THE ETHICS OF ROBOTIC DECEPTION

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

The time of robotic deception is rapidly approaching. While there are some individuals trumpeting about the inherent ethical dangers of the approaching robotics revolution (e.g., Joy, 2000; Sharkey, 2008), little concern, until very recently, has been expressed about the potential for robots to deceive human beings. Our working definition of deception (for which there are many) that frames the rest of this discussion is "deception simply is a false communication that tends to benefit the communicator" (Bond and Robinson, 1988). Research is slowly progressing in this space, with some of the first work developed by Floreano et al (2007) focusing on the evolutionary edge that deceit can provide among an otherwise homogeneous group of robotic agents. This work did not focus on human-robot deceit, however.

As an outgrowth of our research in robot-human trust (Wagner and Arkin, 2008), where robots were concerned as to whether or not to trust a human partner rather than the other way around, we considered the dual of trust: deception. As any good conman knows, trust is a precursor for deception, so the transition to this domain seemed natural. We were able to apply the same models of interdependence theory (Kelley and Thibaut, 1978) and game theory, to create a framework whereby a robot could make decisions regarding both when to deceive (Wagner and Arkin, 2009) and how to deceive (Wagner and Arkin, 2011). This involves the use of partner modeling or a simplistic view (currently) of theory of mind to enable the robot to (1) assess a situation; (2) recognize whether conflict and dependence exist in that situation between deceiver and mark, which is an indicator of the value of deception; (3) probe the partner (mark) to develop an understanding of their potential actions and perceptions; and (4) then choose an action which induces an incorrect outcome assessment in the partner.

While the results we published (Wagner and Arkin, 2011) we believe were modestly stated, e.g., "they do not represent the final word on robots and deception", "the results are a preliminary indication that the techniques and algorithms described in this paper can be fruitfully used to produce deceptive behavior in a robot", "much more psychologically valid evidence will be required to strongly confirm this hypothesis", etc. The response to this research has been quite the contrary, ranging from accolades (being listed as one of the top 50 inventions of 2010 by Time Magazine (Suddath, 2010)) to damnation ("In a stunning display of hubris, the men ... detailed their foolhardy experiment to teach two robots how to play hide-and-seek" (Tiku, 2010), and "Researchers at the Georgia Institute of Technology may have made a terrible, terrible mistake: They've taught robots how to deceive" (Geere, 2010)).

It seems we have touched a nerve. How can it be both ways? It may be *where* deception is used that forms the hot button for this debate. For military applications, it seems clear that deception is widely accepted (which indeed was the intended use of our research as our sponsor is the Office of Naval Research). Sun Tzu is quoted as saying

that "All warfare is based on deception", and Machiavelli in *The Discourses* states that" Although deceit is detestable in all other things, yet in the conduct of war it is laudable and honorable". Indeed there is an entire U.S. Army (1988) Field Manual on the subject.

In our original paper (Wagner and Arkin, 2011), we included a brief section on the ethical implications of this research, and called for a discussion as to whether roboticists should indeed engage in this endeavor. In some ways, outside the military domain, the dangers are potentially real. And of course, how does one ensure that it is only used in that context? Is there an inherent deontological right, whereby humans should not be lied to or deceived by robots? Kantian theory clearly indicates that lying is fundamentally wrong, as is taught in most introductory ethics classes. But from a utilitarian perspective there may be times where deception has societal value, even apart from the military (or football), perhaps in calming down a panicking individual in a search and rescue operation or in the management of patients with dementia, with the goal of enhancing that individual's survival. In this case, even from a deontological perspective, the intention is good, let alone from a utilitarian consequentialist measure. But does that warrant allowing a robot to possess such a capacity?

The point of this paper is not to argue that robotic deception is ethically justifiable or not, but rather to help generate discussion on the subject, and consider its ramifications. As of now there are absolutely no guidelines for researchers in this space, and it indeed may be the case that some should be created or imposed, either from within the robotics community or from external forces. But the time is coming, if left unchecked, you may not be able to believe or trust your own intelligent devices. Is that what we want?

Acknowledgements

This research was supported by the Office of Naval Research under MURI Grant # N00014-08-1-0696. The author would also like to acknowledge Dr. Alan Wagner for his contribution to this project.

References

Bond, C. F., & Robinson, M., (1988). "The evolution of deception", *Journal of Nonverbal Behavior*, 12(4), 295- 307.

Floreano, D., Mitri, S., Magnenat, S., & Keller, L., (2007). "Evolutionary Conditions for the Emergence of Communication in Robots". *Current Biology*, 17(6), 514-519.

Geere, D., (2010). Wired Science, http://www.wired.com/wiredscience/2010/09/robots-taught-how-to-deceive/

Joy, B. (2000). "Why the Future doesn't need us". *Wired*, April 2000.

Kelley, H. H., & Thibaut, J. W., (1978). Interpersonal Relations: A Theory of Interdependence. New York, NY: John Wiley & Sons.

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

Suddath, C., (2010). "The Deceitful Robot", Time Magazine, Nov. 11, 2010, http://www.time.com/time/specials/packages/article/0,28804,2029497_2030615,00.html

Tiku, N., (2010). *New York Magazine*, 9/13/2010, http://nymag.com/daily/intel/2010/09/someone_taught_robots_how_to_l.html

U.S. Army (1988.). Field Manual 90-2, Battlefield Deception, http://www.enlisted.info/field-manuals/fm-90-2- battlefield-deception.shtml

Wagner, A. and Arkin, R.C., (2008). "Analyzing Social Situations for Human-Robot Interaction", *Interaction Studies*, Vol. 9, No. 2, pp. 277-300.

Wagner, A. and Arkin, R.C., (2009). "Robot Deception: Recognizing when a Robot Should Deceive", *Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR.

Wagner, A.R., and Arkin, R.C., (2011). "Acting Deceptively: Providing Robots with the Capacity for Deception", *International Journal of Social Robotics*, Vol. 3, No. 1, pp. 5-26.