could argue that we

9. Evaluating the performance of affective robots is notoriously difficult for many reasons, but most importantly, because the systems are still very brittle, but see [43], [33], [34] for a start.

will eventually build and disseminate complex artificial systems that have their own decision mechanisms, and that it is therefore high time to start investigating the possible social impacts of such systems. Clearly, there is another ethical tension between subjecting humans to all sorts of system evaluations that might cause them harm compared to producing systems that, without proper evaluation, might end up on the shelves of stores and also have the potential to do harm, except without any clear sense of the extent of their negative impact.[10]

Finally, in addition to evaluating and ideally quantifying the individual and social benefits of affect mechanisms, we need to carefully investigate the potentially harmful effects of emergent systemic affective (and non-affective) states and the development of mechanisms to prevent them (if possible). For example, we need to determine what kinds of monitoring mechanisms could be added to a system to detect a particular kind of harmful emergent state (e.g., the frequent interruption of a deliberative process by a lower-level alarm mechanism) or, if detection is not possible, what kinds of architectural provisions one could make to contain the effects of such states or to prevent them altogether, again if possible.

5 Discussion

We departed from the question of whether we should develop affective artificial systems, based on the converging evidence from human-machine interaction research that humans will be looking for affect-sensitivity in and engage in affective communications with artificial agents (at the very least with agents that seem to possess human-like capabilities). We then argued that the answer to this question leads to what we dubbed the "Affect Dilemma for Artificial Agents": there are moral reasons for and against developing artificial agents with affect. The discussion of the dilemma demonstrated that there is no easy answer to the question of whether artificial agents should have affective states. Yet, while there is no clear solution in the sense of pursuing one route that has only benefits and no costs associated – otherwise it would not be a Dilemma – we have argued that a resolution of the dilemma ultimately comes down to evaluating which side of the dilemma is the lesser evil in terms of the potential harm it can inflict on humans and possibly artificial systems. Using five independent reasons to demonstrate the utility of artificial systems with affect, we have proposed that endowing artificial agents with genuine affect should be the preferred option compared to not pursuing affect mechanisms in artificial agents at all. We then also pointed to three major challenges pertaining to affect concepts, their architectural role and implementation, and the evaluation of affective artificial systems that need to be addressed as soon as possible; for historically, applications of technologies usually precede ethical reflections of their use, and research and development in artificial intelligence and robotis is no exception. Even if we do not intend at present to work out criteria for saying when a particular affective state is possible within an architecture and whether it actually is present or not, we will, sooner or later be forced to answer the question independent of the whole affect discussion. This is because there will likely be claims made in the future for complex robots (as there have been already in the past and present) that robot R instantiates state X where X is some mental state that we are typically only attribute to human-like cognitive architectures. Suppose someone claims that R is *aware of itself* (e.g., because R recognizes its mirror image as what it is and not as the image of another agent), then we would really want be certain (to the extent that we can be) that this is *really* an awareness state similar to those hypothesized in humans and not just the attribution of an external observer. This is because we attach ethical principles to systems capable of being self-aware.

Ultimately, it is the responsibility of researchers in the artificial intelligence and robotics communities to confront the ethical implications of building artifacts with "alleged" emotions and feelings – "leaving it to the philosophers" is not an option given that artificial intelligence and robotics research will be implicated. Inevitably, questions will be raised (if not from within scientific community, then from the outside) of whether agents can really have feelings, what kinds of feelings they are capable of having, and whether we should – "should" in a moral sense – produce agents with feelings (there are already some indications, e.g., the movie "AI"). These questions will likely force us sooner or later to revisit very touchy questions that we have shied away from answering (for good reasons, after having been charged with "overstepping" our bounds and "overinterpreting" what artificial systems are like [23]): what kinds of mental states are implemented in a given system and what kinds of states is the system capable of instantiating?

We believe that independent of our current interests in exploring the utility of affect for artificial systems, it will eventually become necessary to take a stance on this issue. One motivation might be to understand the causal potential of different affective states (independent of whatever "labels" they might bear or have been assigned by the programmer) in order to prevent false attributions with all the potentially ensuing consequences about alleged "emotional agents". Ultimately, the more pressing reason will be the non-trivial question of whether it is morally permissible to shut off a system that is capable of having feelings of a particular sort. For the acknowledged presence of mental states in artifacts has far-reaching practical, ethical, and legal implications for the systems *qua system* and is eventually prerequisite to our conferring or willingness to confer *rights* upon them.

10. A case in point is Hasbro's "Baby Alive" that pretends to "love" their owners, see http://www.hasbro.com/babyalive/en_us/demo.cfm (accessed 08/06/2011).

6 Acknowledgments

Thanks to Paul Schermerhorn and Gordon Briggs for helpful comments and suggestions.

REFERENCES

[1] Michael Anderson and Susan Anderson. *Machine Ethics*. Cambridge University Press, 2011.

[2] Ron Arkin. *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall, 2009.

[3] Leda Cosmides and John Tooby. Evolutionary psychology and the emotions. In M. Lewis and J. M. Haviland-Jones, editors, *Handbook of Emotions*, pages 91–115. Guilford, NY, 2nd edition, 2000.

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

[5] Richard J. Davidson. On emotion, mood, and related affective constructs. In [12], pages 56–58. 1994.

[6] Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, editors. *Handbook of Affective Sciences*. Oxford University Press, New York, 2003.

[7] Brian R. Duffy. Anthropomorphism and the social robot. *Robots and Autonomous Systems*, 42(3-4):177–190, 2003.

[8] Brian R. Duffy. Fundamental issues in affective intelligent social machines. *The Open Artificial Intelligence Journal*, 2:21–34, 2008.

[9] E. Eich and D. Macaulay. Fundamental factors in mood dependent memory. In J. P. Forgas, editor, *Feeling and thinking: The role of affect in social cognition*, pages 109–130. Cambridge University Press, New York, 2000.

[10] P. Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, April 1993.

[11] Paul Ekman. Moods, emotions, and traits. In [12], pages 56–58. 1994.

[12] Paul Ekman and Richard J. Davidson, editors. *The Nature of Emotion: Fundamental Questions*. Series in Affective Science. Oxford University Press, New York, 1994.

[13] Nico H. Frijda. Varieties of affect: Emotions and episodes, moods, and sentiments. In [12], pages 56–58. 1994.

[14] Nico H. Frijda. The psychologists' point of view. In [21], pages 59–74. 2000.

[15] Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.

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

[17] C. E. Izard. *The Psychology of Emotions*. Plenum Press, New York, 1991.

[18] W. James. *William James: Writings 1878-1899*, chapter chapter on Emotion, pages 350–365. The Library of America, 1992. Originally published in 1890.

[19] S. Kiesler and P. Hinds. Introduction to the special issue on human-robot interaction. human-computer interaction. *Human-Computer Interaction*, 19:1–8, 2004.

[20] I.J.H Duncan K.P Chandroo and R.D Moccia. Can fish suffer?: perspectives on sentience, pain, fear and stress. *Applied Animal Behaviour Science*, 86:225–250, 2004.

[21] Michael Lewis and Jeanette M. Haviland-Jones, editors. *Handbook of Emotions*. The Guilford Press, New York, 2nd edition, 2000.

[22] Karl MacDorman. Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science*, pages 26– 29, 2005.

[23] Drew McDermott. Artificial intelligence meets natural stupidity. In J. Haugeland, editor, *Mind Design*, pages 143–160. MIT Press, Cambridge, MA, 1981.

[24] Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2003.

[25] Arne Öhman. Fear and anxiety: Evolutionary, cognitive, and clinical perspectives. In [21], pages 573–593. 2000.

[26] N. Oreskes, K. Shrader-Frechette, and K. Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263:641–646, 1994.

[27] A. Ortony and T.J. Turner. What's basic about basic emotions? *Psychological Review*, 97:315–331, 1990.

[28] R. Pfeifer. Artificial intelligence models of emotion. In V. Hamilton, G. H. Bower, and N. H. Frijda, editors, *Cognitive Perspectives on Emotion and Motivation, volume 44 of Series D: Behavioural and Social Sciences*, pages 287–320. Kluwer Academic Publishers, Netherlands, 1988.

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

[30] Hilary Putnam. Robots: Machines or artificially created life? *The Journal of Philosophy*, 61(2):668–691, 1964.

[31] Z. Pylyshyn. The "causal power" of machines. *Behavioral and Brain Sciences*, 3:442–444.

[32] Susan R. Fussell Sara Kiesler, Aaron Powers and Cristen Torrey. Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, 26(2).

[33] Paul Schermerhorn and Matthias Scheutz. Dynamic robot autonomy: Investigating the effects of robot decision-making in a human-robot team task. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, Cambridge, MA, November 2009.

[34] Paul Schermerhorn and Matthias Scheutz. Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-i nitiative task. In *The Fourth International Conference on Advances in Computer-Human Interactions*, pages 236–241, 2011.

[35] Matthias Scheutz. The evolution of simple affective states in multi-agent environments. In Dolores Cañamero, editor, *Proceedings of AAAI Fall Symposium*, pages 123–128, Falmouth, MA, 2001. AAAI Press.

[36] Matthias Scheutz. Architectural roles of affect and how to evaluate them in artificial agents. *International Journal of Synthetic Emotions*, 2011.

[37] Matthias Scheutz. Evolution of affect and communication. In Didem Gökcay and Gülsen Yildirim, editors, *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*, pages 75–93. 2011.

[38] Matthias Scheutz. The inherent dangers of unidirectional emotional bonds between humans and social robots. In Pat Lin, Keith Abney, and George Bekey, editors, *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, 2011.

[39] Matthias Scheutz and Charles Crowell. The burden of embodied autonomy: Some reflections on the social and ethical implications of autonomous robots. In *Proceedings of ICRA 2007 Workshop on Roboethics*, 2007.

[40] Matthias Scheutz and Paul Schermerhorn. Affective goal and task selection for social robots. In Jordi Vallverd and David Casacuberta, editors, *The Handbook of Research on Synthetic Emotions and Sociable Robotics*, pages 74–87. 2009.

[41] Matthias Scheutz and Paul Schermerhorn. Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In *The Fourth International Conference on Advances in Computer-Human Interactions*, 2011.

[42] Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson. First steps toward natural human-like HRI. *Autonomous Robots*, 22(4):411–423, May 2007.

[43] Matthias Scheutz, Paul Schermerhorn, James Kramer, and C. Middendorff. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM International Conference on Human-Robot Interaction*, pages 226–233, 2006.

[44] A. Sloman. The mind as a control system. In C. Hookway and D. Peterson, editors, *Philosophy and the Cognitive Sciences*, pages 69–110. Cambridge University Press, Cambridge, UK, 1993.

[45] A. Sloman. Beyond shallow models of emotion. In Elisabeth Andre, editor, *Behaviour planning for life-like avatars*, pages 35–42. Sitges, Spain, 1999. Proceedings I3 Spring Days Workshop March 9th–10th 1999.

[46] A. Sloman. Architecture-based conceptions of mind. In *Proceedings 11th International Congress of Logic, Methodology and Philosophy of Science*, pages 397–421, Dordrecht, 2002. Kluwer. (Synthese Library Series).

[47] A. Sloman. How many separately evolved emotional beasties live within us? In R. Trappl, P. Petta, and S. Payr, editors, *Emotions in Humans and Artifacts*, pages 29–96. MIT Press, Cambridge, MA, 2002.

[48] A. Sloman, R. Chrisley, and M. Scheutz. The architectural basis of affective states and processes. In J.M. Fellous and M.A. Arbib, editors, *Who needs emotions? The Brain Meets the Machine*. Oxford University Press, New York, forthcoming.

[49] I.P. Wright, A. Sloman, and L.P. Beaudoin. Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126, 1996. Repr. in R.L.Chrisley (Ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol IV, Routledge, London, 2000.

Matthias Scheutz received degrees in philosophy (M.A. 1989, Ph.D. 1995) and formal logic (M.S. 1993) from the University of Vienna and in computer engineering (M.S. 1993) from the Vienna University of Technology (1993) in Austria. He also received the joint Ph.D. in cognitive science and computer science from Indiana University in 1999. Matthias is currently an associate professor of computer and cognitive science in the Department of Computer Science at Tufts University. He has over 150 peer-reviewed publications in artificial intelligence, artificial life, agent-based computing, natural language processing, cognitive modeling, robotics, human-robot interaction and foundations of cognitive science. His current research and teaching interests include multi-scale agent-based models of social behavior and complex cognitive and affective robots with natural language capabilities for natural human-robot interaction.