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.

What if you could no longer believe what your robot assistant was telling you was the truth? Are there circumstances under which that would be acceptable? What if it was for your own good? The time of robotic deception is rapidly approaching. We are being bombarded regarding the inherent ethical dangers of the approaching robotics and AI revolution, but far less concern has been expressed about the potential for robots to deceive human beings.

Deception according to the Turing test for AI is a hallmark characteristic of intelligence, and philosophers such as Dennett (1997) have stated "*"another price you pay for higher-order intentionality is the opportunity [for] ... deception"*. Our working definition of deception (for which there are many) is "deception simply is a false communication that tends to benefit the communicator" (Bond and Robinson, 1988). Several robotics researchers have considered the role of deception for both agent survival (Floreano, 2007) and human-robot interaction (Short et. al., 2010), including our group.

We have successfully demonstrated the value of biologically-inspired deception in four separate cases as applied to robotic systems: (1) pursuit-evasion using interdependence theory when hiding from an enemy (Wagner and Arkin 2008); (2) misdirection based on behavioral changes (Shim and Arkin 2012) ; (3) feigning strength when it does not exist (Davis and Arkin, 2012), and (4) deception used for the benefit of the mark (Shim, 2017). The response to our research at times has been quite striking, 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). This spectrum of response is quite striking. Perhaps, it is *where* deception is used that is the hot button for this debate.

For military applications, it seems clear that deception is widely accepted. Sun Tzu in the *Art of War* said that “All warfare is based on deception”, while Machiavelli in *the Discourses* stated to the effect that “Although deceit is detestable in all other things, yet in the conduct of war it is laudable and honorable”. Indeed, the U.S. Army (1988) has a Field Manual on the subject.

The dangers outside of the military are quite real. And of course, after its development, how is it ensured that it is only used in the context it was designed for? Is there an inherent fundamental right, whereby humans should not be lied to or deceived by robots? Kant’s categorical imperative clearly indicates that lying is fundamentally wrong, as is taught in most introductory ethics classes. But from a consequentialist point of view there are times when deception has societal value, even apart from the military (or adversarial sports), 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 rights-based approach, the intention is good, let alone from a utilitarian or consequentialist formulation. But even then, does that warrant allowing a robot to possess such a capacity?

The point here 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. In particular, the IEEE Global Initiative on Ethics of Intelligent and Autonomous Systems is now confronting these questions among many others. 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 also thanks Alan Wagner, Jaeeun Shim-Lee, and Justin Davis for their contributions.

References

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

Davis, J. and Arkin, R.C. (2012). "Mobbing Behavior and Deceit and its role in Bio-inspired Autonomous Robotic Agents", *Proc. 8th International Conference on Swarm Intelligence (ANTS 2012)*, Brussels, BE.

Dennett, D. C. (1997). "When hal kills, who's to blame? computer ethics," in HAL's Legacy: 2001's Computer as Dream and Reality (Stork, D. G., ed.), Cambridge, MA: MIT Press.

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/ , accessed May 8, 2018.

Shim, J., and Arkin, R.C. (2012). "Biologically-Inspired Deceptive Behavior for a Robot", *12th International Conference on Simulation of Adaptive Behavior (SAB2012)*, Odense, DK.

Shim, J. (2017). *The Benefits of Other-oriented Robot Deception in Human-robot Interaction*, Ph.D. Dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology.

Short, E., Hart, J., Vu, M., and Scassellati, B. (2010). "No fair!!: an interaction with a cheating robot," Proc. 5th ACM/IEEE international conference on Human-robot interaction, HRI '10, pp. 219–226.

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.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.