Autonomy Examined: Research Directions For the Design of Accountable Social Robots

Thomas Arnold, Gordon Briggs, and Matthias Scheutz
{thomasarnold}@alumni.stanford.edu,
{gordon.briggs,matthias.scheutz}@tufts.edu

Human-Robot Interaction Laboratory, Tufts University, Medford, MA USA

Efforts to ban lethal autonomous weapons systems have gained an increasing amount of public attention, both as campaigns and as ethical statements for how autonomous systems should function in human society. Citing laws of war or "humanity" introduced by the prospect of autonomous systems targeting, attacking, and assessing without communication with human command, many prominent figures have suggested that international law must step in to ban such systems before they can begin operating in the field. These advocates include prominent members of the artificial intelligence (AI) community, such as Stuart Russell[1], as well as well-known figures from outside AI. While the potential risks of a reckless and inhumane weapons systems is certainly a threatening cloud in the sky of robotics' future, there is a larger storm front behind it that needs no less study and consideration. Detailing the potential threats and violations of "killer robots" in the military context should not lull us into neglecting the many complex ways robots risk harming human beings in the home, office, school, hospital, and street.

Social robots are rapidly making their way into many contexts of human interaction and needs, and often the conditions and risks of their entry therein have not been adequately foreseen. It was well after Google's self-driving car was being tested that the ethical problems it would face received substantive public discussion in the media – not just how it would act in a *moral dilemma* (e.g., whether it would run over a child in the street or crash into a guardrail instead risking the driver's life), but how its user would be affected no matter what choice it made (would the owner be traumatized by the child's death?). Health-care cases, where a robot might be forced to make real-time decisions about immediate intervention, are even more fraught. And even in cases of domestic companionship, where the robots are engaging, "cute," or marketed as emotionally perceptive, one has to consider the types of dysfunctional, maladaptive, or abusive dynamics that a socially sophisticated system might create with its user [1]. Battlefield destruction may not be at issue, but broad ramifications in public health and societal needs require that domestic, health-care, educational, law enforcement, and public safety contexts not be ignored. In those contexts, and many others yet to be pinpointed, autonomous systems will likely find themselves in morally charged scenarios and be expected to act with moral considerations

[1] https://www.bostonglobe.com/opinion/2015/09/07/ban-lethal-autonomous-weapons/2yI2wF0wWRjHLmNQkPiCpI/story.html

in mind. Moral reasoning cannot be an afterthought across these contexts, nor can it be disowned or deflected – robots will have too much power not to provide safeguards and guidelines for how to exercise it.

So what are the research efforts that autonomy in general presses upon us? For one, HRI must continue to probe the relational and interactive dimensions to ordinary exchanges and dialogue, as it shapes the ability of systems to serve and work socially. Healthcare institutions, for example, need to know what patients are likely to expect, how best robotic work could fit into a provider's treatment plan, and what scenarios could be morally charged for the autonomous system. A patient in agony, with an attending physician unavailble – is that worth breaking protocol to give painkillers? What kind of exchanges between a robotic tutor and students are best suited to that student's development as a scholar and citizen?

In concert with these empirical studies will be more concrete ethical treatments of how tasks and social context relate to a system's attributed status as agent or mere tool. Should robots be expected to deviate from their prescribed tasks to perform an emergency measure – saving a child who has come into the road, for example, even though the robot is a repair robot? What are the basic ethical rules that should inform those interactions? How can the concrete expertise that practitioners in these various fields possess maximally incorporate the abilities and capacities of autonomous systems – not too abstract to be practical but versatile enough to work across standard range of cases.

These empirical and conceptual facets of research must find their way into many applied contexts, including ones not usually feared. What exactly is social domestic companionship with a robot? What is being marketed to users, and what actions will users grow to depend on the system to execute? What are the responsibilities of designers for the abuse and dehumanization that could be channeled, encouraged, deflected, exploited within these relationships? The idea in not that these are dangerous *per se*, but that without adequate anticipation of morally charged situations robots will not give society adequate verification they will work constructively and competently in their particular setting.

Without combining moral accountability, a spectrum of social contexts, and empirical HRI, we will face a sequence of threats and proposed bans (weapons, sex, law enforcement) instead of grasping larger guidelines for autonomous systems' risks and benefits. Bans and proscriptions may be a way to address robotic applications that advocates may consider intrisically "anti-social," but there are many other "pro-social" applications of robotics that are fraught with potential moral import. Autonomy unexamined can spoil a "safe" context quickly, whether self-driving cars or algorithmic market-crashers. Our design and implementation of autonomous systems, as well as our consideration of human ethical and legal responsibilities with regard to the use of these agents, has to be even more comprehensive and integrated to meet such a challenge.

References

1. Scheutz, M.: The inherent dangers of unidirectional emotional bonds between humans and social robots. Robot Ethics: The Ethical and Social Implications of Robotics p. 205 (2011)