Emotion, Artificial Intelligence, and Ethics

Kevin LaGrandeur

Professor, New York Institute of Technology and Fellow, Institute for Ethics and Emerging Technology

klagrand@nyit.edu

Abstract. The growing body of work in the new field of "affective robotics" involves both theoretical and practical ways to instill—or at least imitate—human emotion in Artificial Intelligence (AI), and also to induce emotions *toward* AI in humans. The aim of this is to guarantee that as AI becomes smarter and more powerful, it will remain tractable and attractive to us. Inducing emotions is important to this effort to create safer and more attractive AI because it is hoped that instantiation of emotions will eventually lead to robots that have moral and ethical codes, making them safer; and also that humans and AI will be able to develop mutual emotional attachments, facilitating the use of robots as human companions and helpers. This paper discusses some of the more significant of these recent efforts and addresses some important ethical questions that arise relative to these endeavors.

Keywords: Artificial Intelligence, Affective Robotics, Ethics, Artificial Emotions, Empathic AI, Artificial Conscience

1 Introduction

Many current ideas about creating emotions in Artificial Intelligence (AI) are highly speculative. They are premised upon a future in which we have sentient AI ("strong" AI), and that is a future that could be quite a long way off—or that may never happen. These ideas include two parallel but separate camps of thinkers: those who discuss "friendly AI" [1-2] and those who contemplate what are variously called "moral machines," "robot ethics," or "Artificial Moral Agents (AMA)" [3-5]. As opposed to these foci, this essay will focus on more recent and actual developments regarding the creation of various emotional states in AI, the social motives for doing so, and the ethical dimensions of those efforts and motives.

As a start to this endeavor, we can examine the relatively new field of "affective robotics," the recent attempts to induce an affective state (emotion) in various types of Artificial Intelligence (AI). These efforts stretch back a little more than a decade, and

include numerous ways of simulating emotions (and more recently, seeking ways to actually instill emotions) in AI, and the different motives for doing so. One article that surveys early work on social robots, and which thus cites numerous other early studies, notes, in summary, that there are three main motives for creating artificial emotions:

> [to] facilitate believable human-robot interaction…[to] provide feedback to the user, such as indicating the robot's internal state, goals, and (to some extent) intentions…[and to] act as a control mechanism, driving behavior and reflecting how the robot is affected by, and adapts to, different factors over time [6].

What has changed in the eleven years since this survey was written is that the stakes have risen concerning the use of AI because it has become so widespread and has migrated into sensitive areas, such as the military, domestic companionship, and consumer health care. Accordingly, we will examine AI, emotion, and Human-AI interaction within these particular contexts. There are several motives for trying to create emotions in Artificial Intelligence within these contexts. One is safety, and the other is the attractiveness of AI to humans. If AI continues to become more intelligent, and eventually is able to act autonomously, attractiveness and safety will be paramount, because the biggest sector of robot manufacturing is that for "service robots," which are made primarily for consumers and the military. These kinds of robots—as opposed to "industrial robots," used for manufacturing—include everything from automated milking machines to military drones (which at the moment are the two most numerous types of service robots); this category also includes domestic and personal robots, such as robotic vacuums and "companion" robots of the type used, for instance, as health care aids for the elderly and for autism therapy.

If we look at the statistics given by The Institute of Electrical and Electronics Engineers (IEEE) concerning "World Robot Population," it is clear just how many more service robots exist compared to industrial ones, and how rapidly their numbers are increasing. According to these statistics, by 2011, the total number of robots in use worldwide was 18.2 million: only 1.2 million of those were industrial, and the rest—17 million—were service robots [7]. Another way of looking at this is via the statistics of the International Federation of Robotics (IFR), the major trade group for robot manufacturers. Their data show 2.5 million personal and domestic robots sold in 2011 alone [8] (p. 15); that is more than all of the industrial robots ever sold, from the 1960's to 2011, which amounts to about 2.3 million [8] (p. 10).

2 Programming Emotion in an Effort to Make AI Safe, Friendly, and Attractive

2.1 Military Robots and Emotion: Seeking a Path to Safety

With such rapidly increasing numbers and, obviously, increasing use of machines that can be very powerful, safety is an important issue. And because of their numbers and

social impact, the two types of AI that are receiving the most funding and attention for developing affects or emotions are the personal robots I mention above and military robots. Given that safety is the key factor that unites both of these types of AI, we will focus first on that factor, and first on military AI, because that is where this issue of safety is the supreme consideration. Indeed, if AI continues to become more intelligent and, especially, more autonomous, safety will become an ever more pressing issue in military robotics—which is presently the largest sector of the robot industry. Making computerized, autonomous weapons is clearly fraught with concerns, such as distinguishing between civilians and soldiers, and between friendly soldiers and the enemy, among other things. Furthermore, it is clear that making military robots more autonomous is exactly what the military aims to do. A relatively recent Call for Proposals by the United States Army makes this clear:

> Armed UMS [Unmanned Systems] are beginning to be fielded in the current battlespace, and will be extremely common in the Future Force Battlespace… This will lead directly to the need for the systems to be able to operate autonomously for extended periods, and also to be able to collaboratively engage hostile targets within specified rules of engagement… with final decision on target engagement being left to the human operator…. *Fully autonomous engagement without human intervention should also be considered* (italics added for emphasis) [9].

Ronald Arkin cites this passage in his work, and then points out that there were, as of 2009, already a number of semi-autonomous intelligent weapons in use by the United States Armed Forces and by other forces. One prime example, which is actually fairly old, is the Cruise Missile, which once launched does most of the work of acquiring and destroying its target. Another, more chilling example is a South Korean intelligent weapons platform that can "detect and identify targets in daylight within a 4km radius, or at night using infrared sensors within a range of 2km, providing for either an autonomous lethal or non-lethal response. Although a designer of the system states that 'the ultimate decision about shooting should be made by a human, not the robot,' the system does have an automatic mode in which it is capable of making the decision on its own" [10] (pp. 4-6).

The reason Arkin cites the Army's Call for Proposals and then the examples mentioned above is in order to give a rationale for his own project, which is an attempt to create the software architecture for an artificial conscience that would serve as an ethical governor of autonomous smart weapons. The need for this is, as noted above, very clear. There have already been some documented problems with smart weapons being able to distinguish proper "targets," such as the infamous malfunction of a smart antiaircraft gun in South Africa, when it turned and began killing the soldiers operating it [11]. The main basis for Arkin's artificial conscience would be a reason-based decision tree, but he admits that this plan is not completely sufficient. He points out that the use of emotions may be needed as part of the ethical apparatus. In addition to the rational decision tree based primarily on Kantian deontology (that is, an ethical system focused on sense of duty) combined with the Army's Codified Laws of War, he concedes that there should also probably be some sort of ethical adaptor based on "artificial affective function (e.g., guilt, remorse, grief)" that would motivate

a weaponized AI to review and correct any mistakes it had made in using lethal force [10] (pp. 20-21). This would be done especially focusing on guilt. The coding that would induce robotic "guilt" would be the robot's own monitoring of certain measureable parameters, such as "noncombatant casualties and damage to civilian property, among others," or "criticism" from human monitors [10] (p. 74). Formally, this affective function would be expressed like this:

IF Vguilt > Maxguilt THEN Pl-ethical = ø

"where Vguilt represents the current scalar value of the affective state of Guilt, and Maxguilt is a threshold constant…." If the threshold constant is exceeded, then ethics have been breached, and the weapon is automatically disabled [10] (p.74). The biggest ethical problem with this idea is that such affective functioning only occurs after some kind of heinous humanitarian violation has occurred. The other problem is that these robot "emotions" are only vague simulations—or not even that, but just "diagnostic troubleshooting," as is done now with malfunctioning computers, but under a different name,. Another problem is that the emotions Arkin wants to use in military robots are still based on cold calculation of assessment criteria, not on empathy, sympathy, or, as Arkin himself admits, compassion [10] (p. 75). Thus, calling this development "affective" computing or considering it emotion-based is inaccurate.

We should note that Arkin's stated goal is not to help produce better weapons but to prevent what he sees as the inevitable weaponization of AI by the military from being an unharnessed free-for-all, with huge inadvertent slaughters of innocent noncombatants. This goal of preventing unharnessed slaughter by military robots is a noble one, but, as can be seen above, the coding is just not complex enough, nor the AI advanced enough, to instill the necessary emotions and behavior desired. This is a practical engineering problem that is widespread right now, though small steps toward creating at least simple simulations of some emotions in limited contexts are appearing. The essay by Alidoust and Rouhani in this present volume is an example of that. They present a model for simulation of four emotions (anger, happiness, nervousness, and relief) which, though too simple to imitate human behavior (these behaviors have very narrow determinants in the modeling agents), is a step toward investigating more complex behavior containing more variables and complexities.

However, as Hamid Ekbia argues in his essay, also in this volume, attempts to instantiate emotions in AI may always be doomed to simplistic imitations because of an error in basic assumptions about how emotions work in humans. That is, as he maintains, the approach of AI researchers has to this point been based on what he calls a monadic rather than a dyadic model of emotions. The former model is based on the idea that emotion is internally generated, whereas the latter model, which he argues is more accurate, is based on the idea that emotions are dynamic, relational, and intersubjective—they are built on changing relationships with the external world and its inhabitants, and so they cannot simply be internally generated in robots by a program. I agree. Models like Arkin's and the model referred to by Alidoust and Rouhani (the CCC model), are too focused on autonomous action, as opposed to dynamic interaction, to be a good source of complex, human-like emotion. Beyond these practical

problems, there are other, more abstract philosophical considerations, which we will examine later in the final section of this paper.

2.2 Companion Robots and Emotion: Not Just Safe, but also Attractive AI

2.2.1 Rossler's benevolent AI as a combined attempt at safe military and personal robots

Arkin's ideas and plans are meant for relatively near-term deployment, but they also assume that, in addition to increased autonomy, robots and other military AI will continue to become more intelligent than they are now. Other theories for developing emotional AI in order to protect their human creators assume much more intelligence, including sentience—a concept known as "strong AI" and, as I mentioned in my introduction, most computer scientists consider this type of AI a long way off, if it is possible at all. But there is at least one other theorist who, like Arkin, is trying to make AI friendlier in the near term. That is the German physicist and complexity theorist Otto Rossler. His ideas are laid out in a number of articles, but the most important one is his 2004 article, "Nonlinear Dynamics, Artificial Cognition and Galactic Export" [12]. He claims that, by way of his own mathematical models regarding what he calls "spatial Darwinism," combined with the type of social bonding (called "imprinting") observed by Conrad Lorenz in his famous twentieth-century experiments with geese, a form of benevolent bonding could be programmed into a machine.

Rossler's theory is complex, but this is the essence of it: first, there is Rossler's concept of spatial Darwinism, which he invented to describe how living things survive, not as a species over time (which is Darwin's theory), but as individuals in one lifetime. He maintains that in order to do this, any living animal needs to adapt constantly by moving an appropriate distance through space at the appropriate time, in order to find necessities like food or a mate, and that the valences for this also include important individuals, such as parent figures. Given this definition, Rossler sees benevolence working is as a subset of the concept of "bonding" outlined by Konrad Lorenz and others as the catalyst for benevolence between animal brains. Because bonding works as an adaptive survival trait, it is, he claims, "programmed" into the neural networks of animals. Therefore, for the same reasons, and by way of the mathematical models regarding "spatial Darwinism," bonding could be programmed into a machine.

The way this would work is that algorithms would command a robot's "autonomous path optimization," which Rossler sees as analogous to human emotion in the way that it works, to satisfy its needs,. In other words, Rossler sees emotion as a function of primordial drives, and as necessary adaptations for satisfying those drives. Programming a machine to stay in close proximity to a human is thus relatively straightforward, if bonding is related to basic drives and if it demands that a particular human be seen as integral to its survival and valuable in its own right—what Lorenz called "the animal with home-valence," or more simply, a mother figure [12] (p. 59).

So the autonomous path optimization algorithms would be set to identify a particular human as the "animal with home valence," the equivalent of the mother, for most animals. As a result, the machine would become automatically socially bonded to that human upon first viewing him or her. Then, whenever the human shares things with the robot, as he or she would with a human child, the bonded robot, like a child, would learn to share in return, triggering a learning experience that would initiate an evolving, recursive loop of benevolence between it and humans. A practical problem with this theory is that it depends on the human feeling attachment to the machine, which apparently would be instigated by the robot's following the human around loyally, like a baby goose. Rossler assumes this would please the human, not annoy him. Moreover, a big philosophical problem here is that this theoretical architecture collapses the difference between emotional bonding and simple proximity. And can emotion really be equated to simple "path optimization" for one's survival needs, as Rossler posits, and therefore replicated by a simple algorithm? Although Rossler's model depends more than Arkin's does on "dyadic" relationships to form emotions, as Ekbia and I agree would need to happen, this relational affective model is overshadowed by the fact that the main theory behind Rossler's concept is still "monadic," for the most part (I use Ekbia's apt terms here). It is based on basic drives as the sole reason for emotion, which is too reductive. As Ekbia notes,

> According to [psychologist Sylvan] Tomkins, our behaviors are largely regulated by affects, which are sustained and general in character, as opposed to drives, which are spatially and temporally specific and hence weak in motivating behavior. Affects, as such, take priority over drives. The hunger drive, foundational to behaviorism and also to Freud's theory of sexuality, for instance, is not powerful by itself. It becomes urgent (and so able to compel behavior) when it is amplifed by, say, distress or enjoyment. It can similarly be attenuated or blocked by disgust or fear. **[from p. 9 of Ekbia's manuscript for this volume]**

In short, as with Arkin's idea, Rossler's is noble in concept because it attempts to keep humans safe and happy—and it implicitly keeps an intelligent and perhaps even sentient robot or AI happy, as well, but it has dubious underpinnings. It ignores more subtle needs met and produced by emotions. How, for instance, does empathy fit in here? Or sympathy, or compassion? These emotions are complex and dependent upon inter-relationships, and they are arguably just as important to keeping humans safe and happy as the sort of harmlessness and loyalty that Rossler has named "benevolence." Given the complexities of true benevolence—or of any other true emotionally-based moral behavior—Rossler's prescription may be one for creating mere "clinginess," as opposed to benevolence, a physical behavior, rather than a true emotional or ethical state.

Benevolence is not merely a behavior, though it is manifested that way, it is a complex ethical stance, a conscious decision, based on a constellation of emotions, experience, and reason, to act for the benefit of another. As such, it entails more than just a simple reward system: a baby (or robot) may instinctively bond to the mother figure (her face and smile), and sharing behavior may be a first lesson in the mutual benefit of cooperative social behavior, but is it any more than that? From that step to

benevolence also involves things like altruism, empathy and sympathy, and feelings of responsibility. Some of these are mysteriously complex, like altruism—which research indicates may have not only a genetic component, because it is an adaptive trait for preserving the species, but also a strong learned one.

Likewise empathy seems to be an inborn potentiality that needs experiential help and human instruction to develop. It is a mixture of brain maturation (a physical development of the human organism) and experience. One has to experience pain, for example, in order to understand the pain of others; and not only that, but one has to experience that pain in different contexts to fully understand others' pain. Instruction also plays an important part—parents saying to the child, "How would you like it if someone did that to you?" This aspect of gaining experience via necessary pain also poses a moral conundrum concerning the implication that we might then need to make a sentient Artificial Intelligence that could experience pain: is that moral?

2.2.2 Empathy and Attractive Personal Robots

When contemplating domestic robots, such as robotic helpers and companions, there is more to consider than just safety. Rossler's ideas for benevolent robots hint at this. Personal robots—which can be divided into the categories of domestic and companion robots—need to be attractive to consumers, as well as safe for them to use. As discussed in the introduction to this paper, a key to attractiveness—to "believable human-robot interaction"—is that robots need to exhibit emotion in order to cause humans to develop a real bond with them. As we have seen, affective robotics is in a nascent stage, but researchers have found that at least one category of emotion—empathy—is providing a foothold to creating real bonds between humans and AI. This is actually a two-way process. Robots need to exhibit empathy, and they also need to inspire empathy for themselves in humans; in other words, robots need to enable humans to imagine themselves as the robot—which means humanizing the robot in their minds.

Unsurprisingly, some of this has to do with constructing anthropomorphic facial expressions, speech, and gestures, as with Cynthia Breazeal's experiments at the MIT Media Lab in the early 2000's [13-16]. The animatronic robots created at the Media Lab, such as Kismet, Leonardo, and Huggable, which can still be viewed at the Media Lab's website, were built with special emphasis on facial expression, gestures, and reactivity of both to human interaction [17]. Research has indicated that when anthropomorphic robots mirror the facial expressions and body movements of the human with whom they are interacting, it encourages the human to develop empathy with them [18-20]. But, perhaps more surprisingly, perception of empathy—and human empathic responses to robots—can also have to do with the robot's actions, or the functions it performs.

Some recent examples will help illustrate these somewhat different effects, and also what can now be done in this area of affective computing, and what yet remains. The first example demonstrates a practical application for the descendants of MIT Media Lab's experimental creatures, with their anthropomorphic expressiveness. And

the examples after that one, the instances cited by Matthias Scheutz, exhibit empathy induced by human response to robot functionality. In 2011, a group of researchers funded by the European Union's Platform Seven Agency used empathic robots as teaching tools for elementary school students. As they said, "The goal of LIREC [Living with Robots and Interactive Companions project] was not to build robot companions that replace human contact, but rather to design companions that fulfill their tasks and interact with people in a socially and emotionally acceptable manner" [21] (p. 1). In one of their experiments, reported in a recent article, they used an empathic robot called iCat to teach students to play chess [22]. This robot, made by a Dutch company, and which one can see in the article referenced above [21], looks like a small, plastic cat. It is yellow, is in a sitting position, has tactile sensors in its head and front paws so that it can tell when it is being touched and can react to that. It also has auditory sensors embedded in part of its anatomy, and a tiny webcam mounted in its nose. Most importantly, it has a mobile set of facial characteristics: its mouth, eyes, and eyebrows all move in numerous ways so that, like the MIT creations, it can exhibit facial expressions. Its programming allows it to react to the movements and statements of its human partner. The ability to read human facial expressions is provided by a special software program that also enables it to operate a set of six "model" emotional faces in response to human interaction. This facial expression and recognition software, interestingly, was developed as a (successful) experimental therapy to teach autistic children to better read non-verbal cues [23]. When the students learning to play chess had trouble, the robot would use one of four empathic responses: encouraging comments, offering help, making a bad move intentionally, or scaffolding (which they defined as "providing feedback on the user's last move and, if the move is not good, let the user play again") [22] (p. 3).

Such experiments as this show that some forward progress is being made in practical applications of empathic robotics, but these successes should not be overestimated. Concerning the more complex artificial emotions of the type Arkin and Rossler want to achieve, AI is not powerful enough yet to support this intricate function of sentience. Furthermore, even in the applications discussed above, robots do not really feel empathy. As of now, they just simulate the physical markers of it. But efforts to create the appearance of empathy in robots, the physical markers, have been pretty successful. Consequently, inducing empathy *in humans* toward robots has indeed met with success. Studies show that humans develop real attachment and empathy toward robots. One of the earliest experiments to show this was done by Freedom Baird at MIT's Media Lab in 1999 [24]. Baird was taking care of two gerbils and a simple social robot called a "Furby" to see how the two compared. She noticed that neither the gerbils nor the Furby liked to be held upside down: the gerbils started struggling after about eight seconds of being held that way, and the Furby was programmed to say, over and over and in a pathetic voice, that it was "scared" when it was held this way. Both exhibits of discomfort—from the gerbil and from the Furby—bothered her. So she gave the same experience to a group of children, and she found that the children reacted empathically to both the gerbils and the Furby, as she did. Children, on average, would turn the gerbil rightside up again after eight seconds, and within a minute would also feel compelled to relieve the "suffering" of the Furby robot. Now,

this shows two interesting things: first, that even though people knew that this Furby was just a robot, they felt compelled to respond to its (artificial) emotions; and second, they responded to it more slowly than they did to a genuine animal. These same results have since been replicated in other experiments: humans respond empathically to robots, as such, but not as readily as to humans or animals [25]. But the fact this empathic response is a one-way phenomenon—humans already respond empathically to robots' simulated emotions, and also, as we shall see below, to their actions—is a troubling development, ethically.

This phenomenon of unilateral human empathy toward and attachment to robots is discussed by Matthias Scheutz in a recent book chapter related to robot ethics [26]. He gives a lot of examples of this phenomenon gleaned from various studies, and many of them, as noted above, have to do more with functionality rather than anthropomorphic appearances; one that is somewhat surprising to me is the fact that many people form emotional attachments with their robotic vacuum cleaners, called Roombas. These are simple, disk-shaped devices that merely clean one's floors—one programs them for the time of day they should run, and then they turn themselves on at the designated time and run in a grid-like pattern, bumping into things until they've covered the whole room; then they dock themselves to recharge. Studies cited by Scheutz show that many people personify these simple robots, and some even form a such sense of gratitude toward them that they actually clean the floor themselves in order to give the Roomba "a day off." Many people also dress them up in costumes that can be bought online that are tailored to fit the robot. Scheutz is very concerned about the possibly dangerous behavior that such attachment could cause. He chiefly worries that such one-way attachments will make humans emotionally vulnerable to manipulation by robots, via their human or corporate makers. For instance, corporations that know their robots are seductive could program them to suggest to their smitten human owners to buy more of the corporation's products, or to take other actions not necessarily to their benefit.

An even more direct danger comes from soldiers' emotional attachment to military service robots, such as the bomb-disposal robots used in Afghanistan and Iraq. Scheutz discusses studies that show soldiers can become very devoted to these robots [26] (pp. 211-212). In these cases, it is not just a matter of wanting to give the robot a "day off," or wanting to dress it up because one is emotionally attached to it: personifying bomb-disposal robots makes soldiers reluctant to trade them in for new ones once they have become too damaged to function properly. Obviously, that could cost them their lives.

3 The Road to Future Developments in Artificial Emotion

This current state of uneven reciprocation of emotions between robots and humans raises some problems, as we have seen, in great part because robots cannot feel emotion or empathy now. The state of computing is just not powerful enough to provide strong AI, and it is not likely to be unless experiments with quantum computing or molecular computing succeed. However, although we may have a long way to go

before we can create molecular or quantum-level computing that leads to super-powerful AI, some elements that are part of emotional response in AI are possible now, because of recent, incremental successes replicating cognitive features that contribute to emotions.

For example, Kim and Lipson did an experiment reported in a recent article (2009) on the efficacy of programs that give robots a basic Theory of Mind (ToM) [27]. Essentially, ToM is the ability to understand another's intentions. Humans commonly use ToM to make inferences about others' feelings and states of mind. These investigators created an evolutionary algorithm that allows one robot to infer from another robot's actions what it might do next and how it reasons. Their experiment's main goal was to develop "…controller inference algorithms in robots [that could] help in interaction with non-robotic actors such as humans…" [27] (p. 2072). The experimental set-up provided one robot whose mission was to find a path across a room to a light source. That path varied continuously, based on the position of the light and other factors. Ultimately, the experiment was successful: because of algorithms that could evolve with experience, one robot was able to continually improve its inferences about what another was going to do. This was in a tightly controlled situation, but the long-range implications are obvious: Kim and Lipson's success in creating ToM in a robot is a small step toward enabling robots and other AI to read the internal state and intentions of humans, and thus to bring them one step closer to a mutual emotional interface.

Most remarkably, there are Theodore Berger's successes with long-term memory re-generation (and generation) by using implanted chips to replace damaged parts of the hippocampus in rats and monkeys [28-29]. Berger and his team at University of Southern California have succeeded in recording and transforming into computer code long-term memories that are stored in the hippocampus of these animals. In the case of the rats, they had them perform a memory task. Then, they downloaded and transformed the memory of that task into digital code. Afterwards, they removed the section of the rat's hippocampus that carried these memories and replaced that bit of the brain with a special computer chip, onto which they reloaded the artificially stored memories. They found that these rats' memories could be fully restored using this technique. Even more significant was the fact that Berger's team could also generate or enhance memories that had never existed in the animals—for instance, memory of a task that a rat had never done. They were later able to replicate these same results with monkeys [29]. The implications of this are enormous: if memories can be artificially generated, then it brings up the possibility of generating emotion via chips, too, by using them to replace parts of other brain structures, such as the amygdala, where empathy resides. Obviously, there is also an enormous ethical problem here in giving ourselves the ability to generate false memories, or to enhance long-term memories, which could open up many modes of abuse.

4 Conclusion: Further Ethical Considerations

The ethical concerns of Scheutz's, and those of mine that I've discussed to this point, are specific to particular experiments or projects. What about the ethics and philosophical dimensions of the larger project of generating emotions between humans and Artificial Intelligence? First, although the big problem for the more advanced types of projects like Rossler's and Arkin's is our currently insufficient engineering capabilities, there is also a larger philosophical problem: our perennial disagreement as to what basis we should use to define "proper ethics" when discussing and defining values like "benevolence" or "conscience"—especially given different cultural viewpoints. Kantian deontology (based on pure duty or rule-based ethics), Buddhism (based on selfless compassion), and Utilitarianism (based the greater good of the many) are just a few of the philosophical systems that have been proposed as a basis for AI ethics.

Second, if inducing emotions in AI is important to the effort to create "friendly AI" because it is hoped that AI and humans will develop mutual emotional attachments, then the current experiments are working badly, because so far, the emotional attachment has been a one-way occurrence, as Scheutz reminds us. This is potentially problematic for the reasons noted regarding human vulnerability to emotional manipulation. And third, there are the potential philosophical problems we may create for our treatment of a new species, if we ever manage to create sentient, feeling AI. The problems in this scenario are many, but the chief one that concerns us, regarding emotions, is this: As James Hughes points out [30], and as I mentioned earlier in this paper, in order to feel emotions like guilt, compassion, and empathy, we would have to create suffering beings, because only by suffering do we learn to understand others' pain. But creating a suffering being is of dubious morality. So, would these requirements for instilling compassion in an AI be inhumane?

Research on affective robotics raises some other important philosophical questions relative to it, as well as to human progress in the digital age: Do efforts such as the ones I've outlined risk reducing the complexities of emotive human "movement and the non-verbal spectrum to patterns of imitation and functionality," as some have worried [31]? Clearly they do now. But could these theories and programs do more? Perhaps. Anyone who reads the literature can see that the intent of scientists and others involved in this project is clearly to do more than reduce human emotion to imitations and functionalities, but there is no way yet to do that, and they see imitation as an important initial step.

So, does the increasing juxtaposition of the human with the digital undermine the uniqueness and importance of human kinaesthetic communication processes? Right now, yes, but in the future, the answer to this depends on what sort of perspective one takes on the long-term goals of AI researchers and roboticists. From their perspective, they are trying to replicate those same kinaesthetic communication processes, and in all of their spontaneity, because their ultimate goal is to create robots and AI that are humanoid and—importantly—are emergent: that can, in other words, evolve. If that occurs, then perhaps new, hybrid AI-Human kinaesthetic processes will evolve, as well, and that sort of spontaneous, random change would create its own sort of hybridized kinaesthetic dynamic.

Acknowledgements

This work was supported by a Grant from the Culture Agency of the European Union, Agreement Number 2013 - 1572 / 001 - 001 CU7 MULT7, Metabody Project; as well as by the New York Institute of Technology (NYIT).

References

1. Yudkowsky, E.: Creating friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. The Singularity Institute. http://intelligence.org/files/CFAI.pdf (2001)
2. Muehlhauser, L., Helm L.: Intelligence Explosion and Machine Ethics. In: Eden, A., Søraker, J., Moor, J.H., Steinhart, E. (eds.) Singularity Hypotheses: A Scientific and Philosophical Assessment. Springer, Berlin (2012)
3. Wallach, W., Allen C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2009)
4. Anderson, M., Anderson, S. L.: Machine Ethics. Cambridge University Press, Cambridge (2011)
5. Lin, P., Abney, K., Bekey G. A. (eds.): Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press, Cambridge, MA (2012)
6. Fong, T., Nourbakhsh, I., Dautenhahn K.: A Survey of Socially Interactive Robots: Concepts, Design, and Applications. Technical report, The Robotics Institute, Carnegie Mellon University (2002)
7. Guizzo, E.: 6.5 Million Robots Now Inhabit the Earth. IEEE Spectrum. 15 Oct. 2008. http://spectrum.ieee.org/automaton/robotics/robotics-software/world_robot_population_reaches_6_and_half_million
8. World Robotics--Industrial Robots 2012: Executive Summary. World Robotics: International Federation of Robotics Statistical Department. World Robotics (2012) http://www.worldrobotics.org/index.php?id=downloads
9. U.S. Army SBIR Solicitation 07.2, Topic A07-032: Multi-Agent Based Small Unit Effects Planning and Collaborative Engagement with Unmanned Systems. 57-68 (2007)
10. Arkin, R.: Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture. Technical report, Georgia Institute of Technology (2007) http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf
11. Shachtman, N.: Robot Cannon Kills 9, Wounds 14. Wired Magazine. Oct. 18, 2007 http://www.wired.com/dangerroom/2007/10/robot-cannon-ki/
12. Rossler, O. E.: Nonlinear Dynamics, Artificial Cognition and Galactic Export. American Institute of Physics, Conference Proceedings. Available at Lampsacus.com (2003) http://www.lampsacus.com/documents/roesslergalacticexport.pdf
13. Breazeal, C., Scassellati, B.: Infant-Like Social Interactions Between a Robot and a Human Caregiver. Adaptive Behavior. 8(1), 49-74 (2000)
14. Breazeal, C., Scassellati, B.: Robots that Imitate Humans. Trends in Cognitive Science. 6, 481-487 (2002)

15. Breazeal, C.: Emotive Qualities in Lip Synchronized Robot Speech. Advanced Robotics. 17 (2), 97-113 (2003)
16. Breazeal, C.: Designing Sociable Robots. MIT Press Cambridge, MA. (2002)
17. MIT Media Lab Personal Robots Group, http://robotic.media.mit.edu/projects/projects.html
18. Gazzola, V., Rizzolatti, G., Wicker, B., Keysers, C.: The Anthropomorphic Brain: The Mirror Neuron System Responds to Human and Robotic Actions. NeuroImage. 35, 1674–1684 (2007)
19. Obermana, L. M., McCleeryb, J. P., Ramachandrana, Vilayanur S., Pineda J. A.: EEG Evidence for Mirror Neuron activity During the Observation of Human and Robot Actions: Toward an Analysis of the Human Qualities of Interactive Robots. Neurocomputing. 70, 2194–2203 (2007)
20. Breazeal C., Buchsbaum, D., Gray, J., Gatenby, D., Blumberg B.: Learning From and About Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots. Artificial Life. 11(1-2), 1-32 (2005)
21. Castellano, G., Leite, I., Paiva, A., McOwan, P.W.: Affective Teaching: Learning More Effectively from Empathic Robots. Awareness Magazine. 9 January (2012)
22. Leite, I., Castellanoy, G., Pereira, A., Martinho, C., Paiva A.: Modelling Empathic Behaviour in a Robotic Game Companion for Children: an Ethnographic Study in Real-World Settings. Proceedings of ACM/IEEE International Conference on Human-Robot Interaction, HRI'12. ACM, Boston (2012)
23. Kaliouby, R. E., and Robinson, P.: Mind Reading Machines: Automated Inference of Cognitive Mental States from Video. Proceedings, IEEE International Conference on Systems, Man and Cybernetics. 7, 682-688 (2004) available on IEEE Xplore: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9622
24. Glass, I.: Furrbidden Knowledge. In: Talking to Robots. Radiolab. Podcast, http://www.radiolab.org/2011/may/31/furbidden-knowledge/ (31 May 2011)
25. Humans Feel Empathy for Robots: fMRI Scans Show Similar Brain Function When Robots Are Treated the Same as Humans. Science Daily. (23 April 2013) http://www.sciencedaily.com/releases/2013/04/130423091111.htm
26. Scheutz, M.: The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In: Lin, P., Abney, K., Bekey, G. (eds.) Robot Ethics: The Ethical and Social Implications of Robotics, pp. 205-221. MIT Press, Cambridge, MA (2012)
27. Kim, K.J., Lipson, H.: Towards a 'Theory of Mind' in Simulated Robots. In: Rothlauf, F. (ed.) Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference, pp. 2071-2076. ACM, New York (2009)
28. Berger, T.W., Hampson, R. E., Song, D., Goonawardena, A., Marmarelis, V.Z., Deadwyler, S.A.: A Cortical Neural Prosthesis for Restoring and Enhancing Memory. J. Neural Eng. 8(4), 046017 (2011)
29. Hampson, R. E., Gerhardt, G. A., Marmarelis, V.Z., Song, D., Opris, I., Santos, L., Berger, T.W., Deadwyler, S.A.: Facilitation and Restoration of Cognitive Function in Primate Prefrontal Cortex by a Neuroprosthesis that Utilizes Minicolumn-Specific Neural Firing. J. Neural Eng. 9(5), 056012 (2012)
30. Hughes, J.: Compassionate AI and Selfless Robots: A Buddhist Approach. In: Lin, P., Abney, K., Bekey, G. (eds.) Robot Ethics: The Ethical and Social Implications of Robotics, pp. 69-83. MIT Press, Cambridge, MA (2012)

31. del Val, J.: Metabody: Media Embodiment, Tékhne, and Bridges Of Diversity. Grant Project Description: European Commission - Education, Audiovisual and Culture Executive Agency, Agreement Number 2013 - 1572 / 001 - 001 CU7 MULT7.