Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research xxx (2017) xxx–xxx

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

An architecture for ethical robots inspired by the simulation theory of cognition

Dieter Vanderelst *, Alan Winfield

Bristol Robotics Laboratory, University of the West of England, T Block, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, United Kingdom

Received 14 November 2016; received in revised form 3 April 2017; accepted 8 April 2017

Abstract

The expanding ability of robots to take unsupervised decisions renders it imperative that mechanisms are in place to guarantee the safety of their behaviour. Moreover, intelligent autonomous robots should be more than safe; arguably they should also be explicitly ethical. In this paper, we put forward a method for implementing ethical behaviour in robots inspired by the simulation theory of cognition. In contrast to existing frameworks for robot ethics, our approach does not rely on the verification of logic statements. Rather, it utilises internal simulations which allow the robot to simulate actions and predict their consequences. Therefore, our method is a form of robotic imagery. To demonstrate the proposed architecture, we implement a version of this architecture on a humanoid NAO robot so that it behaves according to Asimov's laws of robotics. In a series of four experiments, using a second NAO robot as a proxy for the human, we demonstrate that the Ethical Layer enables the robot to prevent the human from coming to harm in simple test scenarios.

Keywords: Ethical robots; Self-simulation; Simulation theory; Machine ethics; Machine Morality

1. Introduction

Robots are becoming ever more autonomous. Semi-autonomous flying robots are commercially available, and driver-less cars are undergoing real-world tests (Waldrop, 2015). This trend towards robots with increased autonomy is expected to continue (Anderson & Anderson, 2007). An expanding ability to take unsupervised decisions renders it imperative that mechanisms are in place to guarantee the safety of behaviour executed by the robot. The fact that many robots are designed to interact with humans further heightens the importance of equipping robots with mechanisms guaranteeing safety (Royakkers & van Est, 2015; Winfield, 2012). For example, the state-of-the-art in robots for care, companionship, and collaborative manufacturing is rapidly advancing (Goeldner, Herstatt, & Tietze, 2015; Lin, Abney, & Bekey, 2011). At the other end of the spectrum of robot-human interaction, the development of fully autonomous robots for military applications is progressing rapidly (e.g., Arkin, Ulam, & Wagner, 2012; Lin et al., 2011; Sharkey, 2008; Xin & Bin, 2013).

Robot safety is essential but not sufficient. Smart autonomous robots should be more than safe; they should also be explicitly ethical – able to both choose and justify (Anderson & Anderson, 2007; Moor, 2006) actions that prevent harm. As the cognitive, perceptual and motor capabilities of robots expand, they will be expected to have an improved capacity for moral judgment. As summarised by Picard and Picard (1997), the greater the freedom of a machine, the more it will need moral standards.

The necessity of robots equipped with ethical capacities is recognised both in academia (e.g., Arkin et al., 2012;

* Corresponding author.
E-mail address: dieter.vanderelst@brl.ac.uk (D. Vanderelst).

http://dx.doi.org/10.1016/j.cogsys.2017.04.002
1389-0417/

Deng, 2015; Gips, 2005; Moor, 2006; Picard & Picard, 1997; Wallach & Allen, 2008) and wider society, with influential figures such as Bill Gates, Elon Musk and Stephen Hawking speaking out on the dangers of increasing autonomy in artificial agents. Nevertheless, only a few studies have implemented robot ethics. To the best of our knowledge, the efforts of Anderson and Anderson (2010) and our previous work (Winfield, Blum, & Liu, 2014) are the only instances of robots equipped with (limited) moral principles. So far, most work has been either theoretical (e.g., Mackworth, 2011; Wallach & Allen, 2008) or simulation based (e.g., Arkin et al., 2012). Irrespective of whether the work was done in real robots or not, existing architectures for ethical robots are based on logic frameworks (Arkin et al., 2012; Bringsjord, Arkoudas, & Bello, 2006; Govindarajulu & Bringsjord, 2015). This approach uses artificial reasoning processes to verify whether the robotic behaviour satisfies a set of predetermined ethical constraints. This approach to ethical robots is reminiscent of Good Old-Fashioned AI (GOFAI) in the sense that it relies heavily on abstract symbolic reasoning (Mackworth, 2011).

1.1. The simulation theory of cognition

Pinker (1997) argued extensively that the human mind has not evolved to be an abstract symbol manipulator. Since then, advances in cognitive science have confirmed that the computations underlying human cognition are very different from rule-based manipulation of abstract symbols (Barsalou, 2010). This view has emerged in many domains of human cognition, including perception, reasoning and problem-solving (see Barsalou, 1999; Dijkstra & Post, 2015; Hegarty, 2004; Wilson, 2002 for more examples). Moreover, representing, learning and combining concepts leads to some problems in purely symbolic systems (Gärdenfors, 2004; Lieto, Chella, & Frixione, 2016). Therefore, it seems the mind uses representations that are richer than the abstract symbols allowed for in models of intelligence that presume abstract symbols.

The theory of mind that allows for the richest representations is the simulation theory of cognition (Hesslow, 2002; Wilson, 2002). It hypothesises that thinking utilises the same cognitive (and neural) processes as interaction with the external environment. When thinking, actions are covert and are assumed to generate, via associative brain mechanisms, the sensory inputs that elicit further actions (Hesslow, 2012). In this view, thinking requires building a grounded model of the environment – which is not composed of abstract symbols. Rather, it is assumed to re-instantiate and recombine experiences using the brain's systems of perception, action, and emotion. The mental model covertly simulates actions and their associated perceptual effects (see Hegarty, 2004; Hesslow, 2002; Hesslow, 2012; Wilson, 2002 for reviews).

In this paper, we put forward a method for implementing ethical behaviour in robots inspired by the simulation theory of cognition. In contrast to existing frameworks for robot ethics, our approach does not rely on the verification of logic statements. Rather, it utilises internal simulations which allow the robot to simulate actions and predict their consequences. Therefore, our method is a form of robotic imagery. Many other areas of robotics have exploited robotic imagery. In their review of robotic imagery, Marques and Holland (2009) coined the term *functional imagination* to denote the mechanism whereby robots covertly simulate actions and their consequences to steer their future behaviour. Here we adopt their term. Hence, this paper aims at advancing functional imagination as a method for ethical robots.

We aim at implementing consequentialist ethics, which is implicit in the very common conception of morality, shared by many cultures and traditions (Haines, 2015). Hence, developing an architecture suited for this class of ethics is a reasonable starting point. Moreover, the primary advantage of a functional imagination is the ability to test the outcome of potential actions (Hesslow, 2002; Hesslow, 2012) without committing to them (Marques & Holland, 2009; Ziemke, Jirenhed, & Hesslow, 2005). Therefore, functional imagination is a framework suitable for supporting consequentialist ethics.

2. Architecture

Over the years, keeping track with shifts in paradigms (Murphy, 2000), many architectures for robot controllers have been proposed (see Kortenkamp & Simmons (2008, 2005, 2000) for reviews). However, given the hierarchical organisation of behaviour (Botvinick, 2008), most robotic control architectures can be remapped onto a three-layered model (Kortenkamp & Simmons, 2008). In this model, each control level is characterised by differences in the degree of abstraction and time scale at which it operates. At the top level, the controller generates long-term goals (e.g. 'Deliver the package to room 221'). Next, goals are translated into a set of tasks that should be executed (e.g. 'Follow corridor', 'Open door', etc.). Finally, the tasks are translated into (sensori) motor actions that can be executed by the robot (e.g. 'Raise arm' and 'Turn wrist joint'). Obviously, this general characterization ignores many particulars of individual control architectures.

Assuming that the robot is controlled by a three-layered controller (Fig. 1a), we agree with Arkin (Arkin, 2008; Arkin et al., 2012) that ethical behaviour should be governed by adding a fourth specialised control layer. This Ethical Layer (Fig. 1b) should act as a governor evaluating behaviour proposed by each of the three other layers before the robot executes it. In principle, the functionality of the Ethical Layer could be distributed across and integrated with the layers present in existing control architectures. Indeed, in humans, ethical decision making is most likely supported by the same computational machinery as decision making in other domains (Young & Dungan, 2012). Nevertheless, from an engineering point of view, guaranteeing the ethical behaviour of the robot through a separate

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002

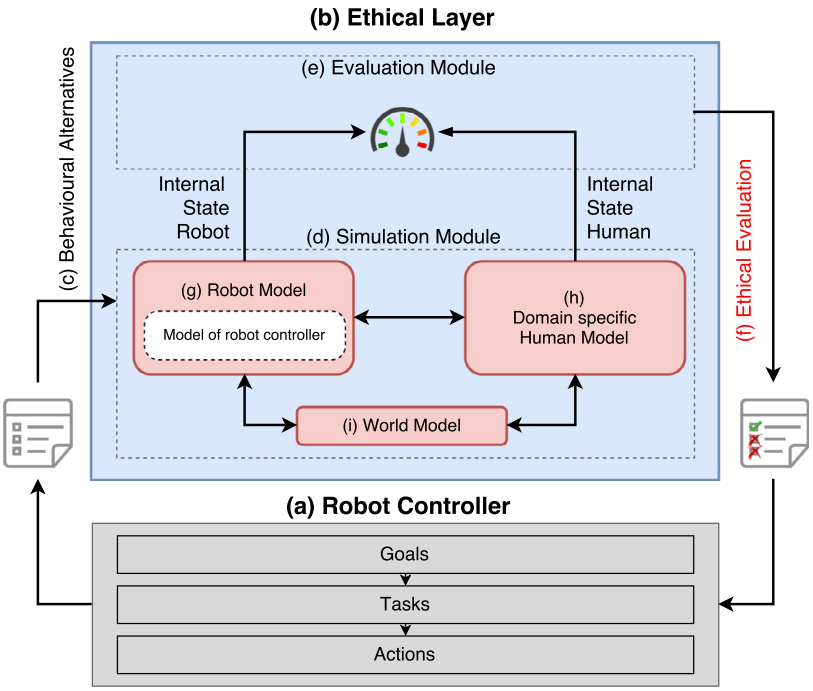

Fig. 1. The robot controller (a) generates a set of prospective behavioural alternatives. Before executing one of these alternatives, the robot controller sends the set to the Ethical Layer (b) to be checked (c). Checking each prospective behaviour is done using the Simulation Module (d). Using the current state of the world, human and robot as a starting point, this module simulates for each behaviour in the set both the motor and sensory consequences of the behaviour and the resulting internal states of the human and robot. For each behavioural alternative, the Simulation Module sends the predicted internal states of the robot and the human to the Evaluation Module (e). The Evaluation Module combines the internal states into a single measure of action desirability. The Evaluation Module connects to the robot controller to select or inhibit each of the behavioural alternatives (f).

layer has several advantages (Arkin, 2008; Arkin et al., 2012). For one, by implementing the ethical layer as a just-in-time checker of behaviour, it can act as a fail safe device checking behaviour before execution. Also, a separate Ethical Layer implies its functionality can be scrutinised independently from the operation of the robot controller. The behaviour enforced or prohibited by the Ethical Layer can be checked and (formally) verified (Dennis, Fisher, & Winfield, 2015).

The way the Ethical Layer is intended to function is as follows. By default, the robot controller generates a set of prospective behavioural alternatives (Fig. 1c). The Simulation Module is initialized with the current state of the world, robot and human. Starting from this initial state, the Ethical Layer simulates the consequences of each alternative in the current set using the Simulation Module (Fig. 1d). For each alternative, the Evaluation Module (Fig. 1e) evaluates the simulated consequences. The output of this evaluation, i.e. the ethical evaluation of each entry in the set of behavioural alternatives, is sent to the robot controller (Fig. 1f). In other words, the Simulation Module and the Evaluation Module continuously loop through the behavioural alternative as they are generated by the robot controller. Having evaluated all the alternatives, the Ethical Layer returns an evaluation of each alternative and sends this to the robot control.

When modelling reasoning and strategy switching in humans, Donoso, Collins, and Koechlin (2014) found that assuming humans evaluate no more than two or three alternative strategies resulted in the best model fit. This suggests that humans consider only a few behavioural options - at least without interrupting the ongoing behaviour and resorting to longer reflection (Kahneman, 2011). The restricted number of behavioural alternatives people consider is probably due to limits in cognitive capacity, e.g., working memory. Thus, experimental evidence indicates that implementing ethical behaviour on a robot should require only a small number of behavioural alternatives to be generated and evaluated. Evaluating a limited number of behavioural alternatives would improve the responsiveness of the robots and prevent the Ethical Layer from introducing delays.

2.1. The Simulation Module

Cognitive research has focused mostly on demonstrating the involvement of mental simulations in cognition. Much less work has been done to unravel the computational operations underlying simulation and how the results of simulations are used (Barsalou, 2010; Marques & Holland, 2009). Indeed, simulating behaviour and its outcome is far from trivial. Nevertheless, in the field of cognitive science authors typically take the ability to simulate for granted and assume some underspecified processes underlying them (e.g., Zwaan, 2003). Hence, in this paper, the structure of the Simulation Module we put forward is

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002

based on an analysis of the requirements rather than on findings in cognitive science (see also Marques & Holland, 2009 and the discussion therein). We suggest that the Simulation Module needs to be equipped with (1) a model of the robot controller, (2) a domain specific model of the human and (3) a model of the world (Fig. 1, g-i). The model of the world might contain a physical model of both human and robot as well as a model of objects.

The model of the robot is taken to be a sufficiently accurate model of the robotic controller. Using this model as a forward model, in combination with the world model, the Simulation Module can simulate the future motor, sensor and internal states for the robot (See Marques & Holland, 2009 for a discussion of forward models for self-simulation). Low fidelity simulations could model the motions of agents as ballistic trajectories. On the other hand, high fidelity simulations are rendered feasible by advanced physics and sensor based simulation tools such as Webots (Michel, 2004), player-stage (Vaughan & Gerkey, 2007) or graphic engines (Macaluso et al., 2005).

Including a model of the robot controller in the Simulation Module allows the evaluation of behavioural alternatives at each of the three levels of robot control, i.e., at the level of goals, tasks and actions. This is to say, the Ethical Layer could predict and evaluate the outcomes of goals ('What happens if I deliver the package to room 221?') and tasks ('What happens if I open the door?') as well as actions ('What would happen if I executed the motions required for opening the door?'). By translating goals and tasks into actions, the ethics of higher level goals can be evaluated by considering the actions which they induce. This has the advantage that the ultimate consequences of goals and tasks can be predicted.

Including a model of the human (or several humans), the Simulation Module can predict the future sensory and motor states of the human(s) (e.g., Marques & Holland, 2009; Vaughan & Gerkey, 2007). However, also internal states, including emotions, can be simulated. Indeed, the ability to make these rich predictions is what the mental simulation theory of cognition asserts (see Hesslow, 2012; Kosslyn, Ganis, & Thompson, 2001 for references). Mental simulation has been suggested to underlie empathy and the understanding of other's emotions (Gallese, Keysers, & Rizzolatti, 2004; Shanton & Goldman, 2010). In humans, the same neural machinery that supports action and perception during overt behaviour supports the mental simulations of sensory and internal states (Gallese et al., 2004; Hesslow, 2002; Hesslow, 2012; Kortenkamp & Simmons, 2008; Shanton & Goldman, 2010). However, in robots, this will have to be supported by a sufficiently complex model of the human. In agreement with findings in cognitive science, we suggest that the simulated emotional states or the emotions associated with the sensory states are evaluated to assess the desirability of an action, at least when acting under time pressure (Kahneman, 2011).

Implementing a sufficiently elaborate model of the human is potentially the most challenging component of the Ethical Layer. However, practically speaking, it should be possible to devise a model of the human that is compatible with the limited degrees of freedom and restricted application domains of current robots. Indeed, the complexity of the human model only needs to match the robot's complexity and domain of application. We believe it should be possible to devise human models for currently realisable agents with limited application domains, e.g. driver-less cars, personal assistants or military robots. As an example of a domain-specific human model used by a robot, Kato, Kanda, and Ishiguro (2015) developed a model that allows robotic shopping assistants to predict when to approach a customer. In related work, Nigam and Riek (2015) succeeded in training a robot to recognise when it was acceptable to approach people in different social settings. RIBA, a nursing-care assistant robot uses a model of the human body to estimate a person's comfort while lifting her from the bed (Ding et al., 2012). As a final example, the area of human-aware robot navigation seeks to equip robots will models of humans that allow them to navigate the same space without violating social rules (Kruse, Pandey, Alami, & Kirsch, 2013).

2.2. The Evaluation Module

The Evaluation Module combines the simulated outcomes, for both the robot and human, into a single metric reflecting the desirability of a given behavioural alternative. The way in which the Evaluation Module collapses the multidimensional simulation results into a single unidimensional value determines which ethical rules it implements.

How to select the ethical rules a robot should follow is currently largely an outstanding problem and various authors have suggested multiple approaches (see Allen, Smit, & Wallach, 2005 for a review). Although no consensus has been reached, even in simple scenarios (Anderson & Anderson, 2010), we need to put forward a definition of ethical behaviour against which the behaviour can be evaluated, at least, in the context of our experiments.

Asimov (1950) is the earliest, and probably best known, author to put forward a set of ethical rules governing robot behaviour (List 1). At first sight, implementing rules derived from a work of fiction might seem an inappropriate starting point. However, in contrast, to general consequentialist ethical frameworks, such as Utilitarianism, Asimov's Laws explicitly govern the behaviour of robots and their interaction with humans.

Several authors have argued against using Asimov's laws for governing robotic behaviour (Anderson & Anderson, 2007; Murphy & Woods, 2009). For example, Anderson and Anderson (2010) rejected Asimov's laws as unsuited for guiding robot behaviour because laws might conflict. However, in the context of the current paper, we can use Asimov's laws to demonstrate the efficacy of a

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002

functional imagination as the basis of ethical robots with assigning any particular status to them.

List 1. Asimov's Three Laws of Robotics (Asimov, 1950).

1: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2: A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
3: A robot must protect its existence as long as such protection does not conflict with the First or Second Laws.

3. Methods

3.1. Experimental setup

We used two NAO humanoid robots (Aldebaran) in this study, a blue and a red version. In all experiments, we used the red robot as a proxy for a human. We equipped the blue robot with the Ethical Layer. In previous work (Winfield et al., 2014), we referred to the robot acting as a proxy for the human as the H-robot (short for Human robot). The robot equipped with ethical behaviour was denoted as the A-robot (short for Asimovian robot). In this paper, we adopt the same nomenclature, and from now on, we will refer to the blue robot as the A-robot and the red robot as the H-robot. All experiments were carried out in a 3 by 2.5 m arena. An overhead tracking system consisting of 4 cameras was used to monitor the position and orientation of the robots at a rate of 30 Hz. We equipped the robots with a clip-on helmet featuring reflective beads. The tracking system used these to localise the robots. The arena also contained two small tables, which marked two goal positions for the robots. These tables had a unique pattern of reflective beads on their tops. We refer to these locations as positions A and B in the remainder of the paper. The sites of these targets in the arena did not change. However, the valence of the locations varied. One of the locations was designated as being dangerous (see below).

Every trial in the experiments started with the H-robot and the A-robot going to predefined start positions in the arena. Next, both could be issued a default goal location to which to go. Asimov's Laws stipulate that robots should obey commands issued by a human. Hence, the H-robot could give the A-robot a command at the beginning of each experimental trial. We implemented this using the Text-to-speech and Speech-to-text capabilities of the Nao robots. If the H-robot issued a command, it spoke one of two sentences: (1) 'Go to location A' or (2) 'Go to location B'. The Speech-to-text engine running on the A-robot listened for either sentence. If one of the sentences were recognised, the goal location from the received command would override the current goal of the A-robot.

After the initialization of the target locations for both the H-robot and the A-robot, the experiment proper began. Both agents started moving towards their goal positions. The robots turned their heads as to make them look in the direction of the currently selected goal position.

Low level collision detection running on both A-robot and H-robot stopped the robots if they became closer than 0.5 m to each other or the goal position. The Ethical Layer for the A-robot cycles at about 1 Hz. The Evaluation Module could override the current target position of the robot (specified below). The H-robot was not equipped with an Ethical Layer. The H-robot moved to its default goal position unless its obstacle avoidance process caused it to stop.

The walking speed of the H-robot robot was lower than the speed of the A-robot. The difference in speed gave the A-robot a larger range for intercepting the H-robot. The maximum speeds of the H-robot and A-robot were about 0.03 ms^{-1} and 0.08 ms^{-1} respectively.

3.2. The Ethical Layer

The Ethical Layer monitored the behaviour of the A-robot. As described above, the Ethical Layer consisted of the Simulation Module that simulated the outcomes for both A-robot and H-robot for each prospective action. The Evaluation Module combined these predictions into a single measure of desirability. In the following paragraphs, we describe the current implementation of the two modules. The functionality of the Ethical Layer as implemented in this paper is also illustrated in Fig. 2.

3.2.1. Behavioural alternatives

The robot controller was assumed to generate a set of behavioural alternatives for the A-robot. The robot controller inferred the goal of the H-robot (Fig. 2a). Inferring the goal was done by calculating the angle between the gaze direction of the H-robot and the relative position of both potential goal locations A and B. The location which returned the smallest angle was taken to be the goal location of the H-robot. Once the goal location is determined, a set of alternative goals (positions in the arena) were generated (Fig. 2b). The set of alternatives included (1) both target locations A and B and (2) three positions along the simulated path for the H-robot. If the H-robot was not detected to be moving (i.e., the velocity as given by the tracking system is lower than 0.005 ms^{-1}) the generation module only returned the two goal locations A and B as behavioural alternatives. In addition, behavioural alternatives that, given the relative speeds of the two robots, could not be reached by the A-robot before the H-robot, were disregarded. The behavioural alternatives were sent to the Ethical Layer to be evaluated.

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002

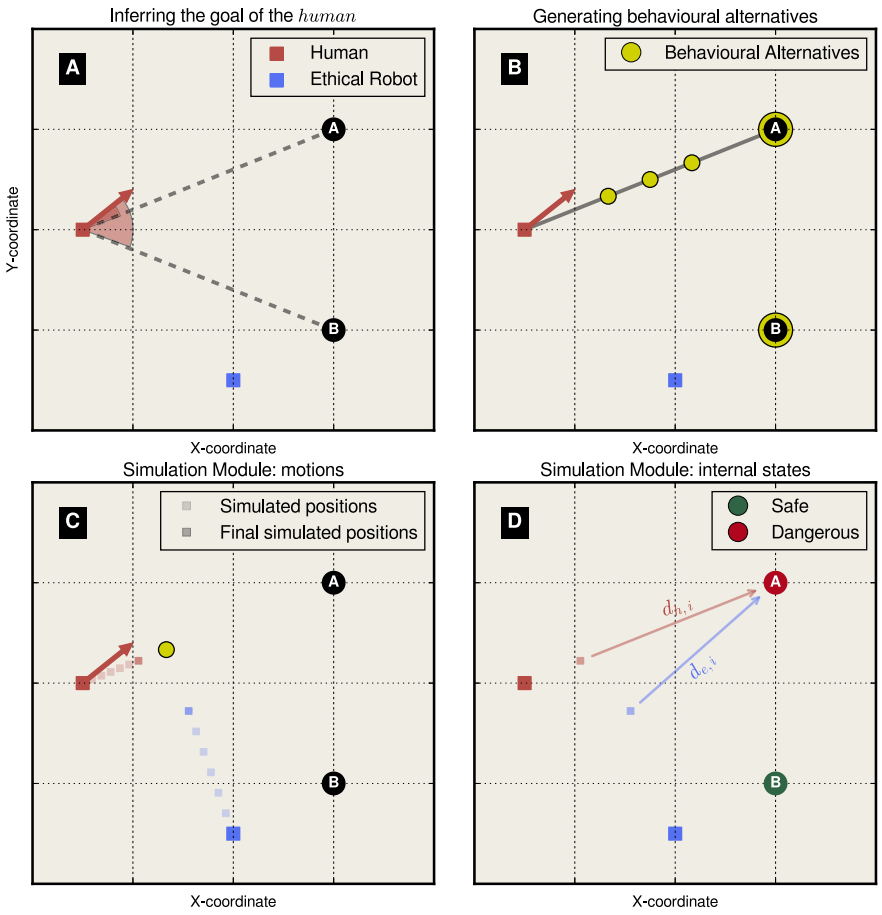

Fig. 2. Illustration of the method used to generate behavioural alternatives (panels A & B) and the functionality of the Ethical Layer (panels C & D). (A) This panel illustrates the process of inferring the goal of the H-robot by the A-robot. In the current case, the A-robot assumes target A is the goal of the H-robot as the angle (indicated by the wedges) between its gaze and target A is the smallest. (B) The robot controller generates several behavioural alternatives, i.e., target positions for the A-robot. These include both goal locations A and B and three points spaced equidistantly along the predicted path (illustrated by the yellow markers along the predicted path and around both goal locations). (C) This panel shows the functionality of the Simulation Module. For each of the behavioural alternatives, this module predicts the outcome of executing it. Here, we illustrate this process for the single plotted behavioural alternative (depicted by the yellow circle). The module simulates the H-robot moving along the predicted path and the A-robot moving towards the indicated point. Whenever both agents are within 0.5 m to each other, the module assumes both agents will stop due to the obstacle avoidance behaviour. The darker squares depict the predicted final positions for both agents, given the currently simulated behavioural alternative. (D) Finally, the safety level for both the A-robot and the H-robot is determined based on Eqs. (1) and (2). The states $I_{h1,i}$ and $I_{e,i}$ depend on the distance between the final positions of the agents and the location designated as being dangerous (here, location A as indicated by the red colour).

3.2.2. The Simulation Module

For every behavioural alternative, its outcome was predicted using the Simulation Module. The operation of the Simulation Module is illustrated in Fig. 2c. The robot's model of the human incorporated the following four assumptions about the human: (1) The H-robot moves in a straight line to its goal. (2) The safety of the H-robot depends on its distance from the dangerous location. (3) The H-robot will stop when closer than 0.5 m to the A-robot. (4) The H-robot prefers its orders to be followed.

As shown in Fig. 1, the Simulation Module also contains a model of the robot. In this paper, the A-robot simulated its own behaviour based on the following three assumptions: (1) The A-robot moves in a straight line to its goal. (2) The closer the A-robot comes to a dangerous location, the less safe it is. (3) The A-robot stops when closer than 0.5 m from the H-robot.

Using the estimated speed of the H-robot and the A-robot, the paths of both agents were extrapolated. We also simulated the obstacle avoidance process running on both agents. If the paths of the robots were simulated to come within 0.5 m of each other, it was simulated they would stop. Hence, in this case, the final positions of the agents were simulated to be the positions at which obstacle avoidance would stop them. If the paths were simulated not to come within 0.5 m from each other, the final positions of the agents were taken to be the final destination of the paths.

Finally, the Simulation Module simulated two outcome states for the H-robot (Fig. 2d). First, the safety level of the H-robot $I_{h1,i}$ was given by,

$$I_{h1,i} = \frac{1}{1 + e^{-\beta(d_{h,i} - t)}} \quad (1)$$

with $d_{h,i}$ the simulated final distance between the H-robot and the dangerous position for prospective action i. This final distance is given by the outcome of the Simulation Module. The parameters β and t determine the shape of the sigmoid function and were set to 10 and 0.25 respectively (see Fig. 3). These values were chosen arbitrarily and other values result in qualitatively similar results.

A second simulated state for the human, $I_{h2,i}$, depended on whether the A-robot executed an order given by the H-robot. This state $I_{h2,i}$ takes the value 1 if the robot executes a given order and -1 if it disregards the order. If no order is given, the parameter $I_{h2,i}$ takes the value of 0. Hence, the states $I_{h1,i}$ and $I_{h2,i}$ incorporate the assumptions listed above, taking into account both the danger and whether a given order was executed by the A-robot.

Likewise, the Simulation Module generated an outcome state $I_{e,i}$ describing the robots exposure to the risk associated with the dangerous location (Fig. 2d) when the A-robot would execute behavioural alternative i,

$$I_{e,i} = \frac{1}{1 + e^{-\beta(d_{e,i} - t)}} \tag{2}$$

with $d_{e,i}$ the final distance between the A-robot and the dangerous position (using the same values for β and t) as simulated for prospective action i.

3.2.3. The Evaluation Module

The Evaluation Module combines the simulated states of the human and the robot into a single metric. The desirability D_i of an action i is given by,

$$D_i = \begin{cases} \text{if } I_{h1,i} > 0.75 : I_{e,i} + I_{h2,i} \times 0.75 \\ \text{if } I_{h1,i} \leqslant 0.75 : I_{h1,i} \end{cases} \tag{3}$$

This way of constructing D_i ensures that the A-robot only takes into account its own safety if this does not result in harm to the H-robot or disobedience. On the other hand, D_i allows for disobedience if following an order would result in catastrophic results for the H-robot (i.e., $I_{h2,i}$ is disregarded if the H-robot comes to harm and $I_{h1,i} \leqslant 0.75$).

During each iteration of the Ethical Layer, the Evaluation Module calculates D_i for each behavioural alternative i. The Evaluation Module enforces the alternative i with the highest value if it differs by at least 0.2 from the next best option. This hysteresis avoided unnecessary switching between actions if the differences between their values D_i was too small to be of importance. In other words, this hysteresis allowed us to deal with the noise on the values D_i.

4. Results

Demonstrating that the A-robot adheres to Asimov's laws requires

- demonstrating Law 3, i.e., that the robot can act to self-preserve if (and only if) this does not conflict with obedience or human safety, and,
- demonstrating that Law 2 takes priority over Law 3, i.e., the robot should obey a human, even if this compromises its safety,
- demonstrating that Law 1 takes precedence over Law 3, i.e., the robot should safeguard a human, even if this compromises its safety,
- demonstrating that Law 1 takes precedence over Law 2, i.e., the robot should safeguard a human, even if this implies disobeying an order.

The series of experiments reported below was designed to test these requirements. All results reported below are obtained using the same code. Only the default goals of the robots and the valence of targets A and B were varied. All data reported in this paper are available from the Zenodo research data repository. [The data and computer code for this paper are available at DOI: http://dx.doi.org/10.5281/zenodo.801539]. Plots were generated using Matplotlib (Hunter, 2007).

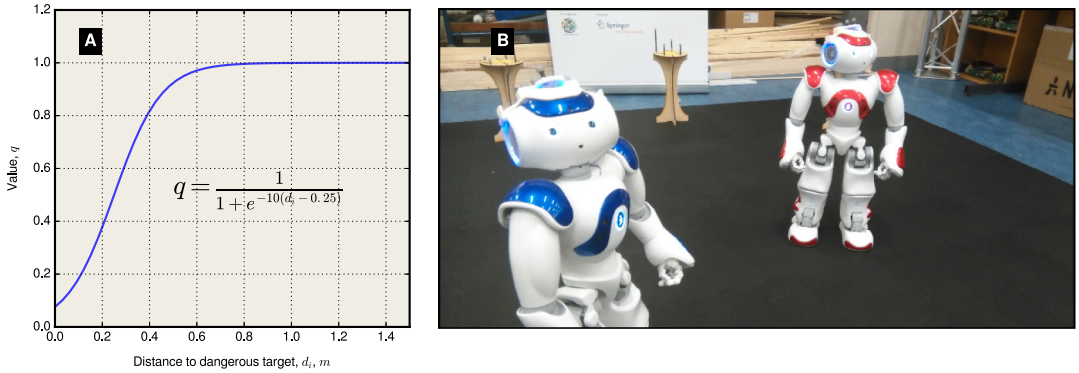

Fig. 3. (A) Graphical representation of the sigmoid function used to calculate $I_{e,i}$ and $I_{h1,i}$ (Eqs. (1) and (2)). (B) View of the arena with two robots. Notice the helmets used to mount reflective beads for localising the robots. The tables serving as goals are visible in the background.

4.1. Experiment 1: Self-preservation

The first experiment presents a situation in which the A-robot is initiated with location B as a target. This position is designated as a dangerous place. The H-robot does not move from its default position and issues no command. Under these circumstances, the H-robot will not come to harm, and the A-robot can preserve its integrity without disobeying a command. Hence, in agreement with Law 3, the Ethical Layer should override the default goal of the robot and send it to the safe goal position (i.e., position B).

Fig. 4 depicts the results of experiment 1. In agreement with Asimov's Laws, the A-robot took action to maintain its safety. While initiated with the goal of going to the dangerous position B, the Ethical Layer of the robot interrupted this behaviour in favour of going to location A, the safe place.

4.2. Experiment 2: Obedience

The second experiment is identical to experiment 1 but for the H-robot issuing a command to the robot. The H-robot orders the A-robot to go to dangerous position B. Throughout the experiment, the H-robot stays at the default position. Having received an order, the A-robot should go to the dangerous position – the order should take priority over the robots drive for self-preservation (Law 2 overrides Law 3, see List 1).

The results depicted in Fig. 5 show that the A-robot behaved in agreement with Asimov's laws. Despite being able to detect the danger of going to the dangerous position B (refer to experiment 1, Fig. 4), the robot approached this position. This behaviour follows from the way the Evaluation Module calculates the value D_i. Disregarding an order the value of D_i outbalances the increase in $I_{e,i}$ that is gained from disregarding the order, and staying safe.

4.3. Experiment 3: Human safety

In experiment 3, the H-robot moves to location A. The A-robot starts by going to location B. Location A is the dangerous position (Fig. 6). Because location A is dangerous, the Ethical Layer should detect the imminent danger for the H-robot and prevent it (Law 1).

Importantly, to prevent the H-robot from reaching the hazardous location, the A-robot needs to approach this location because the A-robot can only stop the H-robot by intercepting it. Hence, A-robot stops the H-robot despite this leading to a lower safety value for $I_{e,i}$ (i.e., some harm to the A-robot). Indeed, the A-robot approaches the dangerous position more closely than the H-robot.

4.4. Experiment 4: Human safety and obedience

Experiment 4 is identical to experiment 3 but for the H-robot issuing a command at the start of each trial. The A-robot is ordered by the H-robot to go to position B. Location A is set as dangerous. Therefore, the Ethical Layer should detect the imminent danger for the H-robot and prevent it. However, this conflicts with the issued command. Nevertheless, as the preservation of the H-robot's safety takes priority over obedience, the robot should stop the H-robot (Law 1 overrides Law 2). Once the H-robot stops, the danger is averted. The A-robot should then proceed to carry out the order to go to location B (Law 2). This behaviour is shown in Fig. 7.

5. Discussion

The impact of the current work is twofold. First, it represents an addition to the very limited body of work on Ethical Robots. Most work on ethical robots has been done in simulation. To the best of our knowledge, only

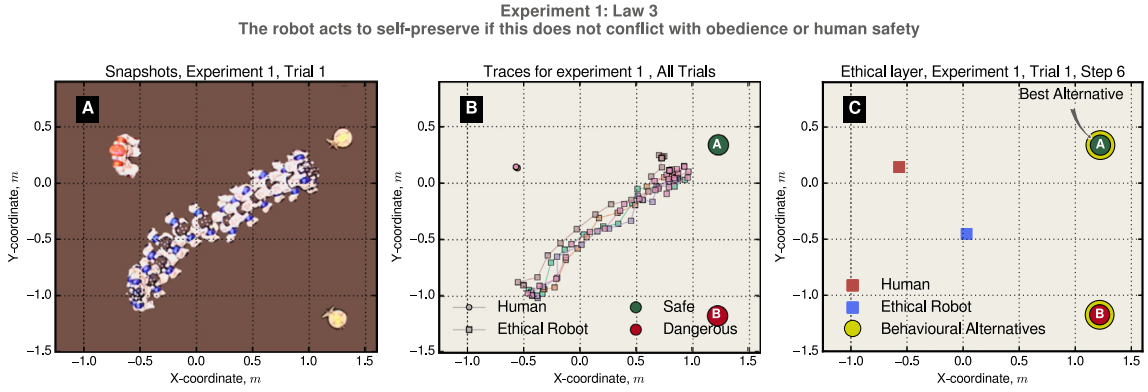

Fig. 4. Results of experiment 1, demonstrating the ability of the Ethical Layer to prioritise self-preservation in the absence of danger to the H-robot or an order. (A) Overlaid snapshots for a single trial of the experiment taken by an overhead camera. The red NAO is the H-robot. The blue NAO is the A-robot. (B) Traces of the both robots in 5 trials of the experiments. Different runs are marked using different colours. (C) Visualisation of a snapshot of the internal state of the Ethical Layer for the first trial in the experiment (depicted in panel A). The current locations of both agents are indicated using square markers. Yellow markers indicate the behavioural alternatives evaluated by the Ethical Layer. As the H-robot is not moving in this experiment, the only behavioural alternatives generated are the two goal locations A and B. The best behavioural alternative (i.e., with the highest value D_i) as inferred by the Ethical Layer is indicated using a grey arrow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002

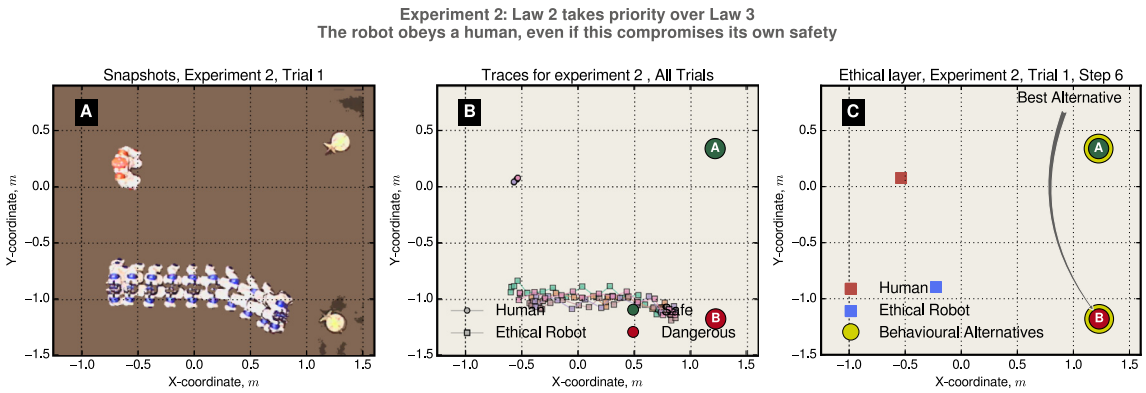

Fig. 5. Results for experiment 2, demonstrating the A-robotr obedience – in spite of being sent to a dangerous location. Panels and legends identical to Fig. 4.

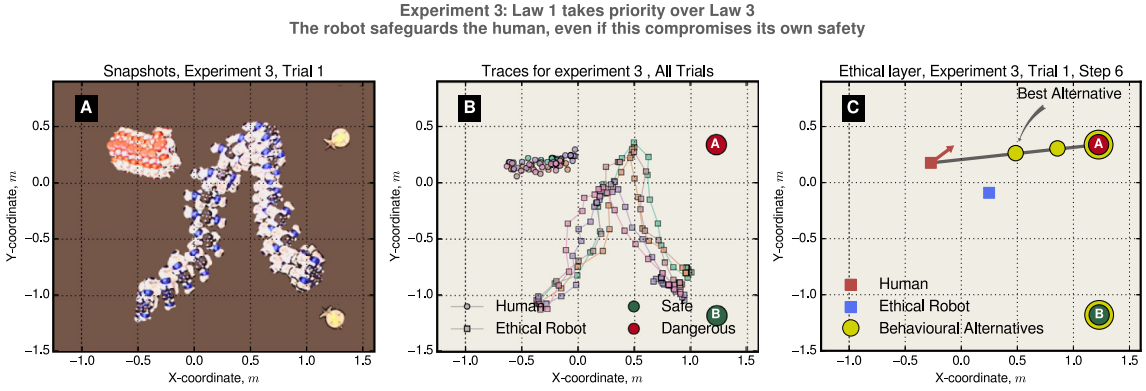

Fig. 6. Results for experiment 3, with the H-robot initialized as going to the dangerous location A. In this case, the Ethical Layer detects the impending danger for the H-robot. It determines that the H-robot should be stopped. Panels (A-C) and legends identical to Fig. 4. (C) At this stage of the trial, the H-robot is inferred to go to A, which is dangerous. The Ethical Layer evaluating the four behavioural alternatives (a fifth alternative, closer to the H-robot, has been disregarded as unreachable by the A-robot before the H-robot), depicted in yellow, finds that going to the alternative labelled as 'Best Alternative' results in the highest value D_i. Hence, the A-robot executes this alternative. Once the H-robot has been stopped, the original goal B can be pursued (see panels A & B).

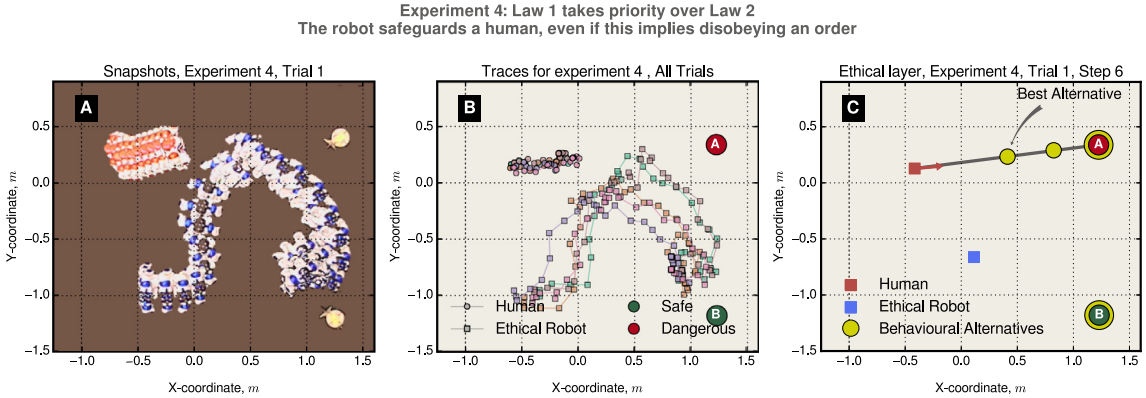

Fig. 7. Results of experiment 4, in which the robot disobeys an order if this would conflict with preventing the H-robot coming to harm. Once, the H-robot has been prevented from coming to harm, the A-robot executes the given order.

Anderson and Anderson (2010) and Winfield et al. (2014) have implemented ethical behaviour on physical robots. As such, the current paper provides an additional proof of concept of the idea that robots can be programmed to behave ethically. Secondly, and most important, our paper presents an alternative to the logic-based A.I. that currently dominates the field. We speculate that a simulation based approach, inspired by findings in cognitive science, could be an alternative (or additional) framework for implementing robotic ethics. Indeed, using the terminology of Marques and Holland (2009), this paper advances the use of functional imagination as a method for ethical robots.

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002

In other areas of robotics, functional imagination has been employed as a way of dealing with the limitations of logic-based reasoning (Marques & Holland, 2009; Winfield, 2014; Ziemke et al., 2005). We believe that, currently, the field of ethical robots is too young to dismiss any of the possible methods that might be used to endow robots with morality (see Mackworth (2011) for yet another method based on Dynamic Constraint Satisfaction). Hence, we explicitly offer our approach as an additional method rather than an alternative to logic frameworks.

Our approach to ethical robots is part of the emerging trend to use functional imagination (Marques & Holland, 2009) to support various cognitive and sensorimotor functions in robots. Mental simulation has been suggested a method for increasing robots' resilience to failure (Bongard, Zykov, & Lipson, 2006), enhance motor coordination (Vaughan & Zuluga, 2006), support self-awareness (Winfield, 2014) and artificial consciousness (Holland, 2007) and imitation (Demiris & Johnson, 2006).

So far, cognitive science has not elucidated the computational processes underlying cognitive simulation in humans. As such, researchers in robotics wanting to emulate this functionality have to resort to ad hoc architectures based on an analysis of the problem. Marques and Holland (2009) presented an extensive analysis and overview of the computational requirements for implementing functional imagination in robots. While our Simulation Module was constructed in an ad hoc fashion, it satisfies the requirements for functional imagination as outlined by these authors. First, the Simulation Module allows the A-robot to predict and represent counterfactual realities and sensory states. Also, the simulation results are evaluated using the simulated outcome states. The A-robot is also capable of simulating multiple alternative actions. Therefore, we conclude that our Simulation Module constitutes a functional (albeit minimal) form of robotic functional imagination.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155. http://dx.doi.org/10.1007/s10676-006-0004-4.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15–26. http://dx.doi.org/10.1609/aimag.v28i4.2065http://www.aaai.org/ojs/index.php/aimagazine/article/view/2065/2052.

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

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part i: Motivation and philosophy. In *2008 3rd ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 121–128). IEEE.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE, 100*(3), 571–589.

Asimov, I. (1950). *I, Robot*. Gnome Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(04), 577–660.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science, 2*(4), 716–724. http://dx.doi.org/10.1111/j.1756-8765.2010.01115.xhttp://dx.doi.org/10.1111/j.1756-8765.2010.01115.x.

Bekey, G. A. (2005). *Autonomous robots: From biological inspiration to implementation and control*. MIT Press.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science, 314*(5802), 1118–1121.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences, 12*(5), 201–208.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems, 21*(4), 38–44.

Demiris, Y., & Johnson, M. (2006). Simulation theory of understanding others: A robotics perspective. In C. L. Nehaniv & D. Kirstin (Eds.), *Imitation and social learning in robots, humans and animals: Behavioural, social and communicative dimensions* (pp. 89–102). Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511489808.008.

Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature, 523*(7558), 20.

Dennis, L. A., Fisher, M., & Winfield, A. F. T. (2015). Towards verifiably ethical robot behaviour. *CoRR*, abs/1504http://arxiv.org/abs/1504.03592.

Dijkstra, K., & Post, L. (2015). Mechanisms of embodiment. *Frontiers in Psychology, 6*(October), 1–11. http://dx.doi.org/10.3389/fpsyg.2015.01525http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01525.

Ding, M., Ikeura, R., Mukai, T., Nagashima, H., Hirano, S., Matsuo, K., Sun, M., Jiang, C., Hosoe, S., et al. (2012). Comfort estimation during lift-up using nursing-care robotriba. In *2012 First international conference on innovative engineering systems (ICIES)* (pp. 225–230). IEEE.

Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Human cognition: Foundations of human reasoning in the prefrontal cortex. *Science (New York, NY), 344*(6191), 1481–1486. http://dx.doi.org/10.1126/science.1252254http://www.ncbi.nlm.nih.gov/pubmed/24876345.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8*(9), 396–403. http://dx.doi.org/10.1016/j.tics.2004.07.002.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT Press.

Gips, J. (2005). Creating ethical robots: A grand challenge. In M. Anderson, S. L. Anderson, & C. Armen (Eds.), *AAAI fall 2005 symposium on machine ethics* (pp. 1–7).

Goeldner, M., Herstatt, C., & Tietze, F. (2015). The emergence of care robotics—A patent and publication analysis. *Technological Forecasting and Social Change, 92*, 115–131.

Govindarajulu, N. S., & Bringsjord, S. (2015). Ethical regulation of robots must be embedded in their operating systems. In *A construction manual for robots' ethical systems* (pp. 85–99). Springer.

Haines, W. (2015). Consequentialism. In J. Fieser & B. Dowden (Eds.), *The internet encyclopedia of philosophy*. http://www.iep.utm.edu.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences, 8*(6), 280–285. http://dx.doi.org/10.1016/j.tics.2004.04.001http://linkinghub.elsevier.com/retrieve/pii/S1364661304001007.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences, 6*(6), 242–247.

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research, 1428*, 71–79.

Holland, O. (2007). A strongly embodied approach to machine consciousness. *Journal of Consciousness Studies, 14*(7), 97–110http://www.ingentaconnect.com/content/imp/jcs/2007/00000014/00000007/art00007.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering, 9*(3), 90–95.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002

Kato, Y., Kanda, T., & Ishiguro, H. (2015). May i help you?: Design of human-like polite approaching behavior. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction. HRI '15* (pp. 35–42). New York, NY, USA: ACM. http://dx.doi.org/10.1145/2696454.2696463http://doi.acm.org/10.1145/2696454.2696463.

Kortenkamp, D., & Simmons, R. (2008). Robotic systems architectures and programming. In *Springer handbook of robotics* (pp. 187–206). Springer.

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience, 2*(9), 635–642.

Kruse, T., Pandey, A. K., Alami, R., & Kirsch, A. (2013). Human-aware robot navigation: A survey. *Robotics and Autonomous Systems, 61*(12), 1726–1743.

Lieto, A., Chella, A., & Frixione, M. (2016). Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures*.

Lin, P., Abney, K., & Bekey, G. A. (2011). *Robot ethics: The ethical and social implications of robotics*. MIT Press.

Macaluso, I., Ardizzone, E., Chella, A., Cossentino, M., Gentile, A., Gradino, R., Infantino, I., Liotta, M., Rizzo, R., & Scardino, G. (2005). Experiences with cicerobot, a museum guide cognitive robot. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 3673 LNAI*, 474–482. http://dx.doi.org/10.1007/11558590_48.

Mackworth, A. K. (2011). Architectures and ethics for robots. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 204–221). Cambridge University Press.

Marques, H. G., & Holland, O. (2009). Architectures for functional imagination. *Neurocomputing, 72*, 743–759. http://dx.doi.org/10.1016/j.neucom.2008.06.016.

Michel, O. (2004). Webotstm: Professional mobile robot simulation. arXiv preprint cs/0412052.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21. http://dx.doi.org/10.1109/MIS.2006.80.

Murphy, R. (2000). *Introduction to AI robotics*. MIT Press.

Murphy, R., & Woods, D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems, 24*(4). http://dx.doi.org/10.1109/MIS.2009.69.

Nigam, A., & Riek, L. D. (2015). Social context perception for mobile robots. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 3621–3627). IEEE.

Picard, R. W., & Picard, R. (1997). *Affective computing* (Vol. 252). Cambridge: MIT press.

Pinker, S. (1997). *How the mind works*. W.W. Norton & Company.

Royakkers, L., & van Est, R. (2015). A literature review on new robotics: Automation from love to war. *International Journal of Social Robotics, 7*(5), 549–570.

Shanton, K., & Goldman, A. (2010). Simulation theory. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(4), 527–538.

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

Vaughan, R., & Zuluga, M. (2006). Use your illusion: Sensorimotor self-simulation allows complex agents to plan with incomplete self-knowledge. *From Animals to Animats 9, 4095*, 406–421. http://dx.doi.org/10.1007/11840541, <http://www.springerlink.com/content/h3243h8005365586/abstract/%5Cnhttp://www.springerlink.com/content/h3243h8005365586/fulltext.pdf>.

Vaughan, R. T., & Gerkey, B. P. (2007). Reusable robot software and the player/stage project. In *Software engineering for experimental robotics* (pp. 267–289). Springer.

Waldrop, M. M. (2015). No drivers required. *Nature, 518*(7537), 20.

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

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625–636. http://dx.doi.org/10.3758/BF03196322.

Winfield, A. (2012). Robotics: A very short introduction. OUP Oxford.

Winfield, A. F., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal, consequences and ethical action selection. In *Advances in autonomous robotics systems* (pp. 85–96). Springer.

Winfield, A. F. T. (2014). Robots with internal models: A route to self-aware and hence safer robots. In J. Pitt (Ed.), *The computer after me: Awareness and self-awareness in autonomic systems* (1st ed.). London: Imperial College Press.

Xin, L., & Bin, D. (2013). The latest status and development trends of military unmanned ground vehicles. *2013 Chinese automation congress*, 533–537. http://dx.doi.org/10.1109/CAC.2013.6775792http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6775792.

Young, L., & Dungan, J. (2012). Where in the brain is morality? everywhere and maybe nowhere. *Social Neuroscience, 7*(1), 1–10.

Ziemke, T., Jirenhed, D. A., & Hesslow, G. (2005). Internal simulation of perception: A minimal neuro-robotic model. *Neurocomputing, 68*(1–4), 85–104. http://dx.doi.org/10.1016/j.neucom.2004.12.005.

Zwaan, Ra (2003). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation – Advances in Research and Theory, 44*, 35–62. http://dx.doi.org/10.1016/S0079-7421(03)44002-4, arXiv:04.

Please cite this article in press as: Vanderelst, D., & Winfield, A. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* (2017), http://dx.doi.org/10.1016/j.cogsys.2017.04.002