The Darker Sides of Robot Social Awareness

Guy Hoffman
Media Innovation Lab
IDC Herzliya
Herzliya, Israel
hoffman@idc.ac.il

Moran Cerf
Kellogg School of Management
Northwestern University
Evanston, IL, USA
m-cerf@kellogg.northwestern.edu

Minimal social awareness required of autonomous interactive robots can include rudimentary behaviors and traits, such as awareness of spatial relationships, individual identities, and interaction histories [1]. These traits likely exist in a broad range of non-human organisms.

However, when considering social robots designed specifically for human interaction, an expanded notion of social awareness should be considered. This notion includes not just basic interpersonal mechanisms, but also those that are particular to the human experience, including cultural and societal aspects of behavior and awareness. In addition, are there not only minimal requirements, but also upper limits on acceptable social awareness?

Some "higher" human social behaviors are generally viewed as positive and beneficial to society at large. These include self-control and delayed gratification [2], altruism [3], as well as perspective-taking and empathy [4].

That said, behavioral research also indicates that human social behavior is marked by a number of "darker" tendencies and biases. Humans are favorable toward in-group members [5], display racial and gender stereotypes [6], engage in deceit [7]; and are liable to fall into group thinking, peer-pressure, and conformity [8].

Given the extent to which Human-Robot Interaction relies on human behaviorial research, both to design robot behavior and to perceive and analyze human behavior, how should roboticists take these darker sides of human social awareness into account?

Should robots mimic negative patterns of social awareness to better pass in human society? Should they cater to them to be more effective in interacting, understanding, and persuading humans? Or could robots present an opportunity to nudge human behavior towards more positive social behavior?

Note that both positive and negative social biases are not necessarily irrational or non-optimal, and could be beneficial and efficient for individuals and groups [9]. Similarly, a perfectly rational robotic agent could also be acting in a socially biased manner.

Consider, for example, the case of an automatic sliding door. By some measure, this is a very simple robot interacting with humans. It has a single sensor and actuator, and makes a straightforward "decision" of opening the doorway for approaching humans. Now imagine this robot enhanced by a camera and face-recognition software and programmed to prevent the entry of recognized shoplifters. Taking this idea one step further, the store's owner could request the installation of software that prevents the entry of people who are classified by a machine learning and pattern recognition algorithm as having a high likelihood of being shoplifters, or even of just having bad credit.

Has such a discriminating autonomous door acquired some level of negative social awareness?

These questions gives rise to three possible ways for researchers in Human-Robot Interaction to address negative social awareness when designing interactive robots:

The first approach is to develop robots that take into account human negative social behaviors. These robots would be more similar to us, incorporating our negative biases, and more adept to us, taking these biases into account when modeling humans.

The second approach would suggest having robots be agnostic or neutral with respect to human social biases. Those robots will be merely functional and will neither provide nor understand social patterns. Such socially lacking robots will not suffer from biases, but may also be less successful in generating a high quality of interaction with humans.

A third approach would be to design robots that are not merely not susceptible to human negative biases, but purposefully embody positive aspects of human social behavior. These robots, personifying the "better angels" of human nature [10], may both interact successfully with humans and also help tame our own negative social behaviors.

That is, rather than acting as proxies for our own social shortcomings, robots can be thought of as tools to support more positive social awareness based on an agreed set of rules, effectively improving on human social awareness. Instead of being bounded by the same biases that humans have a hard time shaking, such as racism, dishonesty, and conformity, researchers can design robots that specifically support values like equality, honesty, and independent thinking. Through interaction, they might shape human behavior and serve as guides for more desirable behavior.

To summarize, we ask not only about the minimum set of social awareness required to simulate consciousness in an interactive robot, but also about acceptable upper bounds, given that human social awareness often leads to negative biases and behaviors. Designing robots guided by some social responsibility may shape their interaction with humans, and in turn steer us towards acting more positively.

REFERENCES

[1] J. Wiles, "Will Social Robots Need to Be Consciously Aware?" *IEEE CIS Newsletter of the Autonomous Mental Development Technical Committee*, vol. 11, no. 2, pp. 14—-15, 2014.

[2] R. F. Baumeister and J. Juola Exline, "Virtue, Personality, and Social Relations: Self-Control as the Moral Muscle," *Journal of personality*, vol. 67, no. 6, pp. 1165–1194, 1999.

[3] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, no. 6960, pp. 785–791, 2003.

[4] B. Underwood and B. Moore, "Perspective-taking and altruism." *Psychological Bulletin*, vol. 91, no. 1, p. 143, 1982.

[5] B. Mullen, R. Brown, and C. Smith, "Ingroup bias as a function of salience, relevance, and status: An integration," *European Journal of Social Psychology*, vol. 22, no. 2, pp. 103–122, 1992.

[6] S. Fiske, "Stereotyping, prejudice, and discrimination," *The handbook of social psychology*, p. 357, 1998.

[7] F. Gino, S. Ayal, and D. Ariely, "Contagion and differentiation in unethical behavior: the effect of one bad apple on the barrel." *Psychological science*, vol. 20, no. 3, pp. 393–8, Mar. 2009.

[8] S. E. Asch, "Effects of group pressure upon the modification and distortion of judgments," *Groups, leadership, and men. S*, pp. 222–236, 1951.

[9] D. K. Levine, "Modeling altruism and spitefulness in experiments," *Review of economic dynamics*, vol. 1, no. 3, pp. 593–622, 1998.

[10] D. C. Lahti and B. S. Weinstein, "The better angels of our nature: group stability and the evolution of moral tension," *Evolution and Human Behavior*, vol. 26, no. 1, pp. 47–63, 2005.