When Robots Say No: Robot Ethics and HRI in Robot Refusal of Human Commands

Sean Welsh

University of Canterbury
sean.welsh@pg.canterbury.ac.nz

Keywords: robot force, robot refusal, robot self-reference, robot affective language

1 Introduction

This short paper seeks to explore ethical and HRI issues involved in cases that are counter to Asimov's Second Law. These are cases where robots are obliged by normative constraints to *disobey* human commands. I focus on cases that might be implemented in a bar robot subject, of course, to policy approval from government.

2 Kinetic Force

The propriety or otherwise of robots using lethal kinetic force has been extensively discussed and publicized (Future of Life Institute 2015). Non-lethal scenarios involving robots restraining violent humans or simply refusing their commands and requests are less extensively discussed.

Would the use of a bar robot as a "bouncer" to eject a violent customer (thus eliminating risk to human bar staff) be justified? Would it be acceptable for a bar robot to break up a brawl? What testing and research would be needed before such a robot could be fielded? Should robots that touch humans to restrain them be banned? Would it be acceptable for a bar robot to refuse service to humans or should it refer problem service cases to a human supervisor? Should there always be humans "on the loop" keeping an eye on bar robots?

The simplest solution (and the most likely in the near future) is to refer problem cases to human supervisors who in extreme cases could call on human security personnel and the police. Getting legal authority for robots to restrain or use force against humans is highly unlikely in the short term. Indeed, it is possible the use of force by robots on humans by police and private security firms could be banned entirely.

3 Language of Robot Refusal

If an intoxicated, disorderly or under age customer asks for an alcoholic drink in a licensed bar, under Australian and New Zealand law, the bar staff are obliged to re-

adfa, p. 1, 2011.

fuse service. If a bar robot were to be implemented, and supposing for the sake of argument it was fitted with sensors that could sense intoxication and disorder in humans (sensing age data from face images has already been done by Microsoft in the Face API in Azure), it would be able to act according to a rule. Namely, if a human customer is intoxicated, disorderly or a minor and asks for an alcoholic drink (such as a beer) then the robot must refuse service to comply with liquor laws.

How should the bar robot refuse service to intoxicated, disorderly or under age customers who ask for beer? What language should it use when it refuses service? Should this language be affectively colored when the message of refusal is spoken? Detailed experimental research into robot refusal of human commands would be interesting from a HRI perspective especially when it involves humans who may have been drinking heavily.

4 Self-Referential and Affective Language

Should robots that do not have personhood speak of themselves in the first person? Should robots use language such as “I’m sorry, sir, but…” when they have no phenomenal self and no capability to feel regret? Would such words not be a black lie? Would robots speaking of themselves as objects not as persons be morally required? If so, what would an “I, me and feeling free” dialect suitable for phenomenally and affectively vacuous robots look like?

Perhaps in such legally charged situations, robots should talk like lawyers, dispassionately and with reference to their “instructions” not themselves. Rather than say, “I’m sorry but,” which might generate an angry response along the lines of “You’re a tin can and I’m a human being with rights!” perhaps a robot should say when refusing service: “this robot apologizes but it is unable to comply, your service request for service must be referred to a human” or some such “depersonalized” language. From a HRI perspective it would be interesting to do experiments to see how personalized and depersonalized language affects anger levels in cases where robots are obliged to disobey human commands.

It would also be interesting to see how the use of “depersonalized” language might reduce unidirectional emotional bonding (Scheutz 2012) where this is unwanted.

5 References

Future of Life Institute (2015). "Autonomous Weapons: an Open Letter from AI & Robotics Researchers." Retrieved 29th July, 2015, from http://futureoflife.org/AI/open_letter_autonomous_weapons.

Scheutz, M. (2012). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. Robot Ethics. P. Lin, K. Abney and G. Bekey. Cambridge MA, MIT Press: 205-222.