Philosophical Concepts, Robot Design, and Ethics

Scott Robbins

Department of Philosophy
Vrije Universiteit Amsterdam
s.a.robbins@vu.nl

1 Introduction

Robot designers use many philosophical concepts to design their robots. Robots have 'knowledge' and 'understanding'. They are 'smart' and more 'intelligent' than humans. They will have their own 'motivations' and 'desires' which will influence their autonomous 'choices'. These words to describe contemporary robotics come with much philosophical baggage.

This makes the debate about the ethical implications of these robots extremely difficult. For example, robots generating and acting upon their own desires indeed sounds like it could be a major ethical problem; however, what does the design of a robot which generates its own desires look like? How does a 'desire' get translated into code?

There is a concern that ethicists may be responding to public fears about robots whose capabilities in no way resemble the capabilities which the public is fearful about or the philosophical concept the ethicist is writing about. This concern is highlighted by three related but distinct epistemological gaps for robot ethics: (1) for ethicists, there is a gap between the philosophical concept used by designers to describe their robot and how the philosophical concept is used in ethics, (2) for designers, there is a cap between using a philosophical concept to describe a robot and knowing the philosophical implications of using such a word, and (3) for the public, a gap between their folk conception of the philosophical concepts used to describe robots and what the robots are capable of doing.

2 Ethicist's Epistemological Gap

Noel Sharkey nicely illustrates the giant gap between how a roboticist uses a philosophical concept, and how the concept is translated into code. In this case the concept was "guilt" and the translation was:

$\texttt{IF } V_{guilt} > Max_{guilt} \texttt{ THEN } P_{l\text{-}ethical} = \emptyset$

V_{guilt} is a counter which gets incremented whenever a perceived ethical violation occurs. When it rises above the maximum (Max_{guilt}) the machine no longer fires its

weapon. Sharkey compares this to a thermostat which cuts the heat when it reaches the threshold temperature.[1]

In this case, the part of the design referred to by the designer by the philosophical concept was rather straightforward (even for someone lacking knowledge of computer programming); however, for each philosophical concept which gets brought up in the context of robotics, how is the ethicist supposed to understand exactly how it is translated into design?

3 The Roboticist's Epistemological Gap

In the European project which generated the idea for this paper, robots are supposed to be able to generate their own "intrinsic motivations" and "values".[2] Roboticists may very well have specific results in mind which would confirm that the robots generated their own "values"; however, what would be the requirements if a philosopher was head of the design team?

What it means for the roboticist that the robot generates values, and the design by which that end state is achieved are of the utmost importance when ethically evaluating the robot. Roboticists using philosophical terms inappropriately (knowingly or unknowingly) will put their projects under ethical scrutiny which may not be warranted. How is the roboticist to know when a philosophical term is appropriate or not?

4 The Public's Epistemological Gap

When robots are described as super-intelligent, autonomous, desiring, emotional machines – there is plenty with which the public has to work with in order to be fearful. Even if robots are being appropriately described, under some minimal definition of a philosophical term, the fear (or excitement) by the public may be unwarranted. If someone were to visit a robotics lab and find out how little robots are capable of doing, would they still have the fears and excitements discussed in the news? How do we make the public informed enough to have justified beliefs about robots?

5 Conclusion

The three epistemological gaps briefly touched upon in this paper are important to overcome if we are to have a realistic, informed, and effective discourse on the ethical implications of robotics. We do not want ethicists responding to public fears about robots whose capabilities in no way resemble the capabilities which the public is fearful about or the philosophical concept the ethicist is writing about.

---

[1] Sharkey, N. E. (2012). The evitability of autonomous robot warfare. *International Review of the Red Cross*, *94*(886), 787–799. http://doi.org/10.1017/S1816383112000732

[2] http://robotsthatdream.eu/