MORAL EMOTIONS FOR ROBOTS

RONALD C. ARKIN
Mobile Robot Laboratory, Georgia Institute of Technology
85 5th ST NW, Atlanta, GA 30332 U.S.A.

As robotics moves toward ubiquity in our society, there has been only passing concern for the consequences of this proliferation (Sharkey, 2008). Robotic systems are close to being pervasive, with applications involving human-robot relationships already in place or soon to occur, involving warfare, childcare, eldercare, and personal and potentially intimate relationships. Without sounding alarmist, it is important to understand the nature and consequences of this new technology on human-robot relationships. To ensure societal expectations are met, this requires an interdisciplinary scientific endeavor to model and incorporate ethical behavior into these intelligent artifacts from the onset, not as a post hoc activity. We must not lose sight of the fundamental rights human beings possess as we create a society that is more and more automated. One of the components of such moral behavior, we firmly believe, involves the use of moral emotions.

Haidt (2003) enumerates a set of moral emotions, divided into four major classes: Other- condemning (Contempt, Anger, Disgust); Self-conscious (Shame, Embarrassment, Guilt); Other-Suffering (Compassion); Other-Praising (Gratitude, Elevation). Allen et al (2006) assert that in order for an autonomous agent to be truly ethical, emotions may be required at some level: "While the Stoic view of ethics sees emotions as irrelevant and dangerous to making ethically correct decisions, the more recent literature on emotional intelligence suggests that emotional input is essential to rational behavior". These emotions guide our intuitions in determining ethical judgments, although this is not universally agreed upon (Hauser, 2006). From a neuroscientific perspective, Gazzaniga (2005) states: "Abstract moral reasoning, brain imaging is showing us, uses many brain systems", where he identifies the locus of moral emotions as being located in the brainstem and limbic system.

The relatively young machine ethics community has focused largely to date on developmental ethics, where an agent develops its own sense of right and wrong in situ. In general, these efforts largely ignore the moral emotions as a scientific basis worthy of consideration. Nonetheless, considerable research has been conducted regarding the role of emotions in robotics, including work in our laboratory over the past 20 years (Arkin, 2005; Moshkina et al 2011). Far less explored in robotics is the set of moral secondary emotions, and their role in robot behavior and human-robot interaction. One example is where De Melo et al (2009) have demonstrated that the presence of moral affect in human-robot interaction is both discernible and enhances the interplay between humans and robot-like avatars.

Our own research (Arkin and Ulam, 2009) in the moral affective space research is illustrated by the use of guilt being incorporated into an ethical robotic software architecture designed for lethal military applications. Guilt is "caused by the violation of moral rules and imperatives, particularly if those violations caused harm or suffering to

others" (Haidt, 2003) and is recognized as being capable of producing proactive, constructive change (Tangney et al, 2007). The specific architectural component we have implemented, referred to as the ethical adaptor, incorporates Smits and De Boeck's (2003) mathematical model of guilt, which is used to proactively alter the behavior of the robotic system in a manner that will lead to a reduction in the recurrence of an event which was deemed to be guilt-inducing. In our initial application, this focuses on the deployment of lethal autonomous weapons systems in the battlefield, with respect to unexpectedly high levels of battle damage. Simulation results demonstrate the ethical adaptor in operation.

For non-military applications, we hope to extend this earlier research into a broader class of moral emotions, such as compassion, empathy, sympathy, and remorse, particularly regarding the use of robots in elder or childcare, in the hopes of preserving human dignity as these relationships unfold in the future. There is an important role for artificial emotions in personal robotics as part of meaningful human-robot interaction, and having worked with Sony Corporation on their AIBO and QRIO entertainment robots (Arkin, 2005), and Samsung for their humanoid robots (Moshkina et al, 2011), it is clear that value exists for their use in establishing long-term human-robot relationships.

There are, of course, significant ethical considerations associated with this use of artificial emotions in general, and moral emotions in particular, due in part to their deliberate fostering of attachment by human beings to non-human artifacts. This is believed to promote detachment from reality by the affected user (Sparrow, 2002). While many may view this as a benign, or perhaps even beneficial effect, not unlike entertainment or video games, it can clearly have deleterious effects if left unchecked, hence the need for incorporating models of morality within the robot itself.

Acknowledgements

This research was supported under Contract #W911NF-06-1-0252 from the U.S. Army Research Office. The author would also like to acknowledge Patrick Ulam for his contribution in software development for this project.

References

Allen, C., Wallach, W., and Smit, I., (2006). "Why Machine Ethics?" *IEEE Intelligent Syst*ems, July.

Arkin, R.C., (2005). "Moving Up the Food Chain: Motivation and Emotion in Behavior-based Robots", in *Who Needs Emotions: The Brain Meets the Robot*, Eds. J. Fellous and M. Arbib, Oxford University Press.

Arkin, R.C. and Ulam, P., (2009). "An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions", *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR.

De Melo, C., Zheng, L. and Gratch, J., (2009). "Expression of Moral Emotions in Cooperating Agents". *9th International Conference on Intelligent Virtual Agents,* Amsterdam.

Gazzaniga, M., (2005). *The Ethical Brain*, Dana Press.

Haidt, J. (2003). “The Moral Emotions”, in *Handbook of Affective Sciences*, Oxford Press.

Hauser, M., (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*, ECCO, HarperCollins, N.Y., 2006.

Moshkina, L., Park, S., Arkin, R.C., Lee, J.K., Jung, H., (2011). "TAME: Time-Varying Affective Response for Humanoid Robots", *International Journal of Social Robotics*.

Sharkey, N. (2008). “The Ethical Frontiers of Robotics”, *Science*, (322): 1800-1801.

Smits, D., and De Boeck, P., (2003). “A Componential IRT Model for Guilt”, *Multivariate Behavioral Research*, Vol. 38, No. 2, pp. 161-188.

Sparrow, R., (2012). “The March of the Robot Dogs”, *Ethics and information Technology*, Vol. 4(2).

Tangney, J., Stuewig, J., and Mashek, D., (2007). “Moral Emotions and Moral Behavior”, *Annu. Rev. Psychol.,* Vol.58, pp. 345-372.