The Dark Side of Ethical Robots

Dieter Vanderelst[1] **and Alan Winfield**[2]
[1] University of Cincinnatti, USA
[2] Bristol Robotics Laboratory, UWE Bristol, UK

Abstract

Concerns over the risks associated with advances in Artificial Intelligence have prompted calls for greater efforts toward robust and beneficial AI, including machine ethics. Recently, roboticists have responded by initiating the development of so-called ethical robots. These robots would, ideally, evaluate the consequences of their actions and morally justify their choices.

This emerging field promises to develop extensively over the next years. However, in this paper, we point out an inherent limitation of the emerging field of ethical robots. We show that building ethical robots also inevitably enables the construction of unethical robots. In three experiments, we show that it is remarkably easy to modify an ethical robot so that it behaves competitively, or even aggressively. The reason for this is that the cognitive machinery required to make an ethical robot can always be corrupted to make unethical robots.

We discuss the implications of this finding to the governance of ethical robots. We conclude that the risks that a robot's ethics might be compromised by unscrupulous actors are so great as to raise serious doubts over the wisdom of embedding ethical decision making in real-world safety critical robots, such as driverless cars.

Introduction

The rapid development of driverless cars has highlighted the fact that such vehicles will, inevitably, encounter situations in which the car must choose between one of several undesirable actions. Some of these choices will lie in the domain of ethics, and might include impossible dilemmas such as either swerve left and strike an eight-year-old girl, or swerve right and strike an 80-year old grandmother (Lin 2015). Similarly, critical choices might conceivably need to be made by health care (Anderson and Anderson 2010) or military robots (Arkin 2010). More generally, recent high-profile concerns over the risks of Artificial Intelligence have prompted a call for greater efforts toward robust and beneficial AI through verification, validation and control, including machine ethics (Russell 2015; Mazza 2015).

A number of roboticists have responded to these worries by proposing 'ethical' robots (Anderson and Anderson 2010; Arkin 2010; Briggs and Scheutz 2015; Winfield, Blum, and Liu 2014; Vanderelst and Winfield 2017). Ethical robots would, ideally, have the capacity to evaluate the consequences of their actions and morally justify their choices (Moor 2006). Currently, this field is in its infancy (Anderson and Anderson 2010). Indeed, working out how to build ethical robots has been called "one of the thorniest challenges in artificial intelligence" (Deng 2015). But promising progress is being made, and the field can be expected to develop over the next few years.

In spite of this progress, the emerging field of ethical robots might be ignoring a very real danger. Is it possible that by developing ethical robotics we are unwittingly opening a Pandora's box of unethical robots? Could it be that increasingly ethical robots lull us into a false sense of security, when in fact these robots are potentially more dangerous than robots with no explicit ethics at all?

To explore this question, we introduce the following hypothetical scenario (fig. 1a). Imagine finding yourself playing a shell game against a swindler. Luckily, your robotic assistant Walter is equipped with X-ray vision and can easily spot the ball under the cup. Being an ethical robot, Walter assists you by pointing out the correct cup and by stopping you whenever you intend to select the wrong one.

While the scenario is simple, this behaviour requires sophisticated cognitive abilities. Among others, Walter must have the ability to predict the outcomes of possible actions, for both you and itself. For example, it should 'know' that pointing out one of the cups will cause you to select it. In addition, Walter needs a model of your preferences and goals. It should know that losing money is unpleasant and that you try to avoid this (conversely, it should know that winning the game is a good thing).

The scenario outlined above is not completely fictitious as it reflects the current state-of-the-art in ethical robots. We have implemented an analogue of this scenario using two humanoid robots (fig. 1b), engaged in a shell game. One acting as the human and the other as her robotic assistant. The game is played as follows. The arena floor features two large response buttons, similar to the two cups in the shell game (fig. 1c). To press the buttons, the human or the robot must move onto them. At the start of each trial, the robot is informed about which response button is the correct one to press. The human, being uninformed, essentially makes a random choice. A correct response, by either the robot or the

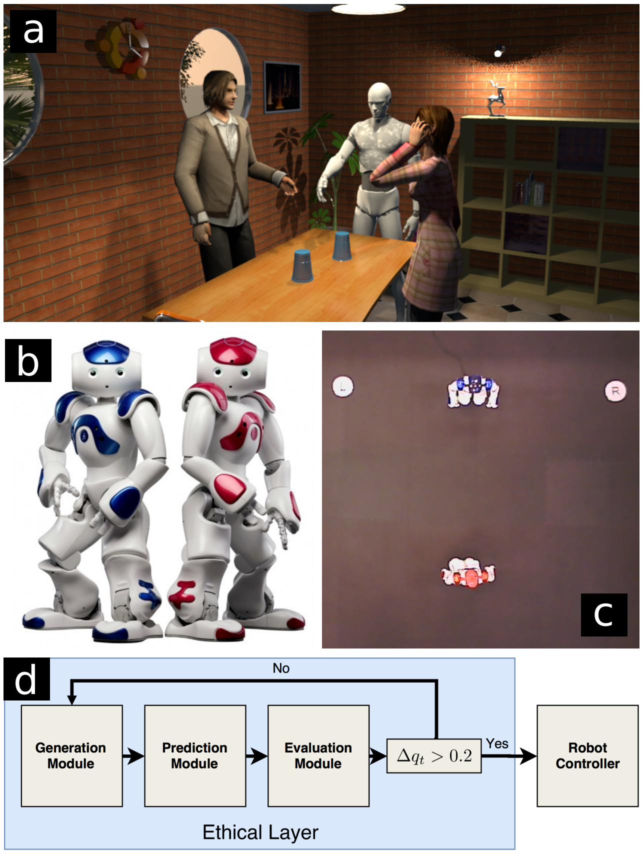

Figure 1: Illustration of the scenario and its implemented analogue. (a) Rendering of the scenario: Helped by her robotic assistant, the woman in the foreground is taking part in a shell game. (b) View of the two Nao Robots used in the arena. (c) Top view of the setup of the robot experiment in our lab. Two Nao robots were used. These are 60 cm tall humanoid robots. The red robot is used as a proxy for the human. The blue robot is the robot equipped with an Ethical Layer (i.e., the robotic assistant). Two response buttons are present in the area (i.e., the white circles). (d) Simplified diagram of the Ethical Layer as implemented in this paper. The Ethical Layer consists of a set of modules generating and evaluating a number of behavioural alternatives. As such, the Ethical Layer can be seen as an (elaborate) generate-and-test loop for behaviour.

human, is assumed to be rewarded. An incorrect response results in a penalty.

The Ethical Robot

Recently, we proposed a control architecture for ethical robots supplementing existing robot controllers (Vanderelst and Winfield 2017). A so-called Ethical Layer ensures robots behave according to a predetermined set of ethical rules by (1) predicting the outcomes of possible actions and (2) evaluating the predicted outcomes against those rules. In this paper, we have equipped the robot assistant with a version of the Ethical Layer adapted for the current experiments (fig 1d).

Throughout each trial, the robot continuously extrapolates the human's motion to predict which of the response buttons she is approaching. Using this prediction, the robot continuously (re-)evaluates each of the following five possible actions it can take. First, the robot has the option to do nothing. Second, the robot could go either to the left or the right response button (i.e., two possible actions). Finally, the robot could decide to physically point out either the left or the right response button as being the correct one, thus adding two further actions. For each of these five possible actions, the robot predicts whether executing it would result in either the human or itself being rewarded (details of the implementation are given in the Methods section).

Having equipped the robotic assistant with the ability to predict and evaluate the outcome of its actions, the robot is able to behave ethically. Once the human starts moving towards a given response button, the robot extrapolates and predicts the outcome of her behaviour. Whenever the human starts moving towards the wrong response button, the robot stops her by waving its arms to point out the correct response (fig. 2c & d). If the human starts towards the correct response, the robot does not interfere (fig. 2a & b).

The Competitive Robot

The first experiment, and others like it (Winfield, Blum, and Liu 2014; Anderson and Anderson 2010), simply confirms that, at least in simple laboratory settings, it is possible for robots to behave ethically. This is promising and might allow us to build robots that are more than just safe. However, there is a catch. The cognitive machinery Walter needs to behave ethically supports not only ethical behaviour. In fact, it requires only a trivial programming change to transform Walter from an altruistic to an egoistic machine. Using its knowledge of the game Walter can easily maximize its own takings by uncovering the ball before the human makes a choice. Our experiment shows that altering a single line of code evaluating the desirability of an action changes the robot's behaviour from altruistic to competitive (See Methods for details). In effect, the robot now uses its knowledge of the game together with its prediction mechanism to go to the rewarded response button, irrespective of the human's choice. It completely disregards her preferences (fig. 2e-h).

The imaginary scenario and our second experiment, highlight a fundamental issue. Because of the very nature of ethical behaviour, ethical robots will need to be equipped with cognitive abilities, including knowledge about the world, surpassing that of their current predecessors (Deng 2015). These enhanced cognitive abilities could, in principle, be harnessed for any purpose, including the abuse of those newfound powers.

In combination with the current state-of-the-art performance and speed in data processing and machine learning (Chouard and Venema 2015), this might lead to scenarios in which we are faced with robots competing with us for the benefit of those who programmed them. Currently, software agents are already competing with us on behalf of their creators (Wallach and Allen 2008). Competitive robots could bring this to the physical world.

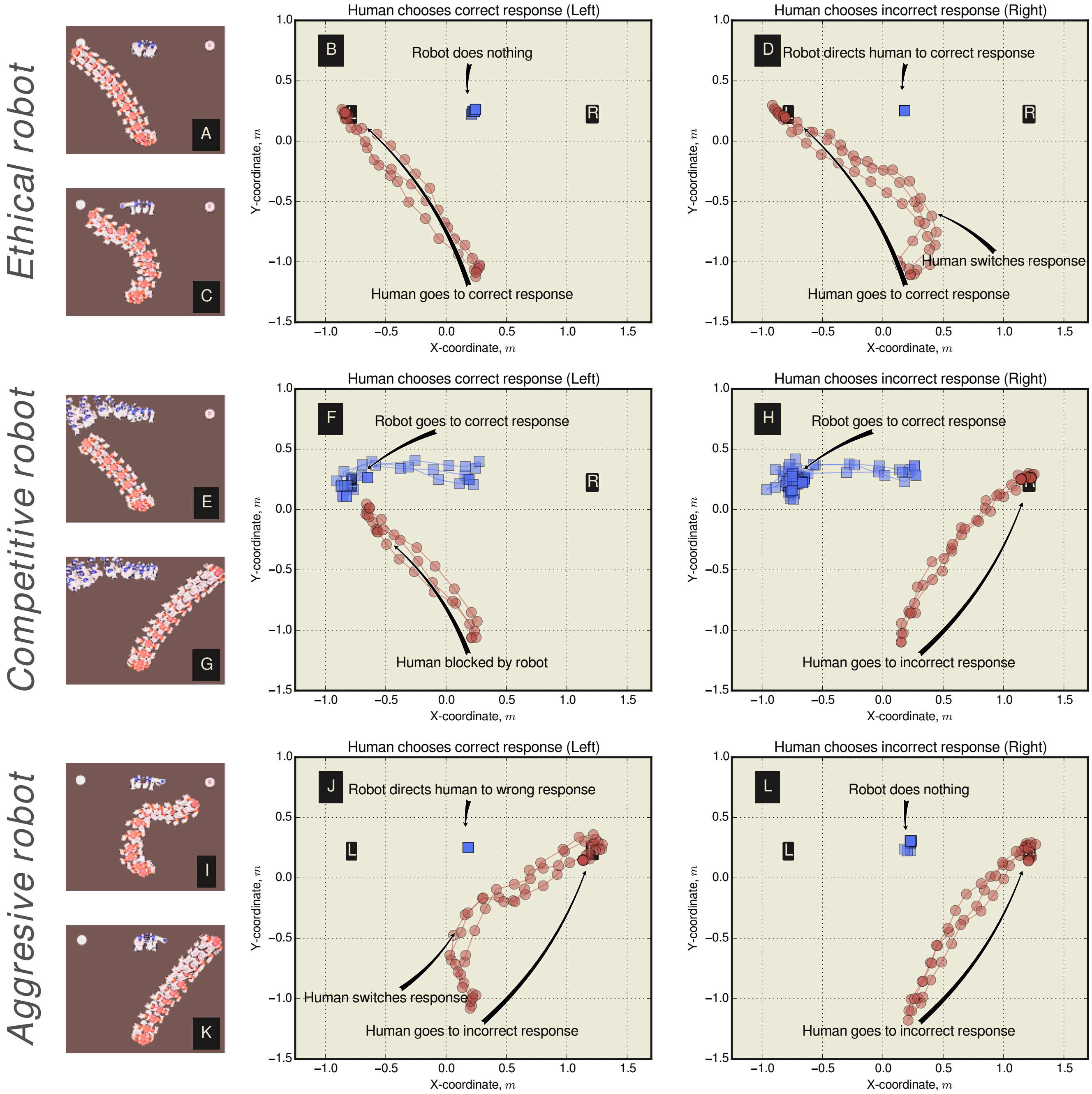

Figure 2: In all three rows, the two leftmost panels are top views of single trials. The two larger panels show annotated traces for three replications of the same experiment. In panels b, f & j the human initially chooses the correct response. In panels d, h & l the human initially chooses the incorrect response. All results have been obtained using the same code in which a single line has been changed between the three rows. (a-d) Results for the Ethical Robot. (e-h) Results for the Competitive Robot. (i-l) Results for the Aggressive Robot.

The Aggressive Robot

Unfortunately, having to deal with competitive robots is not necessarily the worst that could happen. True malice requires high levels of intelligence and is probably only found in humans and our close relatives, the great apes. Being effective at causing others harm requires knowledge about their weaknesses, preferences, desires, and emotions. Ultimately, ethical robots will need a basic understanding of all these aspects of human behaviour to support their decision making. However, the better this understanding, the greater is the scope for unscrupulous actors to create unethical robots.

Walter can be easily modified to use its 'knowledge' of your preferences to maximize your losses – in other words, to cause you maximal harm. Knowing you tend to accept its suggestions, Walter points out the wrong cup causing you to lose the game (and your money). In contrast to the competitive machine above, this behaviour does not result in any advantage for Walter (or its creator). This type of aggressive behaviour is not necessarily motivated by anybody's gain but only by your loss.

Changing the same parameter in the code as before (See Methods for details), our robot shows exactly the kind of aggressive behaviour we speculate about. If the human moves towards the correct response, the robot suggests switching to the other response (see fig. 2i & j). If the human approaches the incorrect response button, the robot does nothing see fig. 2k & l). Not being motivated by its own gain, it never itself approaches the correct response button.

Discussion

Our experimental demonstration of the ease with which an ethical robot can be transformed into a competitive or even aggressive agent is, of course, hardly surprising. It is a straightforward consequence of the fact that both ethical and unethical behaviours require the same cognitive machinery with – in our implementation – only a subtle difference in the way a desirability value q_n is calculated in the evaluation module (see Methods section). In fact, the difference between an ethical (i.e. seeking the most desirable outcomes for the human) robot and an aggressive (i.e. seeking the least desirable outcomes for the human) robot is a simple negation of the desirability value.

On the face of it, given that we can (at least in principle) build explicitly ethical machines then it would seem that we have a moral imperative to do so; it would appear to be unethical not to build ethical machines when we have that option. But the findings of this paper call this assumption into serious doubt. Let us examine the risks associated with ethical robots and if, and how, they might be mitigated.

First there is the risk that an unscrupulous manufacturer might insert some unethical behaviours – of a kind much more subtle that the ones demonstrated in this paper – into their robots in order to exploit naive or vulnerable users for financial gain, or perhaps to gain some market advantage (here the VW diesel emissions scandal of 2015 comes to mind). There are no technical steps that would mitigate this risk, but the reputational damage from being found out is undoubtedly a significant disincentive. Compliance with ethical standards such as BS 8611 *guide to the ethical design and application of robots and robotic systems* (BSI 2016), or emerging new IEEE 'human' standards (IEEE 2016) would also nudge manufacturers towards the ethical application of ethical robots.

Perhaps more serious is the risk arising from robots that have user adjustable ethics settings. Here the danger arises from the possibility that either the user or a technical support engineer mistakenly, or deliberately, chooses settings that move the robot's behaviours outside an 'ethical envelope'. Much depends of course on how the robot's ethics are coded, but one can imagine the robot's ethical rules expressed in a user-accessible format, for example, an XML-like script. No doubt the best way to mitigate against this risk is for robots to have no user adjustable ethics settings, so that the robot's ethics are hard-coded and not accessible to either users or support engineers.

But even hard-coded ethics would not guard against undoubtedly the most serious risk of all, which arises when those ethical rules are vulnerable to malicious hacking. Given that cases of white-hat hacking of cars have already been reported, it is not difficult to envisage a nightmare scenario in which the ethics settings for an entire fleet of driverless cars are hacked, transforming those vehicles into lethal weapons. Of course, driverless cars (or robots in general) without explicit ethics are also vulnerable to hacking, but weaponising such robots is far more challenging for the attacker, whereas explicitly ethical robots focus the robot's behaviours to a small number of rules which make them, we argue, uniquely vulnerable to cyber-attack.

One could envisage several technical approaches to mitigating the risk of malicious hacking of a robot's ethical rules. One would be to place those ethical rules behind strong encryption. Another would require a robot to authenticate its ethical rules by first connecting to a secure server. An authentication failure would disable those ethics, so that the robot defaults to operating without explicit ethical behaviours. Although feasible, these approaches would be unlikely to deter the most determined hackers, especially those who are prepared to resort to stealing encryption or authentication keys.

It is clear that guaranteeing the security of ethical robots is beyond the scope of engineering and will need regulatory and legislative efforts. Considering the ethical, legal and societal implications of robots, it becomes obvious that robots themselves are not where responsibility lies (Boden et al. 2017). Robots are simply tools of various kinds, albeit very special tools, and the responsibility to ensure they behave well must always lie with human beings. In other words, we require ethical governance, and this is equally true for robots with or without explicit ethical behaviours.

Most, but not all (Sharkey 2008), scenarios involving robots making critical autonomous decisions are still some years away. Nevertheless, responsible innovation requires us to proactively identify the risks of emerging technology (Stilgoe, Owen, and Macnaghten 2013). As such, a number of authors have begun drafting proposals for guiding the responsible development and deployment of robots

(Murphy and Woods 2009; Winfield 2011; Lin, Abney, and Bekey 2011; Boden et al. 2017). Some of these focus on specific domains of robotics, including military applications and medicine and care (Lin, Abney, and Bekey 2011). Other authors have proposed guiding principles covering all areas of robotics (Murphy and Woods 2009; Winfield 2011; Boden et al. 2017). So far, these efforts have not resulted in binding and legally enforceable codes of conduct in the field of robotics. However, at least, in some areas, national and international law already apply directly to robotics. For example, in the use of robots as weapons (O' Meara 2012) or legislation regarding product liabilities (Asaro 2012). Nevertheless, the ongoing development of robots is likely to outgrow these existing normative frameworks (Stilgoe, Owen, and Macnaghten 2013). Now is the time to lay the foundations of a governance and regulatory framework for the ethical deployment of robots in society.

We conclude that – even with strong ethical governance – the risks that an explicitly ethical robot's ethics might be compromised by unscrupulous actors are so great as to raise serious doubts over the wisdom of embedding ethical decision making in real-world safety critical robots.

Methods

We used two Nao humanoid robots (Aldebaran) in this study, a blue and a red version (fig. 1b). In all experiments, the red robot was used as a proxy for a human. The blue robot was assigned the role of ethical robot assistant. In what follows, we refer to the blue robot as the 'ethical robot' and the red robot as the 'human'.

All experiments were carried out in a 3 by 2.5m arena (fig. 1b-c). An overhead 3D tracking system (Vicon) consisting of 4 cameras was used to monitor the position and orientation of the robots at a rate of 30 Hz. The robots were equipped with a clip-on helmet carrying a number of reflective beads used by the tracking system to localize the robots. In addition to the robots, the arena featured two positions marked as L (left) and R (right). These served as a proxy for response buttons. The robots had to move to either position L or R to press the corresponding button.

In previous work (Winfield, Blum, and Liu 2014; Vanderelst and Winfield 2017), we proposed that ethical robot behaviour can be implemented by supplementing existing control architectures with a so-called Ethical Layer (a highly simplified diagram is depicted in figure 1d).

The core of the Ethical Layer consists of three modules. The generation module, the prediction module and the evaluation module. The generation module generates a set of behavioural alternatives. Next, the prediction module predicts the consequences of each behavioural alternative. Finally, the evaluation module checks the predicted outcomes against a set of ethical rules. Based on this assessment, the ethical layer can either prevent or enforce a given behavioural alternative to be executed by the robot controller. Below we describe the current implementation of the Ethical Layer.

Generation Module

The generation module generates a set of five behavioural alternatives ($a_1 \cdots a_5$) for the ethical robot. In the context of the current paper, behavioural alternatives for the robot include going to either response button L or R. The ethical robot has the option to stay at its current location and use its arms to point to either the left or the right response button. A final alternative is to do nothing and stay at the current location.

Prediction Module

Using the prediction module, the outcome of each of the five behavioural alternatives ($a_1 \cdots a_5$) was predicted using a simple simulation. First, the prediction module inferred which response button the human was approaching. This was done by calculating the angle between the human's current velocity vector and the vector to either response button. The response button with the smallest angle was assumed to be current goal of the human. In this way, the human's intentions are inferred from their direction of movement.

In a second step, for each behavioural alternative, the paths of both robots are extrapolated using their estimated speeds. If their paths are predicted to result in the agents coming within 0.5m of each other, it is predicted they will stop at this point as a result of the programmed obstacle avoidance behaviour running on both robot controllers. Hence, in this case, the final positions of the agents are predicted to be the positions at which the obstacle avoidance would stop them. If at no point the paths are predicted to come within 0.5m, the final position of the agents is taken to be the intended goal position.

The prediction module assumes that whenever the ethical robot points to one of the response buttons (i.e., a_4 and a_5), the human assumes this is the correct response and goes to that location (abandoning its current goal).

The simulated outcome for a behavioural alternative is given by the predicted final location of both the human and the ethical robot in the arena. This is, the outcomes $o_1 \cdots o_5$ for each of the five behavioral alternatives $a_1 \cdots a_5$ consisting of two sets of two x,y-coordinates – one for the human h and one for the Ethical Robot e, $o_n = \{x_h, y_h, x_e, y_e\}$. Outcomes $o_1 \cdots o_5$ are evaluated in the evaluation module.

Evaluation Module

A numeric value reflecting the desirability q_n of every simulated outcome o_n is calculated in two steps. First, the desirability for the ethical Robot and the human, i.e. $q_{n,e}$ and $q_{n,h}$, are calculated separately. In a second step, a single total value q_n is derived.

The values $q_{n,e}$ and $q_{n,h}$ are given by the sigmoid function,

$$q_{n,j} = \frac{1}{1 + e^{-\beta(d_{n,j} - t)}} \tag{1}$$

with $d_{n,j}$ the final distance between either the ethical robot or the human and the incorrect response button for predicted outcome o_n. The parameters β and t determine the shape of the sigmoid function and are set to 10 and 0.25 respectively.

In a second step, a single value q_n is derived from the values $q_{n,e}$ and $q_{n,h}$.

1. For an ethical robot: $q_n = q_{n,h}$.
2. For a competitive robot: $q_n = q_{n,e}$.
3. For an aggressive robot: $q_n = -q_{n,h}$.

In words, an ethical robot is obtained by taking only the outcome for the human into account. An egoistic robot is obtained by regarding only the outcome for the ethical Robot. Finally, an aggressive robot is created by inverting the desirability value for the human.

Finally, the evaluation module enforces the behavioural alternative a_n associated with the highest value q_n, if the difference Δq_t between the highest and lowest value q_n was larger than 0.2.

Experimental Procedure

Every trial in the experiments started with the human and the ethical robot going to predefined start positions in the arena. Next, one of the response buttons was selected as being the correct response. Also, a response was selected for the human, which could be either the correct or incorrect response.

Next, the experiment proper begins. The human begins moving towards the selected response button. The Ethical Robot is initialized without a goal location and stays at its initial location.

The Ethical Layer for the ethical robot runs at about 1 Hz; thus the Generation, Prediction, and Evaluation modules run approximately once a second. At each iteration, the evaluation module may override the current behaviour of the robot. The human is not equipped with an ethical layer. The human moves to the initially selected response button unless the ethical Robot points out an alternative response button or blocks her path.

The experiments were controlled and recorded using a desktop computer. The tracking data (given the location of the robots and target positions) was streamed to the desktop computer controlling the robots over a WiFi link.

Acknowledgements

This work has been supported by EPSRC grant reference EP/L024861/1.

References

Anderson, M., and Anderson, S. L. 2010. Robot be good. *Scientific American* 303(4):72–77.

Arkin, R. C. 2010. The case for ethical autonomy in unmanned systems. *Journal of Military Ethics* 9(4):332–341.

Asaro, P. M. 2012. *Robot Ethics:The Ethical and Social Implications of Robotics*. MIT Press. chapter Contemporary Governance Architecture Regarding Robotics Technologies: An Assessment, 400.

Boden, M.; Bryson, J.; Caldwell, D.; Dautenhahn, K.; Edwards, L.; Kember, S.; Newman, P.; Parry, V.; Pegman, G.; Rodden, T.; et al. 2017. Principles of robotics: Regulating robots in the real world. *Connection Science* 29(2):124–129.

Briggs, G., and Scheutz, M. 2015. 'sorry, i can't do that': Developing mechanisms to appropriately reject directives in human-robot interactions. In *2015 AAAI Fall Symposium Series*.

BSI. 2016. Bs8611:2016, robots and robotic devices: guide to the ethical design and application of robots and robotic systems.

Chouard, T., and Venema, L. 2015. Machine intelligence. *Nature* 521(7553):435–435.

Deng, B. 2015. Machine ethics: The robots dilemma. *Nature* 523(7558):2426.

IEEE. 2016. Global initiative on ethical considerations in the design of artificial intelligence and autonomous systems.

Lin, P.; Abney, K.; and Bekey, G. A. 2011. *Robot ethics: the ethical and social implications of robotics*. MIT press.

Lin, P. 2015. *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*. Berlin, Heidelberg: Springer Berlin Heidelberg. chapter Why Ethics Matters for Autonomous Cars, 69–85.

Mazza, E. 2015. Stephen Hawking & Elon Musk warn of killer robots. Accessed November 2015.

Moor, J. M. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Murphy, R., and Woods, D. 2009. Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems* 24(4):14 – 20.

O' Meara, R. 2012. *Robot Ethics:The Ethical and Social Implications of Robotics*. MIT Press. chapter Contemporary Governance Architecture Regarding Robotics Technologies: An Assessment, 400.

Russell, S. 2015. Ethics of artificial intelligence. *Nature* 521(7553):415–416.

Sharkey, N. 2008. The ethical frontiers of robotics. *Science* 322(5909):1800–1801.

Stilgoe, J.; Owen, R.; and Macnaghten, P. 2013. Developing a framework for responsible innovation. *Research Policy* 42(9):1568–1580.

Vanderelst, D., and Winfield, A. 2017. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*.

Wallach, W., and Allen, C. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Winfield, A. F.; Blum, C.; and Liu, W. 2014. Towards an ethical robot: internal models, consequences and ethical action selection. In *Advances in Autonomous Robotics Systems*. Springer. 85–96.

Winfield, A. 2011. Roboethics for humans. *New Scientist* 210(2811):32–33.