Georgia State University
ScholarWorks @ Georgia State University

Philosophy Theses | Department of Philosophy

8-12-2014

On the Possibility of Robots Having Emotions

Cameron Hamilton

Follow this and additional works at: http://scholarworks.gsu.edu/philosophy_theses

Recommended Citation
Hamilton, Cameron, "On the Possibility of Robots Having Emotions." Thesis, Georgia State University, 2014.
http://scholarworks.gsu.edu/philosophy_theses/150

This Thesis is brought to you for free and open access by the Department of Philosophy at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Philosophy Theses by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

ON THE POSSIBILITY OF ROBOTS HAVING EMOTIONS

by

CAMERON REID HAMILTON

Under the Direction of Andrea Scarantino

ABSTRACT

I argue against the commonly held intuition that robots and virtual agents will never have emotions by contending robots can have emotions in a sense that is functionally similar to humans, even if the robots' emotions are not exactly equivalent to those of humans. To establish a foundation for assessing the robots' emotional capacities, I first define what emotions are by characterizing the components of emotion consistent across emotion theories. Second, I dissect the affective-cognitive architecture of MIT's Kismet and Leonardo, two robots explicitly designed to express emotions and to interact with humans, in order to explore whether they have emotions. I argue that, although Kismet and Leonardo lack the subjective feelings component of emotion, they are capable of having emotions.

INDEX WORDS: Emotions, Robotics, Artificial Life, Artificial Intelligence, Affective Science

ON THE POSSIBILITY OF ROBOTS HAVING EMOTIONS

by

CAMERON REID HAMILTON

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Philosophy

in the College of Arts and Sciences

Georgia State University

2014

Copyright by
Cameron Reid Hamilton
2014

ON THE POSSIBILITY OF ROBOTS HAVING EMOTIONS

by

CAMERON REID HAMILTON

Committee Chair: Andrea Scarantino

Committee: Neil Van Leeuwen

Daniel Weiskopf

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2014

DEDICATION

To Alan Mathison Turing (1912-1954), whose visions of the future of artificial intelligence have inspired my endeavors towards their actualization.

ACKNOWLEDGEMENTS

I would first like to thank Andrea Scarantino, whose critical analysis of my thesis over the course of a year ensured I produced the strongest work possible. Without his guidance, this thesis would not have the philosophical import it currently has, and for that I am truly grateful. Sandra Dwyer has also been an inspiration throughout my time as an instructor at Georgia State, so I must thank her as well. I would also like to thank Dan Weiskopf and Neil Van Leeuwen for being my committee members and for providing commentary on this thesis. Finally, I would like to thank my friends and family would were willing to listen to and engage in discussion about something as strange as emotional robots.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ..... **v**

LIST OF FIGURES ..... **viii**

1 INTRODUCTION **1**

2 WHAT ARE EMOTIONS? **6**

2.1 Components of Emotion **6**

2.1.1 Appraisal of Events ***6***

2.1.2 Neurophysiological Changes ***7***

2.1.3 Action Tendencies ***8***

2.1.4 Expressions ***9***

2.1.5 Subjective Feelings ***10***

2.2 Contemporary Theories of Emotion **11**

2.2.1 Basic Emotions Theory ***12***

2.2.2 Dimensional Theory ***13***

2.2.3 Cognitivism ***16***

2.2.4 Feeling Theory ***17***

2.2.5 Conclusion ***19***

3 ROBOTIC EMOTIONS **20**

3.1 Kismet **22**

3.2 Leonardo **27**

4 CAN ROBOTS HAVE EMOTIONS WITHOUT FEELINGS? 29

4.1 The Unemotional Robots View 30

4.2 The Emotional Robots View..... 35

5 ROBOTS CAN HAVE EMOTIONS..... 45

5.1 Component Process Model of Emotions 46

5.2 Kismet and Leonardo Have Emotions 47

5.3 Why Do Robots with Emotions Matter?..... 53

6 CONCLUSION 56

REFERENCES..... 59

LIST OF FIGURES

Figure 3.1 Photograph of the robot Kismet 21

Figure 3.2 Kismet expressing fear 22

Figure 3.3 Kismet facial expressions corresponding with recognizable emotions 25

Figure 3.4 Photograph of Leonardo 27

Figure 3.5 Leonardo reaching for a toy while expressing interest 29

1 INTRODUCTION

Whether machines could ever have emotions has been a perennial question explored both in the field of artificial intelligence and throughout the history of science fiction. Many researchers seem to have the intuition robots and virtual agents will never have emotions, claiming emotions emerge from the unique make-up of humans and therefore cannot be replicated in an artificial being. I argue against this intuition by contending robots can have emotions in a sense that is functionally similar to humans, even if the robots' emotions are not exactly equivalent to those of humans. To establish a foundation for assessing the robots' emotional capacities, I first define what emotions are by characterizing the components of emotion consistent across emotion theories. Second, I dissect the affective-cognitive architecture of MIT's Kismet and Leonardo, two robots explicitly designed to express emotions and to interact with humans, in order to explore whether they have emotions. I argue that, although Kismet and Leonardo lack the subjective feelings component of emotion, they are capable of having emotions

The motivation for this project is two-fold: first, to derive a working model of emotions useful for artificial intelligence research, and second, to determine if this model could be realized within an embodied artificially intelligent agent. The first motivation stems from the concern that scientists, philosophers, and laypersons alike disagree about what emotions are and often provide conflicting theories to explain them. For instance, a common intuition is that emotions must involve a phenomenal experience of "what it is like" to be in a particular emotional state (i.e. a feeling; see Adolphs, 2005; Arbib, 2005). Others argue it is possible to undergo an emotional episode without feeling anything at all (Nussbaum, 2001; Roberts, 2003). In addition, some theorists have argued emotions most closely resemble cognitive judgments involving propositional attitudes (Calhoun & Solomon, 1984; Solomon 2004), while critics have pointed out that it is

possible for certain emotions, such as fear, to become activated without awareness of the eliciting stimulus (Scarantino, 2010). A further problem arises when trying to classify emotions: Do emotions occur within discrete categories (e.g. anger, sadness, joy, etc.), or are they instantiated within an affective space constituted by one's level of arousal (i.e. one's alertness) and valence (i.e. one's pleasure or displeasure)? As a purely experimental approach has not yet answered these questions, some further methodological reflection appears needed to resolve the debate on the nature of emotions.

I argue that computational modeling provides the appropriate platform for researchers to test theories of emotion by observing whether a theory's implementation reproduces emotional behavior. For a theory of emotions to be useful, it should be capable of replicating the central function of emotion: using attributions of value to stimuli in the environment to motivate actions that further an agent's progress toward its goal(s) (Scherer, 2005). This is to say an agent encapsulating a model of emotions should retreat from a threatening or noxious stimulus, approach and engage with a rewarding stimulus, etc. Furthermore, the agent should be capable of recognizing the emotional expressions of others, and should display expressions recognizable as corresponding with particular emotional states (see Ekman, 1992 and Freitas-Magalhães, 2012 for a discussion of the universality of facial expressions of emotion). To be clear, I am not claiming that a working model must reproduce emotions in an identical fashion to the way humans do. Rather, I am making the more modest assertion that such models can help researchers to establish plausible mechanisms through which emotions emerge, and to reject or modify those mechanisms that are incapable of appraising and responding to salient stimuli in the environment. Thus, my intent in evaluating computational models of emotions is to determine what theories of emotion can

successfully be implemented to reproduce emotional behavior, so as to reduce some of the uncertainty about what emotions are.

The second motivation for this project is to determine whether robots can have emotions. Skepticism about the possibility of emotional robots is largely drawn from the contention that computational agents cannot have intentionality; the capacity to understand or represent states of affairs (Searle, 1980). The general line of reasoning is that even a computational agent that can carry on a natural language conversation, for example, does not actually have an understanding of what is being discussed, as the agent's behavior is mechanistically determined by the algorithms that underlie its architecture. By extension, robots that possess a model of emotions do not have emotions any more than a model of a rainstorm could get someone wet (Searle, 2002, p. 16). For Searle and his supporters, a model is simply a model, and can never replicate the target phenomenon.

I will argue against the assertion that computational agents cannot have intentionality because their behavior is determined mechanically. The behavior of individual neurons is also determined mechanistically: if a neuron receives enough excitatory stimulation from other neurons to surpass its threshold potential, then the neuron will fire an action potential. Otherwise, the neuron remains at rest. Thus, if we accept that computational agents cannot have intentionality because their behavior is mechanistic, we must also accept that biological life cannot have intentionality, including humans. Proponents of biological realism might reply by asserting that it is the specific machinery of the brain that allows humans to have emotions, or as Searle states, "actual human mental phenomena [are] dependent on actual physical–chemical properties of actual human brains" (1990, p. 29). There is, however, no reason to think human brains are the only

control systems[1] capable of instantiating emotions. I contend that so long as the computational agent is able to attribute value to environmental stimuli, based off their relation to the agent's needs and goals, in order to determine a course of action, the agent is emotional. This is to say that what is necessary for having emotions is not a human physiology , but rather that the agent is able to appraise events in the environment, which in turn predispose the agent towards particular action tendencies and expressions for the sake of maintaining a sense of well-being. In what follows, I will demonstrate there are at least two robots that display these requisite components of emotion by evaluating the affective-cognitive architecture with which they interact with the environment. This evaluation will show that it is possible for robots to have emotions which allow them to autonomously act toward the completion of their goals, without a control system that explicitly resembles a human brain.

The possibility of robots having emotions entails serious consequences for how robots are integrated in human society within the future. Breazeal and Brooks (2005, p. 277-279) see at least four relationships between humans and emotional robots that could emerge: robots as tools, as avatars (human surrogates), partners (as in a team), or as cyborg extensions (prostheses). In the first case, for instance, a search-and-rescue robot could benefit from having interest and fear mechanisms for the sake of balancing the robot's persistence in searching for a missing person with its own self-preservation. Emotional robots could also serve as partners in elderly patient care settings, as these robots could be attuned to the specific signs of distress and anxiety exhibited by the patient, in addition to keeping their company. For those who have no desire to carry on interactions with robots as social partners, emotional robots still hold the advantage over unemotional robots in that they are less frustrating to deal with (Picard, 2003). A robot that is able

[1] Control system refers to a system that manages the behavior of other systems. The brain is a control system in the sense that it takes in information from the sensory systems , circulatory system, etc., in order to determine the proper output signal to send to the motor system.

to recognize that a human is becoming irritated by an interaction by monitoring the human's expressions, for instance, can then modify its own behavior to be more accommodating, such as by asking how it may be of greater service, or by leaving the room. In sum, providing robots with the capacity for emotions may improve the autonomy with which they carry out tasks, as well as the quality and helpfulness of their interactions with humans.

The present work is intended for a wide audience: first, philosophers interested in the metaphysics of mind and emotions may be interested in exploring the embodied computational perspective expressed here. Likewise, artificial intelligence researchers and engineers may find this work useful as a loose guide to how autonomous social behavior and goal directedness may be enabled in artificial agents by including an emotional system within their design. Finally, skeptics of the possibility that robots could have emotions will find this work an appropriate platform for constructing a counterargument against robotic emotions, while supporters of robotic emotions may develop further grounds for their position.

In conclusion, it is entirely possible for robots to have emotions, assuming they are able to attribute value to objects and events within the environment so as to motivate the execution of particular actions and expressions in order to attain a set of goals/sense of well-being. Furthermore, these attributions of value should be comparable to the attributions humans make (i.e. they should include assessments of arousal and valence), if robotic emotions are to be comparable to human emotions. The expressions produced by robots should also be recognizable to humans as corresponding with particular emotional states (e.g. sadness, disgust, etc.), such that the robot is able to coordinate others' behavior in appropriate way (e.g. to elicit sympathy, to indicate the presence of a noxious stimulus, etc.). Of course, robots may have emotions in a different sense than humans do, as human well-being and robot well-being are likely to be distinguished from

each other. By accepting the possibility that robots can have emotions, efforts may be taken to construct robots that are tailored to improve the lives of humans and perhaps the lives of robots as well.

2 WHAT ARE EMOTIONS?

There is disagreement in the affective science community about how to define emotions. No consensus appears to have been reached in the emotion literature as emotion theorists remain divided on whether the theory of basic emotions, core affect theory, or some other theory best characterizes affective phenomena. Despite the variance among emotion theories' accounts of what emotions are, there is some agreement about the components that constitute prototypical emotional episodes and what their functions are. In what follows, I will describe these components, while staying neutral for the time being on which combination is essential for an emotion to be instantiated. I will then describe the contemporary theories of emotion, while highlighting how each theory privileges different components as being integral to an emotional episode. This exposition will serve two functions: 1) to demonstrate that each theory remains subject to unanswered criticism, and 2) to pick out commonalities between theories such that an account of emotions based of prototypical components can be forwarded in Chapter 5.

2.1 Components of Emotion

2.1. 1 Appraisal of Events

An evaluation of events in the environment which leads an agent to have a particular emotion is an emotional appraisal (Ellsworth & Scherer, 2003). Some appraisal theories hold appraisals are only performed consciously (Lazarus, 1991), while others posit a two-process ap-

praisal process, where unconscious appraisals are performed initially, and are followed by conscious appraisals (Smith & Kirby, 2009; Scherer, 2001). Despite the distinctions between these appraisal theories, several commonalities emerge which indicate a shared theoretical understanding. For one, most contemporary theorists view appraisals as processes that evaluate how the current state of the environment relates to the agent's well-being (Moors, Ellsworth, Scherer, & Frijda, 2013). The agent's well-being is in turn determined by whether the environment satisfies or prevents satisfaction of the agent's needs, values, goals and beliefs (Frijda, 1986, 2007; Lazaraus, 1991; Scherer, 2004). Appraisals of the significance of environmental events to the agent's well-being then coordinate changes in action tendencies, neurophysiological responses, motor expressions and feelings so as to promote greater well-being through interacting with the environment (Clore & Ortony, 2000; Roseman & Smith, 2001; Reisenzein, 1994; Scherer, 2001).

2.1.2. Neurophysiological Changes

After an appraisal of the environment has been performed, a set of neurophysiological changes occur in order to prepare the agent for action (Scherer, 2005). The most apparent changes resulting from emotions involve activation of the autonomic nervous system and include changes in heart rate, digestion, respiration, perspiration, and pupil dilation (Purves et al., 2001). Although research has continually shown there are no "emotional centers" in the brain (Fellous & LeDoux, 2005), a number of brain structures have been implicated in emotional processing, including the amygdala (LeDoux, 2002), thalamus (Sherman, 2006), hypothalamus (Cao et al., 2012), and the hippocampus (Fischer et al, 2002), among others. At the neurochemical level, emotions are thought to be coordinated by neuromodulation involving dopamine, noradrenaline, serotonin, oxytocin, cortisol and GABA (Fellous, 2004; Panksepp, 1993). Some theorists have even proposed models for how levels of these neuromodulators correlate with specific emotions,

such as anger with low levels of serotonin, high dopamine and high noradrenaline (Lövheim, 2011). Thus, observing changes in activity at these various levels of neurophysiology may give some indication of the observed agent's emotional state.

2.1.3. Action Tendencies

An individual's cognitive and physiological preparedness to act is defined by states of action readiness that are motivated as responses to significant events (Frijda, 1986; 1987; 2004). For an event to qualify as significant, it must be concerning to the individual or relate to the individual's goals. A typical example of a significant event is when Person A says something slanderous about Person B or her kin (e.g. "Your mother is a whore!"). On Frijda's account, a significant event is appraised on what the event does to the individual as well as what the individual may do to the event. In the case of slander, Person B's appraisal will likely acknowledge "a demeaning offense against me and mine" (Lazarus, 1991), as well the possibility of retaliating against Person A. The appraisal then motivates a change in action readiness that has both a quantitative aspect (activation) and a qualitative aspect (action tendency). Activation refers to the intensity of an individual's neurophysiological arousal; the mobilization of energy towards carrying out an action (Frijda, 1986, p. 90). Actions motivated by emotion have control precedence over other behaviors (Frijda & Scherer, 2009). If Person B is in the midst of a conversation when they are insulted by Person A, for instance, Person B's activation towards retaliating will override any other behavior she had intended. An action tendency is a readiness to execute a particular action or set of actions that have the same intent. Action tendencies allow for behavioral flexibility as a person with an action tendency for retaliation may attack, spit on, insult, walk away from, or return slander to the person that offended them (p. 71). As Frijda asserts action tendencies are equivalent to emotions, feelings refer to an individual's awareness of the for-

mation of their action tendencies as well as their execution. Thus, Frijda provides an account that sets motivation toward particular actions after a significant event as central to what constitutes emotion.

Deliberative action tendencies involve conscious reflection on the behavior to carry out (Frijda, 2010). In contrast, the tendency to retreat when frightened exemplifies an impulsive action tendency, as the scared person automatically withdraws from the feared object without considering alternative courses of action. Frijda distinguishes impulsive actions from reflexes on the grounds that they are intentional; although the appraisals that cause them are formed without conscious deliberation, impulsive actions are directed towards specific objects (e.g. a threatening person or animal) and bear an intent (e.g. to avoid the potential for injury). They are also differentiated from habits and routines such as brushing one's teeth or driving home, as impulsive action tendencies demand immediate attention. Through the formation of action tendencies, both impulsive and deliberative, an agent is driven to modify the environment so as to maintain or further her own well-being.

2.1.4. Expressions

Expressions prompted by an emotional process include facial displays (Keltner et al., 2003; Ekman et al., 1992; Parkinson, 2013), vocal intonations (Russell et al., 2003; Scherer, 2003), and body postures (Wallbott, 1998; Coulson, 2004). As a full account of all forms of expression is beyond the scope of this thesis, I will focus specifically on facial displays, which appear most often in the emotion literature.

Facial expressions appear to communicate social intentions, while also conveying information about the emotion which motivated the display (Ekman, 1994; Fridlund; 1994). For example, bared teeth may reveal an intention for aggressive interaction and convey information

about an underlying state of anger. While expressions are taken to be voluntary on Fridlund's (1994) view, there is reason to think facial displays can be involuntarily produced for the sake addressing a salient stimulus requiring immediate attention. There also appear to be informal social norms within a given culture, known as display rules, which dictate when and where it is appropriate to express particular emotions (Siegler, 2006). For example, displays of anger are less frequently displayed in collectivistic cultures such as Chinese culture, than individualistic cultures, such as the United States, as displays of angers are taken as threatening to group harmony (Miyake & Yamazaki, 1995). In sum, facial expressions serve to communicate an individual's social intentions, such that interactants may coordinate their actions to achieve mutual achievement of their goals.

2.1.5. Subjective Feelings

Feelings are notoriously difficult phenomena to characterize, due to their subjective nature. Unlike facial expressions or physiological changes, feelings are primarily measured through self-reports. This difficulty in measuring feelings thus leads to uncertainty in understanding what role they play in an emotional episode. Nevertheless, feelings are suggested to serve as central representations of the patterns of change occurring for each component, such that the agent has a coherent conception of their emotional state they can use to interact with the environment (Scherer, 2005). Within an emotional episode feelings are thought to serve a monitoring function, whereby the agent is informed of their current emotional state through querying the state of their body (Adolphs, 2005). The agent may then use the information she has obtained about her body state to consciously deliberate on an appropriate action. For example, if an individual walks in on two members of their family committing incest, that individual is able to observe the nausea they feel, which may cause them to avoid their family members for a period of

time. It is the conscious accessibility of feelings that distinguish them from the other components of emotions. While what constitutes an adequate definition of consciousness is hotly debated, the term is generally taken to refer to the experiences an agent has as a subject differentiated from objects and events in the world; a mental process where the subject has a phenomenal awareness of their selfhood and is able to execute control over their thoughts (Farthing, 1992; Van Gulick, 2011). As there is a nearly unlimited number of different ways in which an individual as a subject can interact with different objects in the environment, so too is there diversity in the feelings one can experience as a result of an eliciting stimulus. The varieties of subjective feelings are proposed to be comprised of experiences of what it is like to be in a particular emotional state as an evaluation of the state of one's body, how that state relates to events in the world (i.e the antecedent events of the feeling), and thoughts about the actions to be carried out as a consequence of the eliciting event (Lambie & Marcel, 2002). As feelings are phenomenological in nature, they are often thought to be what most clearly distinguishes man from machine.

2.2 Contemporary Theories of Emotion

New theories of emotion emerge on a continual basis, as emotions remain a yet-to-be-fully-understood class of phenomena. In this section, I will characterize the following prominent theories: basic emotions theory, core affect theory, cognitivism, and feeling theory. In the process of doing so, I hope to make clear how the theories are distinguished from each other in how they categorize affective phenomena and which components they emphasize as most integral to emotional episodes.

2.2.1 Basic Emotions Theory

Some theorists contend there are a number of discrete emotions, each with their own set of associated neurophysiological phenomena (Tomkins, 1962). The theory of basic emotions holds that there is a set of emotions shared by all humans that evolved to deal with ancestral life challenges (Ekman, 1992; Izard, 1992). For example, disgust evolved to deal with the challenge of avoiding noxious stimuli, and fear evolved to deal with the challenge of avoiding dangers. Although the exact number of basic emotions is disputed, happiness, surprise, fear, sadness, anger, and disgust are often thought to comprise the most prototypical basic emotions (see Plutchik, 1980; Ekman, 1999). Much of the supporting evidence offered for the theory comes from experiments that show how certain facial expressions are universally associated with specific basic emotions, regardless of the observer's cultural background. This universality has a production side and a recognition side. On the production side, a particular emotional state is said to elicit a facial expression comprised of a fixed set of facial muscles. For example, Duchenne smiling, which involves contraction of the zygomatic major and the orbicularis oculi muscles, is produced when one is genuinely happy (Duchenne, 1990; Messinger, Fogel, & Dickson, 2001). Individuals who do not report a state of happiness, but smile voluntarily have been found to be capable of only contracting the zygomatic major muscle, not the orbicularis oculi muscle. This finding and related results have led some researchers to postulate that emotional expressions are elicited in an automatic fashion (Ekman, 1977; Griffiths, 1997).

On the recognition side, observers are able to infer the emotional state of the emoter, due to the direct correspondence between emotional states and the facial expressions they cause. The observation that facial expressions can be recognized as corresponding with discrete emotional categories cross-culturally, in addition to evidence of differing ANS patterns between emotions

(Ekman, Levenson, and Friesen, 1983), has also led theorists to infer that there are likely neural substrates that are particular to each basic emotion (Izard, 1992). Thus, basic emotions theory holds neurophysiological changes and expressions as most definitive of an emotional episode.

2.2.2 Core Affect Theory

Others have claimed that emotions should be given a dimensional analysis and be distinguished by their state within an affective space constituted by valence (degree of pleasure) and arousal (degree of responsiveness to stimuli/wakefulness) dimensions (Barrett & Russell, 2009; Fontaine et al., 2007), rather than as discrete or basic emotional categories.[2] Core affect theory, the most prominent theory of this persuasion, dismisses basic emotions as folk psychological concepts, and instead proposes emotions should be conceived of as consciously accessible neurophysiological states constituted by a blend of pleasure-displeasure and energization-enervation feelings (Russell, 2003). In addition, one's appraisal of stimuli in the environment is both influenced by and influences one's core affect. This is to say the more positive one's core affect, the more positively events will be perceived and remembered. For instance, an individual in a state of positive core affect is more likely to think they made a good impression when meeting a stranger than when meeting them in a state of negative core affect (Forgas & Bower, 1987). This idea is supported by experimental findings of a mood congruency effect, such as participants' ability to recall negatively or positively valenced words with greater accuracy in the corresponding affective state (Knight, Maines, & Robinson, 2002). Likewise, participants report lower self-esteem when negative affect is induced and higher self-esteem when positive affect is induced (Smith & Petty, 1995). Russell (2003) proposes affective appraisals of stimuli can also cause changes in one's core affect. For example, the perception that a friend no longer wants to go to

[2] Some theorists have claimed emotions require three or more dimensions to capture all the nuances of emotional phenomena (see Fontaine et al., 2007).

the movies will likely bring on negative affect. This induced state of negative affect may then lead an individual to perceive others as having a disinterest in interacting with them in the future. Thus, dimensional theories argue that appraisals of environmental stimuli (including of one's self) and the changes to one's core affect these appraisals cause are constitutive of an emotional episode. Furthermore, this core affect need not correspond with or be recognizable as one of the basic emotion categories (e.g. happiness, fear, anger, etc.).

Core affect theory's main criticism of basic emotions theory rests on the observation that affective phenomena appear to be both qualitatively and quantitatively diverse. Russell (2003) argues the labels "fear," "anger," "happy," etc., do not capture this diversity. For instance, one might say an individual being chased by an assailant brandishing a knife (Case A), an individual who retreats from a cockroach scurrying across the floor (Case B), and an individual who is concerned they will never find a career that is fulfilling (Case C) are all in a state of fear. On the basic emotions account, an emotional episode involves fixed patterns of neurophysiological and facial expression changes in response to an eliciting stimulus that are distinct between emotions, but are the same within the same emotional category (Ekman, 1992; Griffiths, 1997; DeLancey, 2002). If this were the case, one would expect that the three individuals described above would respond to their eliciting stimuli in the same way, yet homogenous responses between the three cases seem unlikely. Core affect theorists, in contrast, would argue that the individuals in Cases A, B, and C are applying the concept of fear to experience, despite the fact that each individual has a unique core affect. While basic emotion theorists would hold that since all three individuals are experiencing fear, they would execute the same cascade of responses to the stimuli, core affect theorists would contend this is not the case, as each individual bears a core affective state that is distinguished from the other two. For instance, the individual's arousal in response to an

armed assailant in Case A should be greater than the individual in Case B's response to a cockroach, as the former case poses a threat to their life. As a result, the individual in Case A would likely make every effort to escape from the assailant, including trying to negotiate and plead with the assailant, while the individual in Case B would be relatively less dedicated to escaping the cockroach. In sum, core affect theory is compatible with the differences in the cascade of responses to eliciting stimuli, while basic emotions theory only allows for a single fixed cascade of responses to a given emotion.

Russell (2003) further criticizes basic emotion theory for failing to account for affect that lacks object-directedness (p. 147). On a basic emotions account, an emotion is thought to have an intentional object it is directed towards (e.g. I am angry at *you*, I am sad *Sarah* died). Core affect theory argues that core affect may not necessarily be aimed at a particular object: for instance, an individual may feel something like anger towards a friend who cancels plans to hang out in order to spend time with her boyfriend, or they may experience a similar state of physiological arousal without knowing of anything in particular that has offended them. Dimensional models of emotion, such as core affect theory, are therefore capable of accounting for a wider array of affective phenomena than basic emotions theory.

The greatest difficulty dimensional theories of emotion face is determining the [number of] dimensions sufficient for characterizing affective phenomena. Most theorists are in agreement that one dimension is inadequate, as emotions appear to involve degrees of both pleasure-displeasure as well as activation-inactivation (Feldman-Barrett & Russell, 1999). Although a number of researchers have derived these two dimensions from studies of self-reported feelings, and facial expressions (see Russell, 1980, 1991), others have proposed at least two additional dimensions, control and unexpectedness, are needed to capture the important similarities and dif-

ferences in emotion-laden words contained across three languages (Fontaine et al., 2007). On their account, appraisals of control result in feelings of power or weakness, which motivate tendencies to act or refrain from action (p. 1051). Appraisals of unexpectedness involve determining whether a stimulus is familiar or novel, the latter of which often prompts individuals to socially reference others to determine the appropriate course of action Asch, 1952; Schachter & Singer, 1962; Bandura, 1992; Feinman, 1982). Thus, dimensional theories remain subject to skepticism, despite some advantages they provide over basic emotions theory (see Izard, 2007 and Panksepp, 2007 for further discussion).

2.2.3 Cognitivism

The most popular version of cognitivism holds that emotions are special kinds of judgments (Solomon, 1980; Nussbaum, 2001)[3]. On this view, fear of another person, is the judgment that the other person is dangerous. Anger is the judgment that one has been offended. Guilt is the judgment that one has violated a moral standard one holds. On the cognitivist account, emotions are always "about" a particular intentional object; emotions are not simply feelings. For instance, one can be afraid of a stranger approaching, of spiders, or of public speaking, as one may bear the propositional attitude that these objects are dangerous, physically or psychologically (Solomon, 2003). Although cognitivists typically hold that feelings alone are not sufficient for having emotions, some theorists concede feelings are at least partly constitutive of an emotional episode (Solomon, 2003). Whether feelings, desires, or physiological responses comprise emotions is of secondary importance within the cognitivist account which contends emotions are essentially judgments, which are generally assumed to be conscious but may in principle also be unconscious.

[3] There is a variety of cognitivism that identifies emotions as certain sets of belief and desire pairs (Marks, 1982), however this variety is not considered here.

One of the primary criticisms of cognitivism is that the rationality of emotions is not equivalent to the rationality of beliefs, despite the cognitivist requirement that emotions involve propositional attitudes (Ben-Ze'ev, 2000; Goldie, 2000; Elster, 2004). This criticism is closely related to a second criticism: emotions can occur despite an apparent lack of judgments/beliefs. A person might become afraid when encountering a milk snake, for instance, despite having the belief that milk snakes are harmless. In contrast, relatively few people have a fear of driving, despite believing that driving is dangerous. Critics of cognitivism claim the persistence or recalcitrance of emotions despite conflicting propositional attitudes is sufficient to show emotions are not judgments or beliefs (D'Arms and Jacobson, 2003; Brady, 2009; de Sousa, 2013).

That emotions are judgments about some intentional object is further challenged by the occurrence of "blindfright" (Scarantino, 2010). Blindfright is "the activation of the fear system in response to visually presented information registered without awareness" (p. 735), where awareness is taken to be the ability to verbally report on the presented information. In priming studies where the eliciting stimulus is presented 30 ms before the masking stimulus, participants respond to questions about the eliciting stimulus as if they have no awareness of it, yet demonstrate neurophysiological excitation indicative of fear, such increased skin conductance, startle reflex, and amygdala activation (Marcel, 1983). As the fear system has been shown to be activated by subliminal stimuli without participants' awareness undermines the cognitivist claim that judgments are essential to emotions.

2.2.4 Feeling Theory

Feeling theories of emotion share the assumption that emotions are feelings. As I noted above, feelings are commonly defined as conscious subjective perceptions of physical changes in one's self in relation to changes in the environment. Feeling theories originated with William

James (1884), and later Carl Lange (1885), who held, "the perception of bodily changes, as they occur, *is* the emotion" (James, 1884, p. 188). On their account, acknowledging that one has been offended sets off a cascade of sympathetic nervous system activation (e.g. increased heart rate, pupil dilation, etc.), and one's perception of those changes is what constitutes one's anger. This particular feeling theory was later undermined by Cannon (1927) and Bard (1928), who demonstrated that cats with severed connections between their sympathetic nervous system and central nervous system still showed aggressive behavior towards barking dogs. Further evidence against the James-Lange theory includes the fact that the afferent sensory fibers innervating internal organs and glands (i.e. the visceral system) are a tenth of the number of efferent sensory fibers (Langley & Anderson, 1894). Afferent fibers carry information from the sensory receptors of the organs and glands to the central nervous system, while efferent fibers send signals to organs and glands from the central nervous system. That there are fewer afferent fibers than efferent fibers innervating internal organs and glands suggests that less information is being processed by the central nervous system about the current states of these organs and glands than the information processed regarding the motor commands send to the organs and glands. This finding therefore provides reason to think awareness of bodily changes is more minimal than James or Lange proposed. Thus, more contemporary feeling theories assume the recognition of physiological responses to an eliciting event comes after recognizing that one is experiencing an emotion.

A further problem with feeling theories in general is that emotions are not differentiable by their corresponding perceptions of physiological changes alone. In a seminal study by Schachter and Singer (1962), participants were injected with epinephrine and placed in room with a confederate who behaved either angrily or playfully with the participant. Results showed participants' classification of their own emotional state tended to reflect the confederate's behav-

ior: participants in the room with the aggressive confederate reported feeling angry, and participants with the playful confederate reported feeling happy. This study suggests that perception of physiological changes alone is not sufficient to constitute an emotion, as similar physiological changes can be shared between emotions.

Goldie (2009) asserts that the division classical feeling theories have drawn between cognition and feelings is misguided. On his account, feelings do not pertain only to changes in the state of one's body, but can also be about objects in the world outside the body, which is a view shared by Lambie & Marcel (2002). He refers to these types of feelings as *feelings towards* and holds they cannot be separated from cognitive processes. To support this notion, Goldie presents a thought experiment involving an ice scientist, Irene, who has never before slipped on ice, but has a deep theoretical understanding of how dangerous ice is. Goldie argues that once Irene slips on the ice and hurts herself, she will have gained the *phenomenological* concept that ice is dangerous, where before she only had the *theoretical* concept that ice is dangerous (p. 234). While Irene previously had the belief that ice was dangerous and the desire to avoid falling on ice, she now bears a phenomenological concept of her fear of ice which can neither be reduced to her theoretical concept of ice's danger, nor to her bodily feeling alone. Irene's thoughts and actions thus become different from those she would have had she never slipped on the ice, according to Goldie. Therefore, feelings may be said to extend beyond mere perception of bodily sensations to include intentionality towards objects in the world that is not reducible to beliefs or desires.

2.2.5 Conclusion

There is no consensus on what emotions are, except on the fact that emotions involve a set of components. Although the lack of an agreed upon definition for emotions is apparent in the contemporary theories of emotion I have characterized here, several themes have emerged which

are common to each theory. For one, emotions are thought to be preceded by an eliciting event as appraised by the emoter, such as the noxious odor of vomit or the presence of a threatening person ,which encourages an immediate cascade of responses. This contention is held by all emotional theories except psychological constructionism (core affect theory), which asserts an individual observes a pattern of components and from that infers they are in a particular emotional state (Russell, 2003, p. 152). Second, the response is determined by the individual's appraisal of the event, whereby the individual determines how the event relates to their sense of well-being. Third, the appraisal of the event motivates neurophysiological changes which prepare the individual to execute a particular behavior, in order to modify the individual's relationship to the event. Finally, some theorists see emotions as involving a subjective experience, which allows the individual to reflect on the emotional episode and regulate it. This shared theoretical understanding will serve as the working definition of what emotions are throughout the course of this thesis.

3 ROBOTIC EMOTIONS

In order to explore the possibility of robotic emotions, I will consider two robots, Kismet and Leonardo. While many robots give the appearance of having emotions by producing expressions recognized as emotional by human observers, Kismet and Leonardo were designed with drives that motivate the robots to socially communicate with others, in order to get help satiating these drives (Breazeal & Brooks, 2005). Kismet has three drives which represent the robot's needs: a need to be stimulated by toys (stimulation drive), a need to interact with people (social drive), and a need to rest (fatigue drive). The comparisons made between the fulfillment of these drives and the robots' appraisal of the current state of the environment is what determines the ro-

bots' behavioral response. Breazeal claims these comparisons determine the robots' emotive state (p. 292), while also noting the robots' emotions are not equivalent to humans'. Since specifying all the details of the robots' affective-cognitive architecture are beyond the scope of this paper, I will focus only on denoting the aspects integral to how the robots deliberate on possible behaviors, and communicate with human interactants, for the sake of determining whether these robots have emotions in Chapter 5. Although the robots have similar architectures, I identify the important differences that allow Leonardo greater social and affective functionality.

3.1 Kismet

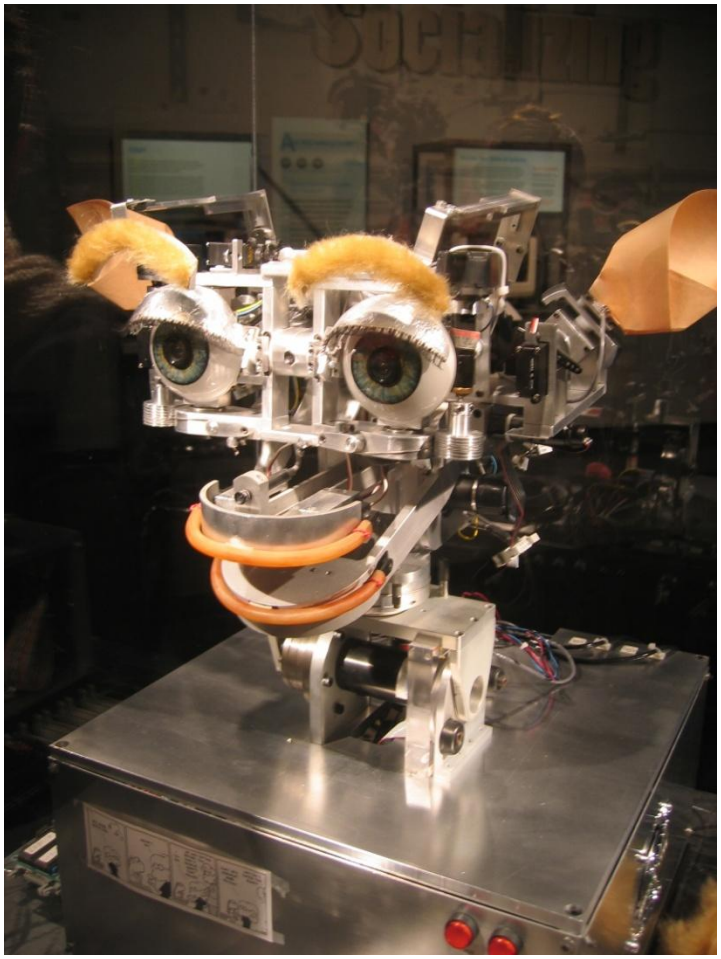

Figure 3.1. Photograph of the robot Kismet. Photograph taken by Jared C. Benedict on 16 October 2005. Copyright © Jared C. Benedict.

Kismet is an anthropomorphic robot developed at MIT by Cynthia Breazeal for the purpose of demonstrating how the capacity for face-to-face social interaction can be achieved through machines (Breazeal & Brooks, 2005). The robot was modeled after the interaction style between an infant and adult human, Kismet communicates what Breazeal calls its emotive state through the prosody of its vocalizations which are pre-linguistic 'babblings,' rather than through explicit description of its states as human adults do (Breazeal, 2002; Menzel & D'Aluisio, 2000). In addition to vocalizations, Kismet communicates its current state through facial expressions and gaze direction. An object that is brought towards Kismet too rapidly and too close to its cameras, for instance, will elicit fear in Kismet who will produce a corresponding facial expression (Figure 3.2)

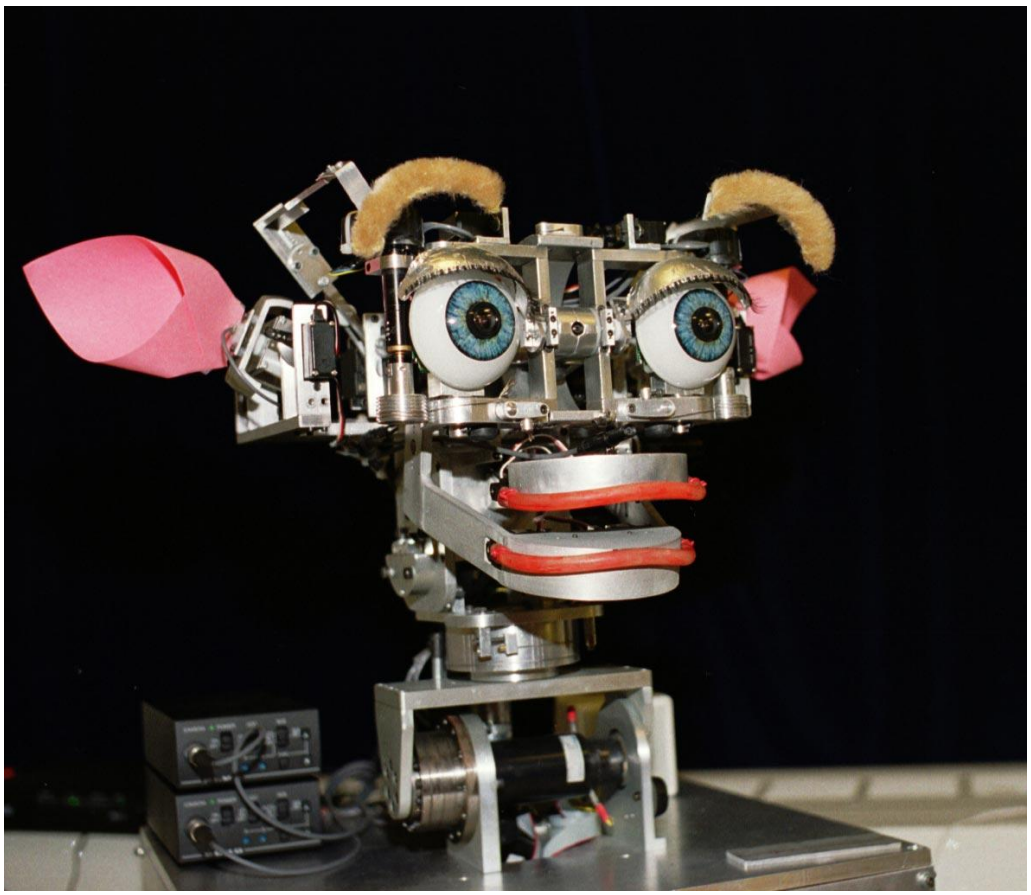

Figure 3.2. Kismet expressing fear. Reproduced from http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html

Kismet's expression resembling fear is typically enough to cause interactants to back away from the robot (Breazeal & Brooks, 2005, p. 289). Thus, by indicating intention through its facial expressions, Kismet is able to encourage interactants to maintain its state of well-being (i.e. a state where its drives are satiated).

To understand how Kismet's emotive state and subsequent behaviors are generated, consider an example where a toy is presented to Kismet. Before proceeding with the example however, it is important to note the term "emotive state" is used by Breazeal and her colleagues to denote the internal state of the robot which promotes particular facial expression, vocal intonation and posturing (Breazeal & Brooks, 2005). I will continue to use the term here, without yet endorsing the robot as having emotions.

When Kismet first sees an object, it performs a feature extraction, noting the object's size, direction and speed of motion, and color, noting whether that color is within the range of human skin tone. Kismet's cognitive system then assesses these features in conjunction to determine whether a toy or person percept should be formed. An Elmo puppet, for example, is easily classified as a toy by Kismet, as the object is about a foot tall and brightly-colored. Percept information is then passed on to the robot's affective releasers which further classify the toy as a desired, undesired or threatening stimulus, based both on the percept and on which of Kismet's drives are active. Kismet's appraisal module will then assign affective tags to the toy based on the robot's current goal, which is again determined by its active drive. Affective tags are assigned values along three dimensions: how stimulating-enervating Kismet finds the object (arousal), how attractive-aversive Kismet finds the object (valence), and whether Kismet is inclined to withdraw from or approach the object (stance). A toy that is bright-colored and moving slow enough for Kismet to easily track will likely get coded as a desired stimulus by Kismet's

affective releasers, particularly if Kismet's stimulation drive is active. If the toy is desired, it will likely receive moderately high appraisal tags for arousal, valence (positive), and stance (approach). In contrast, Kismet's affective system will likely code the toy as undesired if its drive for rest is active, and will further appraise the toy with lower values for arousal, valence (negative), and stance (withdraw). A toy that is presented such that it crowds Kismet's cameras will be coded as a threatening stimulus, causing the toy to be appraised with high arousal, low valence and low stance (withdraw) values. The net values of the appraisal tags will then be passed on to emotion elicitors, which would then initiate an emotive state - happiness in the first case, disgust in the second, and fear in the third.

Once what Breazeal calls happiness is activated, Kismet will then display a facial expression that maps onto the valence, arousal, and stance values appraised of the stimulus (see Figure 3). Although Breazeal explicitly states "Kismet's emotive system is strongly inspired by various theories of basic emotions" (p. 293), she also asserts, "Inspired by [Smith & Scott's (1997) theory of facial expressions], Kismet's facial expressions are generated using an interpolation-based technique over a three-dimensional affect space" (p. 282). Breazeal adds the dimension of stance to the traditional two-dimensional model, which roughly parallels the dimension of control Fontaine and his colleagues (2007) posit, as appraisals along these dimensions are said to motivate agents to approach or withdraw from a stimulus. According to Breazeal, the stance dimension characterizes how approachable or repulsive Kismet finds the stimulus (p. 300). Stance is distinguished from valence, which corresponds with how favorable Kismet appraises the stimulus to be, though the researchers concede that both dimensions tend to increase/decrease in accordance with each other. On her view, there are nine basic facial expressions that span this affect space, though other expressions can be produced that resemble a blend of two or more basic expres-

sions. One might say the model of emotions employed in Kismet was constructed with the tacit assumption that basic emotions are particularly recognizable or common positions within the affective space constituted by these three dimensions. From a dynamical systems perspective, basic emotions could be conceived of as attractors within this affective space. While Breazeal endorses basic emotions theory, it is clear Kismet's model of emotions hybridizes elements from both basic emotions theory and dimensional theory (core affect theory).

Figure 3.3. Kismet's facial expressions corresponding with recognizable emotions. By "Tired", Breazeal means Kismet is in an emotive state more commonly referred to as boredom. Reproduced from http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html

As Kismet's emotive state is constituted by appraisal along valence, arousal, and stance dimensions, changes within the state space result in coordinated changes in Kismet's expressions.

For instance, as arousal increases, Kismet's ears move upwards, while its mouth and eyes widen. Likewise, as arousal declines, Kismet's ears sink down, its mouth closes, and its eyes shut. When a desired toy is first presented to Kismet, the robot's ears will perk up, and its mouth and eyes will widen with interest, as shown in Figure 3 above. However, after the interactant uses the toy to play with Kismet for an extended period of time, Kismet will attenuate to the stimulus (i.e. its fatigue drive will become active), such that it finds the toy less arousing, and will produce an expression of tiredness. In this way, Kismet's expressions remain synchronized with its emotive state and active drives.

As Kismet is an anthropomorphic robot with facial features similar to humans, its emotional state can be inferred even by the naïve human observer (Breazeal & Brooks, 2005; Breazeal & Scassellati, 1999). Breazeal states "…Kismet's expressive behavior is effective because it is readily understandable and predictable to the person who interacts with it. This follows from the fact that Kismet's emotive responses are modeled after basic emotions that are universally understood by people (Ekman, 1992)," (Breazeal & Brooks, 2005, p. 303). In a certain sense, Kismet's emotive state is more transparent to an observer than a human's, as Kismet forms no intentions for deception; all of its expressions are directly correlated with its current emotive state and its intention to satisfy active drives. Through the transparency of its emotive state, Kismet is able to encourage human interaction in order to fulfill the robot's goals.

3.2 Leonardo

Figure 3.4 Photograph of Leonardo. Reproduced from http://spectrum.ieee.org/automaton/robotics/robotics-software/leonardo

Leonardo, another social robot developed by Cynthia Breazeal and her colleagues, bears an affective-cognitive architecture that is similar to Kismet's (Breazeal, 2003). Unlike Kismet, however, Leonardo has a torso with operating arms and hands with which it is able to complete tasks, such as memory tasks involving button pressing or assembling blocks (Berlin, Gray, Thomaz, & Breazeal, 2006). The robot is also more sophisticated in its social interactions, as it is capable of recognizing the emotional states of its human interactants through an implementation of simulation theory (Thomaz, Berlin, & Breazeal, 2011). Like human infants, the robot learns to understand its interactant's emotional state by imitating their facial expressions (Meltzoff, 1996), which incites in Leonardo an emotional state corresponding with the expression produced. If the robot observes a human interactant upturning the corners of its mouth and forming wrinkles at the corners of her eyes, Leonardo will execute motor changes that replicate

this facial configuration. Feedback from Leonardo's face is signaled to the robot's emotive system, which informs the robot that it is happy. Thus the robot is able to form the belief that the human interactant is currently happy.

The interactant first teaches Leonardo to recognize her facial expressions by imitating Leonardo's facial expressions. Once Leonardo has learned a mapping of its own facial features to those of its interactant, the robot can then imitate the interactant's facial expression by blending features of its seven basic facial expressions to reach an expression closest to the interactant's. When Leonardo produces a facial expression, an output signal is sent to its affective system which activates the emotional state that corresponds with that expression. As the robot's affective system is based on Kismet's affective system, Leonardo tags perceptual and internal states with valence (positive or negative) and arousal (high or low) tags, which bias its attention toward/away from particular stimuli and motivate subsequent behavior. In addition, the robot is able to assign novelty values to the stimulus, based off whether that stimulus has been appraised in the past or not. Thus, by imitating an interactant's facial expression, Leonardo is able to induce an emotional state within itself analogous to the interactant's, such that the robot is able to understand the interactant's behavior in much same way that humans do.

Figure 3.5 Leonardo reaching for a toy while expressing interest. Screenshot taken from http://www.ted.com/talks/cynthia_breazeal_the_rise_of_personal_robots.html

4 CAN ROBOTS HAVE EMOTIONS WITHOUT FEELINGS?

As we have seen, Breazeal and her colleagues openly state that Kismet and Leonardo have emotions, albeit in a more limited sense than humans do. How are we to interpret their claims? I will distinguish two conflicting positions on whether robotic emotions could exist which I will refer to simply as the "Unemotional Robots View (URV)" and the "Emotional Robots View (ERV)." The core claims of URV are that robots cannot have emotions, because they cannot have feelings and feelings are necessary for having emotions. I will reject this argument as inconclusive, and defend ERV instead by specifying in what sense robots like Kismet and Leonardo have emotions. At the same time, I will acknowledge that there are important differences between human and robotic emotions, namely that robots currently lack any sort of subjective feelings. In addition, Kismet and Leonardo's emotional capacities parallel those of human infants, and thus lack I will support ERV, and provide arguments against URV, which holds feelings are necessary for having emotions. I conclude with a discussion of how Kismet and Leo-

nardo bear the necessary components for having emotions in a functional sense, and the ways in which their emotions differ from humans.

4.1 The Unemotional Robots View

Critics of affective computing have claimed robots will only ever be capable of the *appearance* of having emotions on the grounds that robots do not and may never have subjective feelings (Picard, 1997; Velik, 2010; Turkle, 2010; Feil-Seifer & Matric, 2011). This view permeates both the academic community and popular culture, as the possibility of machines having the capacity to feel sorrow after the loss of a close friend, or to feel fear about death, clashes with the notion that machines obey a fixed set of mechanistic instructions. While robots and virtual agents can certainly give the impression they have emotions through producing expressions indicative of an emotional state, Adolphs argues "such constructions would be rather limited and susceptible to breakdown of various kinds" (2005, p. 10). That a robot is capable of producing behaviors often classified as emotional, such as smiling, frowning, retreating from intense stimuli, etc., is not sufficient reason for thinking the robot has emotions, according to Adolphs. He states, "Behavior, physiological response, and feeling causally affect one another; and none of them in isolation is to be identified with the emotion, although we certainly use observations of them to infer an emotional state" (p. 16). Although 'feeling' is a notoriously difficult term to define, Adolphs (2005) describes a feeling as "one (critical) aspect of our conscious experience of emotions, the aspect that makes us aware of the state of our body - and through it, often the state of another's body" (p. 21). Feelings thus serve a monitoring function where the agent consciously recognizes their emotional state and uses it to make inferences about how that state relates to the environment. For a robot to have feelings, it must replicate "the relevant internal details at a

level below that of radio transmitters but above that of actual organic molecules" (p. 12). Although Adolphs concedes affective science is still unsure what these relevant internal mechanisms are (p. 12), he posits the mechanism should be capable of maintaining a self-model that is recursively updated in response to new information in the environment and available to other cognitive processes.

Adolphs (2005) contends the view that feelings are not necessary for defining emotions is problematic for two reasons: First, if feelings are no longer part of our conception of emotions, then there is little way of telling which behaviors were onset by a particular emotion, according to Adolphs. He contends that since any number of behaviors could be enacted when one is in a given emotional state, it would be difficult to say whether such a state was present without feelings to serve as a reliable indicator of that state. Adolphs is apt to note that behavioral equivalence between two individuals does not guarantee internal equivalence; there is an indeterminate number of different behaviors that could be carried out as the result of an emotion, say anger, such as by cursing an offending party, physically assaulting that party, or even silently seething in the confines of one's bedroom. It is also possible for two individuals to behave the same way, despite having different emotions. For instance, Cathy may cry because her favorite pet has passed away, or because she is overwhelmed by having just won the lottery. As two or more individuals could execute the same kinds of behavior while bearing distinct internal states, Adolphs maintains an account of emotion must include experience of emotions, since this experience gives an indication of why the behavior was performed. With regard to robots, it is entirely possible for them to execute behaviors that appear emotional, such as fleeing from a threatening stimulus, without undergoing an emotional state. Thus, Adolphs holds feelings are a necessary component of emotions.

Second, Adolphs claims by leaving considerations of conscious experience out of the components necessary for constructing an emotional robot, one is committed to a behaviorist account (p. 10). Once a strict behaviorist perspective on emotions is adopted, feelings are no longer attributed any causal power over the agent's action selection. He argues this cannot be the case, as feelings serve two clear functions: 1) to monitor one's current emotional state so that one can take action to regulate it, and 2) to empathize with others so as to understand their emotional state. Adolphs likens feelings to a means for obtaining information about one's own emotional state, by serving as representations of one's physiological changes and how they correspond with changes in the environment (p. 21). If one registers an experience of disgust after smelling fish, for instance, then one is likely to consciously avoid consuming fish in the future. In addition, research supports the theory that people understand another's emotional state by simulating that emotional state within themselves (Adolphs, 2002; Jeannerod, 2005). Studies have demonstrated that observing the facial expression of another can cause changes in one's feelings (Schneider, Gur, Gur & Muenz, 1994; Wild, Erb, & Bartels, 2001). By experiencing another person's emotion, one's subsequent actions become attuned to the social circumstances. For instance, if Gerry makes a comment about Annette that causes her to furrow her brow and clench her fists, Gerry may recognize she is angered by his comment by simulating Annette's emotional state using his observations of her body state. On his account, feeling empathy towards another involves representing the physiological changes of another person, which helps one to understand their emotional state. As feelings allow one to monitor the emotional state of one's self and others, and constitute the experiential aspect of emotions, Adolphs believes feelings are necessary for an agent to have emotions proper. According to Adolphs, a robot could " have emotions in a narrow sense in the absence of feelings. However, such constructions would always be rather limited and

susceptible to breakdown of various kinds" (p. 10). What Adolphs means to say is that without feelings, robots that are capable of convincingly reproducing the facial expressions, posturing and behaviors that are indicative of emotions in humans will eventually reveal that these expressions and behaviors were designed with the intent of fooling humans into believing the robots have emotions, when in fact their actions are just mediated by a set of clever algorithms (p. 18). Adolphs reasons that since these capacities for expression and behavior associated with emotion are explicitly built into the robots, that there likely is some unanticipated situation(s) in which the system responsible for generating convincing responses to social situations will fail, showing that the robot was merely simulating emotion.

Adolphs presents yet another criticism of robotic emotions: even if the robots were to execute behaviors that corresponded with what would be expected of a human in a similar scenario, the robot would still lack the biological, and more specifically, neurological underpinnings that produce those behaviors in humans (2005, p. 12-13). A robot that was able to fool an interrogator on the Turing Test into thinking it was a human still would not necessarily possess the internal mechanisms that would have produced those same responses by a human on the test. Without undergoing physiological changes that are roughly equivalent to humans' during an emotion, Adolphs asserts robots cannot have emotions. While Adolphs does not explicitly state that robotic emotions are impossible, it is clear he thinks building a robot with an internal structure capable of replicating human physiological changes and changes of feeling is highly implausible.

Arbib (2005) espouses a similar position to Adolphs: He states that while robots may be able to make emotional appraisals, they cannot experience the "heat" of subjective feelings, since they lack any foundation in biological evolution (p. 374). While Arbib does not provide an explicit definition, his discussion suggests "heat" is the visceral experience aspect of emotion ena-

bled by the particulars of human biology. Arbib believes it is highly unlikely robots will ever be designed to experience this aspect of emotions, due in part to the difficulties inherent in replicating the biological system(s) necessary for feeling. A similar criticism is furthered by Velik (2010) who holds that robots will always lack emotions unless their physical implementation includes a visceral system analogous to humans'. On her account, a robot's perception of a change in its body does not qualify as a feeling; the robot must experience the sensation of a bodily change in the way humans do. According to Velik, robots will lack emotions until a human-like visceral system is integrated in the robots' physical implementation.

Arbib also sees robotic feelings as implausible due to the difference in ecological niche between robots and humans. Like all biological life, humans are thought to be motivated by the four Fs: feeding, fighting, fleeing and reproduction (Pribram, 1960). Robots, in contrast, are not impelled by these drives, and need not be, as these machines are presumably engineered with other functions besides replicating all aspects of living organisms in mind. Rather than survival, the robots' most basic drives will be tailored to the specific tasks they are designed for. Arbib also notes many robots do not have control and sensory systems that are able to recognize human emotional expressions, and are thus incapable of empathy. As empathy allows one to recognize and understand another's emotional state, many emotion theorists take empathy to be a necessary capacity for having emotions at all (Stueber, 2006). Although Arbib is noncommittal about whether a robot lacking empathy could be said to have emotions, it is apparent he does not believe they can in any deep sense.

4.2 The Emotional Robots View

Adolphs is apt to note that just because two agents behave in the same way does not guarantee that the internal mechanisms controlling their behavior are identical (2005, p. 12). As behavior does not guarantee the presence of a particular mechanism that generates it, Adolphs asserts that observation of behavior normally associated with emotion is not sufficient to conclude the agent has emotions. Nevertheless, behaviors do provide a useful heuristic for estimating another's emotional state, even if they cannot guarantee accuracy. When someone is angered, for example, they will typically engage in aggressive behavior in retaliation to the offense committed against them. An angered individual who smiled at the person who had insulted them and offered to shake their hand would be in conflict with what anger is conceived to be: a retaliative response to a perceived offense (Videbeck, 2006). Although aggressive behavior is more commonly recognized as associated with anger, withdrawal behaviors can also follow an appraisal that leads to anger (Novaco, 1986). These two types of behavior give the observer an initial indication of the emotional state of the person they are observing. As Adolphs would likely note, however, fear has also been shown to elicit withdrawal behavior as well (Buss et al., 2003). How then can emotional state be inferred? I argue there are two other contributory ways for determining an agent's emotional state: by observing their expressions and by observing their neurophysiological changes to a stimulus. Of course, robots do not have neurophysiology, as they are not biological[4], so robots must be constructed with systems analogous to the neurophysiological systems of humans. The question that then follows is this: How similar must the control system (i.e. the robot's 'nervous system') and body of the robot be in order for the robot in question to have emotions comparable to human emotions? The answer to this question, I argue, is that

[4] It is possible for hybrid agents with both biological and artificial components comprising their control system and body to exist (i.e. cyborgs), and in fact, these sorts of agents already do exist (e.g. in individuals with pacemakers or neuroprosthetics). However, I will not consider these sorts of agents in this thesis.

the robot's control system and body must consistently prepare the robot for the same sorts of action tendencies that human neurophysiology prepares humans for, such that observations of control system and bodily changes reliably indicate emotional appraisals and behavior.

Before determining how changes in the control systems and bodies of robots designate emotions, let us first consider how facial expressions and neurophysiological changes serve as reliable indicators of emotional states in humans. As Paul Ekman (1971, 1993, 1997) has argued, there is evidence to suggest there are particular facial expressions that correspond with particular emotions, and that these facial expressions can be recognized cross-culturally. If this is the case, then we are able to distinguish an individual who is afraid from an individual who is angry, as the facial expressions characterizing the respective emotions are markedly distinct. Anger and fear share common physiological responses, such as elevated heart rate and increase in perspiration (Ekman, 2004), though differ in the level of neuromodulators observed during the respective emotional states (Novaco, 2000). Norepinephrine is more strongly activated during anger than epinephrine, while epinephrine is the more dominant neuromodulator during fear. In addition, neuroimaging studies have found the lateral orbitofrontal cortex to be the brain structure most commonly implicated in anger (Portegal & Stemmler, 2010), while the amygdala are argued to play the central role in fear processing[5] (LeDoux, 2002; Olsson & Phelps, 2007). Taken together, observations of behaviors, expressions and neurophysiological changes provide an indicator of the observed individual's emotional state, without needing to speculate on that individual's subjective experience. Thus, Adolph's first criticism can be dismissed as applied to humans.

How then can bodily changes and facial expressions of robots serve as reliable indicators of emotional states in robots? This question raises the further question of what role bodily changes play in the emotions of humans. As was established in Chapter Two, bodily changes fol-

[5] The amygdala have also been implicated in anger, as well as a number of other emotions (Phelps & LeDoux, 2005)

low from an appraisal of a salient stimulus in order to prepare the agent for action (Scherer, 2005). Likewise, the robot's appraisal of a given stimulus should elicit a cascade of bodily changes in order to prepare the robot for an action that is an appropriate response to the appraisal (i.e. a response that will assist the robot in reaching its goal). Furthermore, there should be unique cascades of bodily changes for each of the robot's emotions, as is the case with humans. So long as these conditions for bodily changes are fulfilled, the physical differences in how bodily changes occur between robots and humans are irrelevant, as bodily changes would serve as reliable indicators of emotion in both types of agents. In addition, the emotional robots should have facial expressions that both correspond with their emotional state and that are easily recognizable to humans as indicative of those emotional states. As there are no apparent physical limitations on designing these sort of robots, Adolphs' first criticism may also be dismissed as applied to robots.

Adolphs' second criticism misses a critical distinction between the conception of emotion held by theorists in agreement with Fellous and LeDoux (2005), and the behaviorist position when he equates the two positions: the approach to understanding emotions proposed by Fellous and LeDoux, Rolls, and like-minded researchers relies on observations of neurophysiological state, expressions, and behavioral changes in order to draw conclusions about the agent's emotional state. In contrast, the behaviorist simply refers back to the agent's past responses to particular stimuli in order to predict future behavior (Graham, 2010). The behaviorist argues inner states of the agent should not be used to explain a given behavior, as the behaviorist is committed to the notion that external factors (i.e. the environment) determine behavior. As cognitive neuroscientists postulating on how the brain structures implicated in emotional processing relate to attention, memory, and other processes, Fellous and LeDoux are certainly far from committed to

anything resembling behaviorism; in fact, they maintain an opposing position. The approach they espouse claims emotions should be understood as largely unconscious processes originating in well-defined neural circuitries. Thus, Adolphs mischaracterizes the position of Fellous, LeDoux, Rolls and like-minded researchers when he states they define emotion as "behavior without conscious experience" (2005, p. 27).

In humans, feelings do appear to serve the function of monitoring changes in one's emotional state over time relative to environmental events as Adolphs suggests. However, this is not sufficient reason to think such a monitoring function is necessary in order for an agent to achieve its goals or to maintain a state of well-being. Breazeal argues Kismet's "emotive system is responsible for perceiving and recognizing internal and external events with affective value, assessing and signaling this value to other systems, regulating and biasing the cognitive system to promote appropriate and flexible decision making, and communicating the robot’s internal state (p. 292)." None of the robot's deliberation on action selection is conscious; the selection of a behavior for execution is determined by an algorithm which evaluates the particular behavior that has received the most activation from Kismet's active drive and emotive state. For instance, if Kismet's drive for rest is active and a human interactant presents a toy to the robot, Kismet's emotive state corresponding with disgust will activate, causing the robot to turn up the corner of its mouth and turn away from the toy. This informs the human interactant Kismet is not interested in playing. Thus Kismet fulfills its drive to rest. This however brings up the question of whether it is appropriate for Breazeal to refer to the internal state of Kismet as an "emotive state." If Breazeal is correct in referring to Kismet's internal state as an emotive one, as I argue in the next section, then it would appear as if it is entirely possible for robots to achieve their goals and maintain a state of well-being without conscious deliberation.

While Adolphs (2005) and Arbib (2005) are right to note the lack of biological grounding apparent in most current robots (see Warwick et al., 2010, for discussion of a robot-rat brain hybrid), there is no principled reason to think biological materials are the only materials capable of producing the functions emotions serve. If we are committed to the physicalist notion that all psychological phenomena originate from the interaction of physical entities (Stoljar, 2009), then there appears to be nothing contradictory in asserting the possibility of designing mechanisms that reproduce emotions in robots, even artificial feelings, so long as researchers and engineers have a sufficient understanding of what interactions need to occur between the agent's body and the environment for such phenomena to arise. This is not to deny the challenges inherent in replicating something as intricate as the human visceral system, but rather to acknowledge its physical possibility. Furthermore, it is not clear to what extent robots would have to have a physical implementation parallel to the human body in order to have emotions, despite Arbib (2005) and Velik's (2010) insistence. While I concede emotions are exceptionally complex processes, there are a number of sophisticated functions and abilities that can be performed by entities with little resemblance to each other: a human chess player might use associative memory to recognize how the current state of the board resembles a board state of a previous game and choose a move based on what worked (or did not work) in the previous game. In contrast, a computer player will likely employ an algorithmic search for a move that has the greatest probability of minimizing the possibility of an opponent checkmating them (Shannon, 1950). Unless the goal is to model human emotions with all the exact biological details, then the possibility that emotions could be functionally realized is closer in reach than thought by proponents of the URV.

Arbib's (2005) contention robots should have empathy to be considered emotional is endorsed by Scherer (2010) who argues an agent is competent in perceiving emotions when she can

recognize another's emotional state through observations across several sensory modalities (i.e. face, posture, vocal intonation, verbal content), despite some cues being hidden from the observer (p. 8). While many theorists assume subjective feelings are necessary for having empathy, Damiano and colleagues (2011) argue empathic relations can be achieved between humans and robots, so long as the emotional states between these agents are coordinated. By coordinated, she means the emotional expression of one agent influences the expression displayed by the other agent in an exchange, and vice versa. For instance, an enraged individual who confronts their housemate for never completing household chores might evoke a startled expression in the housemate who was not expecting the dispute. The housemate's expression may shift to guilt after realizing their neglect, which in turn could cause the angered individual to soften to a more neutral expression. Damiano asserts emotional expression is an ongoing process in humans that is not caused by a feeling, but rather precedes feelings, as some have contended (Dumouchel, 2008). The central idea is humans are embodied in the environment such that their emotional expressions are contingent upon their interaction with a communicative partner, particularly the partner's own expressions; emotions are thus relational processes as their formation requires first observing an agent or event for the emotions to be directed towards. Damiano and colleagues conclude a robot that is able to coordinate a human's expressions through its own expressions, and is likewise influenced by human expressions in a reactive fashion has the necessary capacities for artificial empathy.

Damiano and colleagues are right in asserting the emotional expressions of others directly influence the expressions one produces as a consequence. When Agent A starts rapidly approaching Agent B with a fixed stare and furrowed eyebrows, Agent B either takes action to avoid the oncoming aggressor or else Agent B matches Agent A's level of aggression. In either

case, the fact that Agent A has expressed aggression towards Agent B necessitates Agent B sending a reciprocal message back to Agent A. We would suspect a deficiency in the emotional capacities of an individual who, for example, did not return a smile when smiled at by others, or who displayed a blank stare and replied in a monotone voice when verbally berated, since humans are inclined to respond to another's expressions. What is more, there is reason to think these exchanges in affective expression are reflexive to one's emotional state. Ekman (1997) holds that once an affect program (i.e. a preprogrammed, automatic response to relevant events both shared and unique across cultures (Ekman & Cordaro, 2011)) is initiated, the corresponding facial muscles are directly stimulated, meaning these facial muscles will contract every time the affect program activates (p. 324). Although individuals may attempt to suppress expressions by imposing voluntary control on their face, Ekman maintains they cannot prevent the neurological signals from being sent to the corresponding facial nerves of the expression. This is not to suggest that Agent B has a fixed expressive response to Agent A's display of aggression, or that all other agents will respond with the same expressions as Agent B. This is rather to contend that expressions are displays produced as reactions to another’s actions or expressions, which bear information both about the expresser's social intentions (Fridlund, 1997) and their emotional state (Ekman, 1997). A robot that is capable of recognizing and classifying the emotional state of a human, and as a result is able to tune its own expressions to match those of a healthy human, should be said to function in a socially-appropriate way. As Leo is able to adjust its expressions in this manner through imitating human interactants (Breazeal, Buchsbaum, & Gray, 2009; Berlin, Gray, Thomaz, & Breazeal, 2006), and using them to social reference (Thomaz, Berlin, & Breazeal, 2005), Adolphs and Arbib's criticisms appear unsubstantiated.

From the account of expressional exchanges provided above, it appears as if an individual can have an unconscious emotion in the absence of a feeling. However, a critic of this position might assert that feelings are still present within an emotional episode, even if one is not able to sense them. In response to this discrepancy, Lacewing (2007) organizes the positions held on the relationship between unconscious emotions and feelings into three groups. The first group holds that unconscious emotions do involve conscious feelings. Ben-Ze'ev (2000) and Greenspan (1988) maintain that individuals are aware of a conscious feeling during an unconscious emotion, but do not understand the connection between the emotion and the feeling. In contrast, Goldie (2000) asserts that unconscious emotions are consciously felt, despite that the individual remains unaware of the conscious feeling. The second group deny that unconscious emotions require feelings at all. Nussbaum (2001) argues that emotions are effectively judgments that do not involve any non-cognitive processes, such as feelings. Roberts (2003) defines feelings as immediate perceptions of one's self as being in a certain emotional state, but contends emotions need not be consciously recognized. On his view, feelings do not occur when the emotions they correspond with could be psychologically damaging, or else masking feelings are elicited that distract from the emotion (e.g. feelings of superiority to mask fear of failure [Lacewing, 2007, p. 95]). The third group contends unconscious emotions have unconscious feelings. Gardner (1993) and Wollheim (1984) defend this position on the grounds that emotions are useful to us only if they are felt. These authors argue that the experience of a mental state and its representational content are inseparable, yet their presence need not be detected by consciousness in order to persist. As these three accounts appear equally plausible from his perspective, Lacewing concludes no answer to the question of how feelings relate to unconscious emotions yet exists. However, since we are able to provide an explanation for how an unconscious emotion alone can

moderate social interactions and select appropriate actions, I argue feelings are not necessary for characterizing unconscious emotions.

There are a number of researchers who hold that feelings are too poorly understood to be included within the broader definition of emotion at all. In addition, many theorists do not consider feelings strictly necessary for having emotions. Fellous and LeDoux (2005) insist the study of emotions should remain focused on the neurophysiological aspects of emotion that are well-defined, such as the circuits implicated in emotional processing related to fear (p. 105). Within a fear conditioning paradigm, subjects are presented with a painful unconditioned stimulus (e.g. a foot shock) immediately following a neutral conditioned stimulus (e.g. a tone), such that they learn to fear the conditioned stimulus when it is presented in isolation. The researchers are then able to operationalize the test subject's fear through measurements of blood pressure, freezing responses, pituitary-adrenal stress hormones and activation of the specific processing circuits implicated in fear (p. 87). Through these studies, LeDoux and colleagues have found the amygdala to play a central role in the coordination of fear responses, particularly through signals output to the brainstem. Fellous and LeDoux argue a greater understanding of the neural circuitry involved in facilitating fear responses enacted by the body could provide insight into how a "fearful" robot might be constructed.

Feelings are exceedingly difficult to study empirically as subjective experience cannot be directly observed. If feelings are necessary in order to say an agent has emotions, then the question of whether nonhuman animals have emotions remains unanswered, as researchers have no means of accessing these creatures' phenomenology (LeDoux, 2002). This is to say, if an organism's subjective experience cannot be directly observed, then we have deficient grounds for knowing whether the organism actually underwent such an experience. While we cannot be cer-

tain our fellow humans have subjective experience based on this conditional, we have sufficient reason for believing a friend is actually excited about an upcoming concert, for instance, because they are able to describe their excitement at various levels of detail. If pressed to explain why they are excited about the concert, the friend may say they enjoy the style of music the performers will play, the fact that other friends will be present, and that there will be an after-party following the show. This is not to suggest individuals' introspective accounts of their feelings are entirely veridical, however. On the contrary, a number of studies have shown people to be inconsistent in reports about their motivations for a particular behavior (Nisbett & Wilson, 1977; Wilson, 2002; Pronin, 2009). A recent experiment (Johansson, Hall, Sikström, & Olsson, 2005) demonstrated 74% of participants failed to detect when a choice they made had been switched for another option. Furthermore, participants provided confabulated accounts for why they made a choice they never actually made. This failure to detect changes in choices one has made, known as choice blindness, was found for choices of facial attractiveness (Johansson, Hall, Sikström, & Olsson, 2005; Johansson, Hall, Sikström, Tarning & Lind, 2006), tea and jam preference (Hall, Johansson, Tärning, Sikström, & Deutgen, 2010), and moral judgments (Hall, Johansson, & Strandberg, 2012). Such findings undermine the assumption that introspection provides accurate accounts of one's subjective experience. That said, the fact humans are able to provide these sorts of accounts at all differentiates them from other animals for whom we have only physiological and behavioral measures to assess their mental state. Thus, while subjective feelings appear to be a part of human emotions, we do not yet have a clear enough understanding of feelings to assume they are necessary for any agent to have emotions.

5 ROBOTS CAN HAVE EMOTIONS

Thus far, I have provided rebuttals to the existing arguments for why robots cannot have emotions, focusing in particular on dismissing the claim that robots cannot have emotions because emotions require feelings, which robots cannot possess. My critique of this Unemotional Robots View has relied on an analysis of the positions defended by Adolphs (2005), Arbib (2005), and on the literature on unconscious emotion that evaluates the role of feelings in emotions. In this section, I will argue robots can have emotions, and in fact, Kismet and Leonardo already do. As all contemporary theories of emotion remain subject to unanswered criticism as I illustrated in Chapter 1, I propose that the best way of understanding emotions is through the component process model of emotions developed by Scherer. This theory builds upon the consensus among theories as to what components comprise emotions, despite the emphasis these theories place on particular components. The component process model of emotions also provides a clear advantage over other theories of emotion as it accommodates for the wide variety of ways in which emotions are instantiated. Most importantly, the component process model allows for the conclusion that robots can have emotions even if they do not have feelings, as only a majority of the components of emotion need be present within an emotional episode on this account. I will then demonstrate that Kismet and Leonardo both instantiate emotion through four of the five components: appraisals, physiological changes, expressions and action tendencies. I will also briefly discuss the additional features of Leonardo that allow the robot to form beliefs about others' intentions with which the robot can perform social and task learning. I conclude with a discussion of what Kismet and Leonardo's capacities mean to the possibility of robots having emotions at large and suggest future directions for exploring affective computing.

5.1 Component Process Model of Emotions

As stated earlier, most theorists and affective scientists are in agreement on the components that comprise emotions, despite the fact that no consensus has emerged on what emotions are. The component process model of emotions (Scherer, 1987; 2000; 2005; 2010b) characterizes emotions as a series of coordinated responses driven by an appraisal of the environment across multiple objectives. Scherer states an "emotion is defined as an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism (Scherer, 1987, 2001)." On Scherer's account, the states of the subsystems he refers to are the components of emotion: cognitive appraisal, neurophysiological changes, action tendencies, expressions and feelings (2005, 2010b). The sorts of responses executed by the coordination by the components depend upon how salient internal and external stimuli are appraised across multiple objectives. The first appraisal objective is to determine how relevant a given stimulus is to the agent or other members of its group (2010b, p. 51). Satisfaction of this objective includes checks for the novelty and intrinsic pleasantness of the stimulus, in addition to its relevance to the agent's goals and needs. The second related objective is to evaluate how the stimulus might affect the agent's survival, well-being and long-term goals. To meet this objective, the agent appraises the sorts of causes and probable outcomes of a given stimulus and assesses whether dealing with it is a high or low priority. Third, the agent appraises the ways she could cope with the stimulus and what the consequences of such actions would be. This appraisal motivates the action tendencies the agent will later carry out. The final appraisal objective is to determine how the stimulus relates to the agent's self-concept as well as social norms and values. By forming appraisals that take into consideration all the details that are important to the agent's well-being,

the agent is able to flexibly coordinate the remaining components of emotion so as to motivate adaptive responses to the stimulus in question[6].

It is important to note that Scherer concedes an episode that involves changes in only a majority of the components qualifies as an emotion, so long as the synchronization of these components produces an adaptive response to a significant event (2010b, p. 49). Thus, contrary to traditional accounts which hold emotional categories have necessary and sufficient conditions (Clore & Ortony, 1991), the component process model recognizes that not all the components need to be instantiated for an emotion to occur. Although feelings are important on the component process model account for monitoring the agent's internal state and their interaction with the environment, the agent can still have emotions without feelings, so long as she is able to coordinate adaptive responses to stimuli appraised as relevant to her needs, goals, and values. As Scherer (2010b) asserts, the component process model of emotions provides a framework for which computational models of emotion can be constructed and conceptualized. By embodying computational models of emotion that accord with the component process model within a robot, emotions are no longer simply modeled, but instantiated through the robot's interaction with the physical environment.

5.2 Kismet and Leonardo Have Emotions

Kismet demonstrates its capacity for appraisal through its ability to assign valuations of arousal, valence, and stance to low-level percepts it forms through observing salient stimuli, in addition to determining whether the stimuli are undesired/desired/threatening. Like humans, the robot compares its appraisal of the environment to its internal state (active drives, emotions) in

[6] Scherer is clear that most of coordination of the components of emotion occurs unconsciously on different processing levels that work in parallel (2010b, p. 47).

order to determine how to coordinate future action, expressions, and overall emotional state. It is important to note these appraisals pertain only to toys and human faces, as those are the only objects Kismet can classify. In Kismet's world, these are only types of objects that matter, and everything else is undesired or threatening. Kismet was modeled after prelinguistic infants (0-13 months; Shaffer, et al., 2002), which are not expected to have classification schemes for objects beyond animate and inanimate at this young of an age (Kuhlmeier, Bloom, & Wynn, 2004), so Kismet's limited categorization is fitting. As Kismet's appraisals of the environment directly influence the actions and expressions the robot executes, Kismet satisfies requirements for the first component of emotion.

Although Kismet lacks many of the neurophysiological systems that undergo change in humans (i.e. a biological brain, circulatory and endocrine systems), in addition to lacking limbs and appendages for manipulating the environment (the robot is only an anthropomorphic head!), the robot may still be said to adapt its face and neck configuration after an appraisal has been formed. As noted earlier, as the robot's appraisal of the stimulus' arousal increases, Kismet's ears move upwards, while its mouth and eyes widen. Likewise, as arousal declines, Kismet's ears sink down, its mouth closes and its eyes shut. This is to say the appraisals Kismet forms have direct influence over the operation of motors in its ears, eyes, mouth and neck. Furthermore, these adaptations in body configuration characterize the fulfillment of action tendencies. The process by which a particular action is selected to be carried out is best illustrated by an example: If Kismet's social drive is active and there are people present in the room, Kismet's attention will be biased towards observing their faces, and the robot will ignore stimuli perceived as toys. Already Kismet's distinguishing between people and toys motivates the robot's attention, which involves manipulating the motors controlling its eyes. An individual that is appraised as desirable

(positive valence, approach stance) will further the action tendency to engage that person, while an individual who is too loud and/or approaches Kismet too rapidly will be appraised as intense, and will encourage avoidance tendencies. Kismet's demonstrates its ability to form action tendencies in the sense that the robots appraisals of the environment are weighed in conjunction with its active drives and emotive state to determine the action that is carried out. (p. 291). For example, if the individual receives high valence and stance (approach) ratings, Kismet will then assess their proximity, which determines whether the robot calls out to the person or engages in vocal play with them. Thus, through forming appraisals of salient stimuli, Kismet is able to deliberate among action tendencies to execute the behavior most congruent with achieving its current goal. What is more, Kismet is able to interact socially with humans through its facial and vocal expressions such that the robot's intentionality is clear to its interactants. By displaying sorrow when a desired toy is taken away or by turning away from a rapidly approaching stimulus, Kismet encourages its interactants to modify their actions in order to maintain the robot's state of well-being. Thus, Kismet can be said to instantiate emotions through forming appraisals of salient stimuli, undergoing bodily changes, producing expressions and carrying out action tendencies to facilitate the completion of its own goals.

As Leonardo bears a similar affective-cognitive architecture to Kismet, the robot is able to perform the same functions as its predecessor, but to an even fuller extent. First, Leonardo has a body with moving arms and hands with which it can manipulate the environment. This means the robot is able to formulate a far wider range of action tendencies based off its appraisals of itself and other objects in the world. Second, Leonardo is able to appraise whether a stimulus is novel or has been encountered before, in order to determine whether any appraisals have been made of the stimulus in the past that could be referenced (Breazeal & Gray, 2009). Third and

most importantly, Leonardo is able to able to maintain ongoing beliefs about another's emotional state by imitating another's expressions so as to invoke an analogous emotional state within itself. This additional capacity allows Leonardo to have an understanding of its interactants behavior, which has been shown to promote task learning in the robot (Breazeal, Buchsbaum, & Gray, 2004; Thomaz, Berlin, & Breazeal, 2005). On the component process model account, Leonardo satisfies the first three appraisal objectives outlined by Scherer (2010b) as the robot (1) assesses the relevance of stimuli, (2) determines how they might affect the robot's 'well-being' and long-term goals, and (3) evaluates how best to cope with these stimuli as a result. As Leonardo is able to use its multi-level appraisals to coordinate the components of its emotional system in order to deal with stimuli within a dynamic environment, the robot qualifies as emotional, according to the component process model.

In sum, there are already clear indications of emotions present in some robots, such as Kismet and Leonardo, and the possibility for expanding these capacities in the design of future artificial agents. These robots are able to use emotions to select behaviors autonomously and communicate with humans to coordinate the achievement of their goals. What is more, Leonardo can utilize the embodiment of its own affective system to simulate the emotional state of others, such that the robot has an empathic understanding of its interactant. With empathy comes a more accurate consideration of how others deliberate among actions, which means empathic robots may coordinate their own behaviors more successfully with humans in order to facilitate cooperation between parties.

These robots are however emotionally limited in the sense that they were modeled after human prelinguistic infants and thus lack the functionality emotions provide to human adults. Kismet categorizes objects as only human faces or toys, and thus lacks awareness of other object

categories that might be relevant to the robots' flexible and adaptive behavior in its environment, such as young humans, old humans, other robots, animals, vehicles, etc. Of course, the sorts of objects that are salient to a given robot should be contingent on the purpose the robot was meant to serve. An emergency response robot, for instance, should have the ability to distinguish between an injured and a non-injured person; the extent of the injury appraised should determine the state of urgency the robot bears. In addition, Kismet and Leonardo are able to evaluate prosody of speech directed at them (e.g. whether it is approving, prohibitive, etc.), but do not understand the semantic content of what is being said. With the ability to extract emotion-laden language from speech, future robots would be able to fully participate in human conversation. This is a vital ability for most robots in the service industry, such as server robots or personal assistant robots, whose occupation requires following explicit verbal directions. Future robots with these improvements to their capacity for appraisal will be able to more fully participate within the human social realm.

Arbib (2005) is astute to note that the difference in ecological niche between humans and robots means that their emotions are not likely to be equivalent. So much of human emotion ultimately results from the satisfaction of or failure to satisfy our inherent drive for reproduction, like all other biological life. While there is little reason to think a drive for reproduction could not be reproduced within a robot constructed with an artificial reproductive system, creating such a reproductive system and drive seems pointless if robots can more efficiently be manufactured on a factory assembly line. However, if the goal is to create a robot for the sake of better understanding human emotions, then it would be worthwhile to include these details within its design.

Furthermore, considerations should be made for future robots' likeness to human bodies and faces if these robots are to be accepted as social partners. It is particularly important that

roboticists are mindful of the uncanny valley effect which states that as robots come to resemble humans more and more closely, human observers will treat these robots increasingly more positive and empathic until a certain resemblance is reached where human observers become strongly repulsed by the robot. While it is currently unclear why the uncanny valley effect occurs, a number of explanations have been proposed: due to an evolved mechanism for avoidance of mates with undesirable traits (Green, MacDorman, Ho, & Koch, 2008; Rhodes & Zebrowitz, 2002), because such entities remind humans of mortality (MacDorman & Ishiguro, 2006), and because these provide conflicting perceptual cues as to the sorts of entities they are (Saygin, 2012) As the robots continue to appear more like humans after passing through the uncanny valley however, human observers' response returns to being increasingly positive and empathic (Mori, 1970/2012). Despite having anthropomorphic facial features, Kismet and Leonardo avoid the uncanny valley effect by being easily distinguishable from humans; Kismet is distinguishable by lacking a body and having a face constructed from metal, and Leonardo by resembling a foreign creature. In contrast, many have considered the android constructed by Hiroshi Ishiguro to resemble himself, known as *Geminoid*, to fall within the uncanny valley, as the android nearly resembles its creator, but is still recognizably different (Becker-Asano, Ogawa, Nishio, & Ishiguro, 2010). Of course, an android that was physically indistinguishable from a human in appearance, movement and by touch would be positioned on the opposite side of the uncanny valley as Leonardo and Kismet, and would achieve maximal empathic relations with humans, according to Mori. An android constructed with limbs that moved closely to the fashion of humans' would achieve further empathy from human interactants as it would be capable of common social and emotional gestures such as handshakes and hugging, which are essential to how humans develop

relationships with each other. By taking these considerations into account with future robots, we can ensure a fuller integration of robots into the social realm of humans.

In conclusion, Kismet and Leonardo can be said to have emotions through their instantiation of the majority of components that comprise emotions, despite lacking subjective feelings. However the emotional capacities of these robots parallel the capacities of human infants and thus, future efforts should be taken to scale their capacities up to the level of human adults. As subjective feelings serve to monitor one's emotional state within humans, future experiments may look to uncover the neurophysiological mechanisms underlying feelings, though this is a more long-term goal for affective science. As the fields of affective science and robotics progress, the recognition of some robots as emotional will continue to grow until the sophistication of their emotional capacities is beyond what can be ignored.

5.3 Why Do Robots with Emotions Matter?

Skeptics of robotic emotions might wonder whether the bar has been set too low for the phenomena that count as emotions. They might argue that robots having emotions as defined by the component process model (Scherer 2005, 2010) is trivial, as the definition of emotions provided the CPM does not match the lay conception of emotions; the lay conception of emotions requires that the agent experiences a feeling during an emotional episode, while the CPM does not. The concern is that if feelings are not required for having emotions on the present account, yet feelings are central to how most people define emotions, that the present account has merely adapted an arbitrary definition of emotions.

There is also a related concern regarding the importance of classifying these robots as having emotions. Even if Kismet, Leonardo, and other like robots have emotions on the present account, what is the theoretical advantage of categorizing them as emotional? Categories allow

observers to make inferences about the entity in question, based on what categories it falls within. For instance, classifying a whale as a mammal allows the observer to conclude that whales give birth to live young, are warm blooded, and the females have mammary glands, among other features. So what then can be inferred about these robots based on their having emotions?[7]

When considering whether the present account provides an arbitrary definition of emotions, it is important to note that this definition was derived from a shared theoretical understanding of what emotions are. While there is no consensus among researchers as to what components of emotions are the most important for the occurrence of an emotional episode within humans, there is general agreement on what components are involved; these components are appraisals of events, neurophysiological responses, action tendencies, expressions, and feelings. However, despite the insistence that feelings are necessary for having emotions on the lay account, there is disagreement among researchers as to whether this is actually the case. As was argued earlier, there are a number of researchers who hold that feelings are too poorly understood to be included within the broader definition of emotions at all (see Fellous & LeDoux, 2005), largely because there is no means for directly accessing an individual's phenomenology (LeDoux, 2002). More important is the fact that these robots are able to continually perform what I argue is the central function of emotions: to attribute value to stimuli in the environment with respect to how those stimuli relate to the agent's goals/needs, in order to autonomously select an action or set of actions to satisfy those goals/needs. While other robots may be capable of fulfilling goals through non-emotional processes, Kismet and Leonardo utilize the coordination of the components of emotion in order to meet their goals through emotional processes. Thus, these robots have emo-

[7] I would like to extend a special thanks to Neil Van Leeuwen and Dan Weiskopf for bringing these concerns to my attention during the defense of this thesis.

tions in a non-arbitrary sense, even if the layperson may not recognize Kismet and Leonardo as having emotions.

The theoretical advantage of classifying Kismet and Leonardo as emotional is that doing so allows researchers to treat these robots as a viable platform for studying how emotions occur in humans and in other biological life. While it is true that human emotions and the emotions of these robots are far from equivalent, it is important to remember that members of a category need not share every feature in common. For instance, platypuses lay eggs even though they are classified as mammals. If one were to make the prediction that platypuses give birth to live young based on their status as mammals, one would be incorrect. Nevertheless, platypuses' status as mammals allows one to correctly infer that they are warm-blooded, have fur, and breath through lungs. Therefore, a category can be useful, even if some of its members have exceptional features or lack certain features. As the components of emotion are coordinated within these robots such that they are able to perform the central function of emotions, predictions may be made about how the absence or malfunction of one of these components will affect the agent's ability to perform this central function. The capacities of these robots also show that feelings are not necessary for autonomous appraisal and action deliberation towards the fulfillment of goals/needs. Furthermore, the difference between human emotions and the emotions of these robots gives researchers an indication of the sorts of mechanisms that must be added to the robots' architecture, such as pain receptors and a means for determining novel actions (i.e. actions that are not preprogrammed), in order for the robots' emotions to more closely resemble humans. Therefore, despite their dissimilarity to human emotions, there is sufficient similarity in the emotions of these robots to use them as a means for studying human emotions.

6 CONCLUSION

In this thesis, I have argued that it is possible for robots to have emotions, despite lacking subjective feelings. I first demonstrated that while there is no consensus among contemporary theories of emotion on what emotions are, a number of components of emotion consistently appear within the literature: appraisals of events, physiological changes, action tendencies, expressions and subjective feelings. Second, I dissected the affective-cognitive architecture of MIT's Kismet and Leonardo, two robots explicitly designed to express emotions and to interact with humans, in order to explore whether they have emotions. Before concluding on their emotional capacities, however, I first considered objections to the possibility of robots having emotions. The majority of these objections were based on the assertion that feelings are necessary for having emotions, and that it is impossible/implausible for robots to ever have feelings due the differences in physiology between humans and robots. I contended that while reproducing the exact details of human physiology that allow for subjective feelings is currently beyond our available modeling techniques, socially-competent robots can be constructed without having subjective experiences of emotion. More importantly, I argue in support of the component process model of emotion (Scherer, 2010), which asserts an agent may be said to have emotions if she can instantiate at least a majority of the components of emotion. As Kismet and Leonardo are able to instantiate all of the components of emotion except for feelings, I hold that they qualify as emotional agents, albeit with emotional capacities that are not equivalent to human emotions. Since these robots demonstrate emotional capacities analogous to human infants, I suggested ways in which future robots may be extended and scaled up in order to have more prosperous interactions with human adults.

It is crucial to make clear the distinction between robots that have human emotions and robots having emotions at all. In the former sense, it is trivially true that robots cannot have emotions, as robots are not humans. To have human emotions, a robot would have to have an equivalent physiology, control structure, ecological niche, influence from human social norms, etc. In essence, to have identical emotions to humans, a robot would have to be human. While humans serve as the central inspiration for emotional robots and virtual agents, it would be impossible to construct a robot identical to humans in every way, and it would make little sense to do so. A robot could be constructed with emotional capacities that allowed it to socialize with humans in a socially-appropriate way, despite differences in its physiology and control structure. Kismet and Leonardo have emotions in this sense: they are able to recognize emotion expressed through facial displays and vocal intonations such that the robots coordinate their subsequent expressions to encourage further participation with human interactants. What is more, Leonardo is able to gain an understanding of its interactants' emotional state by inducing an analogous state through reproducing the interactants' expressions and inputting the feedback into its own emotional system. These robots' emotional capacities make them less frustrating to deal with, which is a central concern of Picard's (2003); in fact, interactants have reported enjoying their play with the robots and a desire to continue a pleasant interaction (Breazeal & Scassellati, 2000; Breazeal, 2002a, 2003a). As robots become more mainstream, it will be an undeniable advantage if humans enjoy their experiences with robots, especially in industries where humans and robots have frequent, intimate interactions, such as in health care and education.

In sum, robots with emotions already exist. Despite having emotional capacities analogous to human infants, these robots are able to attribute value to objects and events in the environment, in order to coordinate the preparation of their bodies for actions that will modify their

relationship to a constantly changing environment so as to further the pursuit of their goals. As Scherer (2010) contends, this capacity to adapt to environmental contingencies for goal-pursuit is widely held within the emotion literature to be central to classifying what emotions are. Thus, if we are interested in constructing robots that are able to autonomously complete tasks in dynamic, dangerous environments (e.g. deep ocean or planetary surface exploration, search-and-rescue; Breazeal, 2005), emotional capacities are of utmost importance. As the fields of robotics, artificial intelligence and emotion theory continue to develop, the possibility of emotional robots will become a greater reality.

REFERENCES

Adolphs, R. (2005). Could a robot have emotions? Theoretical perspectives from social cognitive neuroscience In Fellous J.M. (Ed.) & Arbib, M. A. (Ed.), *Who needs emotions?: The brain meets the robot* (pp. 9-28). Oxford: Oxford University Press.

Arbib, M. A. (2005). Beware the passionate robot neuroscience In Fellous J.M. (Ed.) & Arbib, M. A. (Ed.), *Who needs emotions?: The brain meets the robot* (pp. 333-384). Oxford: Oxford University Press.

Becker-Asano, C., Ogawa, K., Nishio, S., & Ishiguro, H. (2010). Exploring the uncanny valley with Geminoid HI-1 in a real-world application. In *Proceedings of IADIS International Conference Interfaces and Human Computer Interaction* (pp. 121-128).

Ben-Ze'ev, A. (2000). The subtlety of emotion. Cambridge, MA: MIT Press.

Berlin, M., Gray, J., Thomaz, A. L., & Breazeal, C. (2006). Perspective taking: An organizing principle for learning in human-robot interaction. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 2, p. 1444). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press.

Brady, M. S. (2009). The irrationality of recalcitrant emotions. *Philosophical studies*, *145*(3), 413-430.

Breazeal, C. (2002). Regulation and entrainment in human-robot interaction.*The International Journal of Robotics Research*, *21*(10-11), 883-902.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*(1), 119-155.

Breazeal, C., & Scassellati, B. (2000). Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*, 8, 47–72.

Breazeal, C., & Brooks, R. (2005). Robot emotion: A functional perspective In Fellous J.M. (Ed.) & Arbib, M. A. (Ed.), *Who needs emotions?: The brain meets the robot* (pp. 271-310). Oxford: Oxford University Press.

Breazeal, C., Gray, J., & Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *The International Journal of Robotics Research*, *28*(5), 656-680.

Buss, K. A., Schumacher, J. R. M., Dolski, I., Kalin, N. H., Goldsmith, H. H., & Davidson, R. J. (2003). Right frontal brain activity, cortisol, and withdrawal behavior in 6-month-old infants. *Behavioral neuroscience*, 117(1), 11.

Cao, B., Huang Y., Lu J., Xu F., Qiu Y., & Peng Y. (2012). Cerebellar fastigal nuclear GABAergic projections to hypothalamus modulate immune function. *Brain Behavior and Immunity*, 27.

Clore, G. L., & Ortony, A. (1991). What more is there to emotion concepts than prototypes?. *Journal of Personality and Social Psychology*, 60(1), 48-50.

Clore, G. L., & Ortony, A. (2000). Cognition in emotion: Always, sometimes, or never. *Cognitive neuroscience of emotion*, 24-61.

Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, *28*(2), 117-139.

Menzel, P., & D'Aluisio, F. (2001). *Robo sapiens: Evolution of a new species*. MIT Press.

D'Arms, J., & Jacobson, D. (2003). VIII. The significance of recalcitrant emotion (or, anti-quasijudgmentalism). *Royal Institute of Philosophy Supplement*, *52*, 127-145.

Damiano, L., Dumouchel, P., & Lehmann, H. (2012). Should empathic social robots have interiority?. *Social Robotics* 268-277.

DeLancey, C. (2002). Passionate engines. *What emotions reveal about mind and artificial intelligence. New York*.

de Sousa, R. (2013) "Emotion", The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.). Retrieved November 20 2013 from http://plato.stanford.edu/archives/spr2013/entries/emotion/.

Duchenne, G. B. (1990). *The mechanism of human facial expression*. Cambridge university press.

Dumouchel, P. (2008). Social Emotions In Canamero L., Aylett R. (eds.), *Animating Expressive Characters for Social Interaction* (pp. 1-20). Amsterdam-Philadelphia: John Benjamins.

Ekman, P. (1992). Facial Expression of Emotion: New Findings, New Questions. *Psychological Science*, 3, 34-38.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*,*48*(4), 384.

Ekman, P. (1997) Expression or Communication About Emotion. *Uniting Psychology and Biology: Integrative Perspectives on Human Development.* 48, 384-392.

Ekman, P. (2007). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Macmillan.

Ekman, P. E., & Davidson, R. J. (1994). *The nature of emotion: Fundamental questions*. Oxford University Press.

Ekman, P. & Friesen, W.V. (1971) Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17, 124–129.

Ekman, P., Rolls, E. T., Perrett, D. I., & Ellis, H. D. (1992). Facial expressions of emotion: An old controversy and new findings [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*,*335*(1273), 63-69.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic.*Emotion Review*, *3*(4), 364-370.

Ellsworth, P.C., & Scherer, K.R. (2003). Appraisal processes in emotion. In R.J. Davidson, K.R. Scherer, & H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 572—595). New York: Oxford University Press.

Elster, J. (2004). Emotion and action. In Robert C. Solomon (ed.), *Thinking About Feeling: Contemporary Philosophers on Emotions.* Oxford University Press. 19-36.

Farthing, G. W. (1992). *The psychology of consciousness*. Prentice-Hall, Inc.

Feil-seifer, D., & Matarić, M. J. (2011). Ethical principles for socially assistive robotics. *IEEE Robotics & Automation Magazine, Special issue on Roboethics*, 18(1), 24-31

Fellous, J. M. (2004). From human emotions to robot emotions. *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence*, 39-46.

Fellous, J. M., & Ledoux, J. E. (2005). Toward basic principles for emotional processing: What the fearful brain tells the robot. *Who needs emotions*, 79-115.

Fischer, H., Wright C., Whalen P., Mcinerney S., Shin L., & Rauch S. (2002). "Brain habituation during repeated exposure o fearful and neutral faces: a functional MRI study". *Brain Research Bulletin* 59.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, *18*(12), 1050-1057.

Freitas-Magalhães, A. (2012). Facial expression of emotion. In V. S. Ramachandran (Ed.), *Encyclopedia of Human Behavior* (Vol. 2, pp.173-183). Oxford: Elsevier/Academic Press.

Fridlund, A. J. *Human facial expression: An evolutionary view.* 1994. San Diego, CA: Academic.

Fridlund, A. J. (1997). The new ethology of human facial expressions. *The psychology of facial expression*, *103*.

Frijda, N. H. (1986). *The emotions.* New York: Cambridge University Press.

Frijda, N. H. (2007). *The laws of emotion.* Mahwah, NJ: Lawrence Erlbaum.

Frijda, N. H., Kuipers, P., & Ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of personality and social psychology*, *57*(2), 212.

Gardner, S. (1993). Irrationality and the philosophy of psychoanalysis. Cambridge, England: Cambridge University Press.

Goldie, P. (2000). The emotions. Oxford, England: Oxford University Press.

Green, R. D., MacDorman, K. F., Ho, C.-C., & Vasudevan, S. K. (2008). Sensitivity to the proportions of faces that vary in human likeness. *Computers in Human Behavior*, 24(5), 2456–2474.

Graham, G. (2010) Behaviorism In *The Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/archives/fall2010/entries/behaviorism/>.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the Marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, 17,54-61.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS one*,*7*(9), e45457.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, 99(3), 561-565.

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3), 260-280.

Jeannerod, M. (2005). How do we decipher others' minds In Fellous J.M. (Ed.) & Arbib, M. A. (Ed.), *Who needs emotions?: The brain meets the robot* (147-169). Oxford: Oxford University Press.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310, 116-119.

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition: An International Journal*, 15(4), 673-692.

Keltner, D., Ekman, P., Gonzaga, G. C., & Beer, J. (2003). Facial expression of emotion In Richard J. (Ed), Scherer, Klaus R. (Ed), Goldsmith, H. Hill (Ed), *Handbook of affective sciences. Series in affective science* (pp. 415-432). New York, NY, US: Oxford University Press.

Kuhlmeier, V. A., Bloom, P., & Wynn, K. (2004). Do 5-month-old infants see humans as material objects?. *Cognition*, *94*(1), 95-103.

Lacewing, M. (2007). Do unconscious emotions involve unconscious feelings?. *Philosophical Psychology*, 20(1), 81-104.

Lambie, J. A., & Marcel, A. J. (2002). Consciousness and the varieties of emotion experience: a theoretical framework. *Psychological review*, *109*(2), 219.

Lange, C. G. (1885). The mechanism of the emotions. *The Classical Psychologists. Boston: Houghton Mifflin*, *1912*.

Lazarus, R. S. (1991) *Emotion and Adaptation*. New York: Oxford University Press.

LeDoux, J. (2002). Synaptic Self: How brains become who we are. *Viking Penguin, New York*.

Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, *78*(2), 341-348.

Öhman, A., Flykt, A., & Lundqvist, D. (2000). Unconscious emotion: Evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. *Cognitive neuroscience of emotion*, *296*.

MacDorman, K. F.& Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive science research. *Interaction Studies*, 7(3), 297-337.

Meltzoff, A., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. Early Development and Parenting, 6, 179–192.

Messinger, D. S., Fogel, A., & Dickson, K. L. (2001). All smiles are positive, but some smiles are more positive than others. *Developmental Psychology*,*37*(5), 642.

Mori, M. (1970/2012). The uncanny valley (K. F. MacDorman & N. Kageki, Trans.). *IEEE Robotics & Automation Magazine*, 19(2), 98–100.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of personality and social psychology*, *64*(5), 723.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Novaco, R. W. (1986). Anger as a clinical and social problem. *Advances in the study of aggression*, *2*, 1-67.

Novaco, R. W., & Taylor, J. L. (2000). Anger. *Encyclopedia of psychology*, *1*, 170-174.

Nussbaum, M. (2001). Upheavals of thought. Cambridge, England: Cambridge University Press.

Miyake, K., & Yamazaki, K. (1995). Self-conscious emotions, child rearing, and child psychopathology in Japanese culture. *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride*, 488-504.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*,*5*(2), 119-124.

Novaco, R. W. (2000). Anger. *Encyclopedia of Psychology*, Oxford University Press.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature neuroscience*, *10*(9), 1095-1102.

Panksepp, J. (1993). Neurochemical control of moods and emotions: Amino acids to neuropeptides In Lewis, Michael (Ed); Haviland, Jeannette M. (Ed), *Handbook of emotions* (pp. 87-107). New York, NY, US: Guilford Press.

Panksepp, J. (2007). Neurologizing the Psychology of Affects. *Perspectives on Psychoogical Science*, 2: 281–296.

Parkinson, B. (2013). Contextualizing Facial Activity. *Emotion Review*, *5*(1), 97-103.

Picard, R. W. (2003). What does it mean for a computer to “have” emotions. *Emotions in humans and artifacts*, 87-102.

Pronin, E. (2009). The introspection illusion. *Advances in experimental social psychology*, *41*, 1-67.

Potegal, M., & Stemmler, G. (2010). Constructing a neurology of anger. In *International Handbook of Anger* (pp. 39-59). New York: Springer.

Pribram, K. H. (1960). A review of theory in physiological psychology. *Annual Review of Psychology*, 11, 1–40.

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A. S., McNamara, J. O., & Williams, S. M. (2001). Emotions: Physiological Changes Associated with Emotion. Neuroscience, 2nd Edition. Sinauer Associates Inc.

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology*, *67*(3), 525.

Rhodes, G., & Zebrowitz, L. A. (2002). *Facial attractiveness: Evolutionary, cognitive, and social perspectives* (Vol. 1). Ablex Publishing Corporation.

Roberts, R. (2003). Emotions: An essay in moral psychology. Cambridge, England: Cambridge University Press.

Rolls, E. T. (1999). *The brain and emotion* (Vol. 4, p. 1619). Oxford: Oxford University Press.

Rolls, E. T. (2005). *Emotion explained*. Oxford University Press.

Roseman, I. J., & Smith, C. A. (2001). Appraisal theory: Overview, assumptions, varieties, controversies.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145.

Russell, J. A. (2009). Emotion, core affect, and psychological construction. *Cognition and Emotion*, *23*(7), 1259-1283.

Russell, J. A., Bachorowski, J. A., & Fernández-Dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual review of psychology*, *54*(1), 329-349.

Saygin, A.P. (2012). The Thing That Should Not Be: Predictive Coding and the Uncanny Valley in Perceiving Human and Humanoid Robot Actions. *Social Cognitive Affective Neuroscience* 7: 413–22.

Scherer, K.R. (1987). Toward a Dynamic Theory of Emotion: The Component Process Model of Affective States In *Geneva Studies in Emotion and Communication* (pp. 1–98); available at: http://www.unige.ch/fapse/emotion/genstudies/genstudies.html

Scherer, K. R. (2000). Psychological models of emotion. *The neuropsychology of emotion*, *137*, 162.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, *92*, 120.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40*(1), 227-256.

Scherer, K. R. (2004, April). Feelings integrate the central representation of appraisal-driven response organization in emotion. In *Feelings and emotions: The Amsterdam symposium* (pp. 136-157). Cambridge,, UK: Cambridge University Press.

Scherer, K.R. (2005). What are emotions? And how can they be measured? *Social Science Information* 44, 693-727.

Scherer, K.R. (2010a) Emotion and emotional competence: conceptual and theoretical issues for modelling agents. In Scherer, K. R., Bänziger, T., & Roesch, E. (Eds.), *A Blueprint for Affective Computing: A sourcebook and manual*. Oxford University Press.

Scherer, K. R. (2010b). The component process model: Architecture for a comprehensive computational model of emergent emotion. *Blueprint for affective computing: A sourcebook*, 47-70.

Schneider, F., Gur, R. C., Gur, R. E., & Muenz, L. R. (1994). Standardized mood induction with happy and sad facial expressions. *Psychiatry research*,*51*(1), 19-31.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3 (3): 417–457.

Searle, J. (1990). Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association* 64 (November): 21–37.

Searle, J. (2002). *Consciousness and Language.* Cambridge University Press.

Shaffer, D. D. R., & Kipp, K. (2007). *Developmental psychology: Childhood and adolescence*. Cengage Learning. Chicago

Shannon, C. E. (1950). XXII. Programming a computer for playing chess. *Philosophical magazine*, *41*(314), 256-275.

Sherman, S. M. (2006). Thalamus. *Scholarpedia*, *1*(9), 1583.

Siegler, R. (2006). *How children develop: Exploring child develop student media tool kit & scientific American reader to accompany how children develop*. New York: Worth Publishers.

Smith, C. A., & Kirby, L. D. (2009). Putting appraisal in context: Toward a relational model of appraisal and emotion. *Cognition and Emotion*, *23*(7), 1352-1372.

Smith, C., & Scott, H. (1997). A componential approach to the meaning of facial expressions. In J. Russell & J. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 229–254). Cambridge: Cambridge University Press.

Calhoun, C. & Solomon, R. C. (eds.) (1984). *What is an Emotion?: Classic Readings in Philosophical Psychology*. Oxford University Press.

Solomon, R.C. (2004). Emotions, thoughts, and feelings: Emotions as engagements with the world. *Thinking About Feeling: Contemporary Philosophers on Emotions*. Oxford University Press. 1-18.

Stoljar, D. (2009). Physicalism In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from <http://plato.stanford.edu/archives/fall2009/entries/physicalism/>.

Stueber, K. R. (2006). *Rediscovering empathy: Agency, folk psychology, and the human sciences*. Cambridge, MA: MIT Press.

Thomaz, A. L., Berlin, M., & Breazeal, C. (2005). An embodied computational model of social referencing. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on* (pp. 591-598). IEEE.

Tomkins, S. S. (1962). Affect, imagery, consciousness: Vol. I. The positive affects.

Turkle, S. (2012). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Van Gulick, R. (2011)., Consciousness In *The Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/archives/sum2011/entries/consciousness/>.

Velik, R. (2010). Why machines cannot feel. *Minds and Machines*, *20*(1), 1-18.

Videbeck, S. L. (2006). *Psychiatric mental health nursing (3rd ed.)*. Lippincott Williams & Wilkins.

Wallbott, H. G. (1998). Bodily expression of emotion. *European journal of social psychology*, *28*(6), 879-896.

Warwick, K., Xydas, D., Nasuto, S. J., Becerra, V. M., Hammond, M. W., Downes, J., & Whalley, B. J. (2010). Controlling a mobile robot with a biological brain. *Defence Science Journal*, 60(1), 5-14.

Wild, B., Erb, M., & Bartels, M. (2001). Are emotions contagious? Evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences. *Psychiatry research*, *102*(2), 109-124.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Psychology*, *55*.

Winkielman, P., & Zajonc & Norbert Schwarz, R. B. (1997). Subliminal affective priming resists attributional interventions. *Cognition & Emotion*, *11*(4), 433-465.

Winkielman, P., Schwarz, N., Reber, R., & Fazendeiro, T. A. (2000). Affective and Cognitive Consequences of Visual Fluency: When Seeing is Easy on the Mind. *Visual Persuasion*.

Wollheim, R. (1984). The thread of life. Cambridge, England: Cambridge University Press.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American psychologist*, *35*(2), 151.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, 39(2), 117-123.

Zajonc, R. B. (1998). Emotions In Gilbert, Daniel T. (Ed); Fiske, Susan T. (Ed); Lindzey, Gardner (Ed), *The handbook of social psychology, Vols. 1 and 2 (4th ed.),* (pp. 591-632). New York, NY, US: McGraw-Hill.

Zajonc, R. B. (2000). Feeling and thinking: Closing the debate over the independence of affect In Forgas, Joseph P. (Ed), (2000). *Feeling and thinking: The role of affect in social cognition. Studies in emotion and social interaction, second series.*, (pp. 31-58). New York, NY, US: Cambridge University Press