Sexbots as Ethical Agents: On the Possibility of Ethical Machines

Christopher J. Headleand[12] and **Llyr Ap Cenydd**[3] and **William J. Teahan**[4]

Abstract. The development of robots as synthetic sexual partners is associated with a number of risks. This has led to individuals calling for the outright ban on their development, or legislation regarding their use. In this paper, we approach this argument from a complementary perspective, and ask whether sexbots should be developed as artificial ethical agents. We open by exploring the concepts of morality in non-human agents, before exploring machine morality in detail. We explore some of the key practical and philosophical objections to the development of moral machines. We then explore definitions for both Artificial Moral Agents (AMAs) and Artificial Ethical Agents (AEAs), establishing that while moral machines may not be possible at this time, ethical machines are. This is followed by a brief overview of the current work in the development of Artificial Ethical Agents. The paper concludes by arguing for further research in this area.

1 Introduction

The concept of 'sexbots' and artificial sexuality has recently awoken discussions over the ethics of robotics. This has included calls to ban, or legislate the development of these anthropomorphic adult products [47]. However, it is not our intention to enter into the debate on the ethics of their use. In this paper, we will discuss the possibility of sexbots being developed as ethical agents in their own right.

One of the principal concerns about these devices is that they could lead to the further objectification, and dehumanisation of sexual relations. This was illustrated particularly powerfully in the second episode of the TV series *Humans* [59]. In this episode, the female 'synth' called Niska had assumed the role of a prostitute in a robot brothel. This culminated in a dramatic scene when a 'client' asked her to act *young and scared*, a request she took umbrage to. While this particular fictional encounter ended violently, it does raise an interesting question, namely, "should sex robots have an understanding of right and wrong?", and importantly should they be able to recognise unethical behaviour. If we conclude that they should, then we must also ask whether imbuing a robot with the ability to make ethical judgements is possible?

This paper seeks to explore this question through a review of the literature on the subject. We start by looking at an argument for morality in non-human animals, in order to introduce the idea that humans may not be the only moral (or ethical) agents. We will then describe the current literature in Artificial Moral Agents, including an overview of four key philosophical objections. We then move to define a difference between Artificial Moral Agents (AMAs) and Artificial Ethical Agents (AEAs), which we will provide clear definitions for, before discussing the practical implementation of Ethical Agents.

In the following sections the words 'agent' and 'moral agent' will be used frequently. For clarity, an agent is defined as:

> *an entity with the capacity to act in a given environment,*

a moral agent is defined as:

> *an entity with the capacity to act with reference to right and wrong.*

2 Morality in non-Human Animals

When considering broad concepts such as morality and ethics in machines, it is prudent to consider the debate regarding other non-human entities.

Typically, when attempting to describe animal behaviour as academics, we do not assign them with any more qualities than we absolutely have to [48]. It was once taboo to use the words *animal* and *cognition* within the same sentence [23] in any academic context and philosophers have traditionally denied the possibilities of non-human animals being moral agents [16].

An example of this traditional way of thinking is provided by Himma [35], who describes a dog which attacks someone wearing red because that is how it is trained. Although the dog was the direct cause of its behaviour (in the sense that its mental state resulted in its actions), Himma argues that "it has not freely chosen its behaviour *because dogs do not make decisions in the relevant sense* (emphasis added)". As such, the moral responsibility (and by extension, the agency) remains with the person who trained it.

Rowlands[49] also considers the concept of responsibility[5] as being central to the issue of agency, which he defines as:

> *X is a moral agent if and only if X is (a) morally responsible for, and so can be (b) morally evaluated (praised or blamed, broadly understood) for, its motives and actions.*

Arguably, it is probably mistaken to classify animals as full moral agents, but even with that being considered, they could be moral subjects.

> *X is a moral subject if and only if X is, at least sometimes, motivated to act by moral considerations.*

[1] Bangor University, United Kingdom, email: chris@chrisheadleand.com
[2] University of Lincoln, United Kingdom, email: cheadleand@lincoln.ac.uk
[3] Bangor University, United Kingdom, email: llyr.ap.cenydd@bangor.ac.uk
[4] Bangor University, United Kingdom, email: w.j.teahan@bangor.ac.uk

[5] Importantly, the definitions focus on responsibility, rather than accountability [11].

Additionally, some philosophers have questioned the traditional way of thinking, renewing the debate on animal morality. It has been argued [38] that the behaviours of humans and non-human primates point towards a shared evolutionary background towards morality. For example, non-human primates have demonstrated similar methods to humans for preventing and resolving conflicts, which have been described as the building blocks of moral systems [27]. One of these traits is empathy, which has often been revered as a human trait; however, a significant body of research now suggests that other animals exhibit this phenomenon [25].

There are also various examples in the literature where behaviour has been described that if it had been attributed to humans would have been assumed to have been moral (for examples see [26, 15, 46], and note that this is a relatively small sample of the accounts available). This forces us to question if it is an unfair standard to not credit an animal with any greater quality than we absolutely have to.

Coeckelbergh [18] notes that in the study of human morality we tend to rely on observation. We argue that it would be more consistent if we applied the same standards to the assessment of other entities. By assuming an observational standpoint (even an anthropometric one), we open up new opportunities in the study of behaviour, by comparing our behaviour with that of primates (such as the work by Frans de Waal [24]) or through simulation.

3 Artificial Moral Agents

There are many reasons why we may wish to build Artificial Moral Agents. Firstly, building agents with a moral capacity may help to avert dangerous or unethical behaviour (such as the prophesied AI Armageddon [22]). An alternate position is that, if we accept that ethics is a cognitive pursuit, it is possible and logical to assume that a superintelligent AI could make ethical decisions better than humans [12]. Also, robots as simple models of agency could be used to help understand more complex cases of human ethical judgements [21].

More importantly, society is becoming increasingly driven by technology. There are now many examples of tasks that would have originally required a human that have now been delegated to machines and algorithms. This handover of control in our society has in many cases (such as banking) placed computers in situations where they can affect the moral rights of humans [51]. There is even discussion over whether algorithms themselves are value laden, and should be considered from an ethical position [39].

One of the central questions is whether an artificial entity can ever be a moral agent [35, 36, 51, 54]? This has led to a range of debates regarding how we should be viewing robots and other autonomous agents. For example, the increasing use of robotics in the military arena has led to a growing body of research on the ethics of autonomous military machines (so-called 'killer robots') and a plethora of literature on the subject [19, 33, 40, 41, 43, 57]. There is also significant fear that the development of artificial intelligence could be dangerous, so much so that it could be incompatible with human life.

Even the word 'robot' comes from a dystopian fantasy, coined as the name for artificial workers who rise up and overthrow their human masters [6]. This Frankenstein complex, that the things we create will ultimately destroy us, is persistent (arguably re-enforced by numerous works of science fiction), and if nothing else has forced us to evaluate the moral implications, and responsibilities of machines. This theme plays a central role in the works of Isaac Asimov, best highlighted in the 'Three Laws' of robotics [8, 42][7].

However, there is an argument that machines could be preferential to humans in some scenarios. For example, ethical standards are greatly challenged on the battlefield [56], with the best that can be done is to provide soldiers with rules of engagement and hope they do not break them. It is possible that the situation is improved by taking humans out of the loop, leaving an opportunity for roboticists to design systems that could potentially do better [55].

It has been stated that the primary goal of the study of artificial morality is the design of agents that act as if they are moral[1], although an important distinction to make is the difference between acting moral, and being a genuine moral agent. If we were able to produce agents that behaved according to moral guidelines then maybe some of our doomsday fears would be alleviated. At the very least, humans may be more comfortable collaborating with machines that follow basic principles of right and wrong.

But is artificial morality even possible? Sullins[53] argues that it is "absolutely certain" that under certain conditions, artificial life (ALife) programs can exhibit artificial morality. Sullins also states that the study of 'Wet ALife' (the use of bio-components in artificial life research) could result in the fields of bio and computational ethics sharing concerns.

However, there are many arguments which oppose this position on the grounds that synthetic moral agency is either impossible or undesirable. These four objections are specifically focused on artificial morality, and do not include the wider objections to true AI. This is not intended to be a complete survey of the arguments against AMAs, but an informative sample.

The Information Processing Objection: [51] argues that *information* is a core requirement of moral agency which precludes computers from achieving it. The argument is that computers in their current form are processors of data, rather than processors of information.

Frankenstein Objection: If we believe that moral reasoning requires consciousness, and the ability to exercise emotions, then producing an AMA in the foreseeable future is unlikely. If this is the case, then it would be better to avoid building highly intelligent autonomous agents as their lack of morality may make them dangerous [18, 28]. This concept of a robotic uprising is often based on the fear of machines without morality [42]. Conversely, building true moral agents may never be desirable. A genuine moral agent must be capable of recognizing and acting *immorally* [10], and such an entity that simply simulates moral behaviour may be preferable to genuine moral agent.

The Scapegoat Objection: If machines are considered to be true moral agents, then they would be morally accountable. One fear is that people may begin to use machines to avoid personal responsibility [29, 33, 37, 51] essentially blaming machines for their actions. Similarly there is a fear that decision makers may rely on machines too much, circumventing individual responsibility entirely [60].

The Free Will Objection: According to Kant, moral agency requires both rationality and personal freedom. Supporters of this argument state that computer programs do not have free will, and

[6] The play in question is R.U.R (Rossum's Universal Robots) written by Czech author Karel Capek [14]. The word *Robot* is derived from the Czech word 'robota' which translates as servitude.

[7] The Three Laws are: (i) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (ii) A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law. (iii) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

thus they can never be independent moral agents [35, 36, 58]. Another interpretation of the free will objection is the 'Mousetrap Objection': if an intelligent agent is capable of closing a loop between a sensor and effector without human intervention [50], then by extension a mousetrap is an intelligent agent. However, a mousetrap is not morally responsible (and as such, not a moral agent) because it closes its loop entirely at the volition of one or more humans. Responsibility for the actions of the mousetrap remains with the humans who armed it. "If an agent is to be morally praiseworthy, then its rules for behaviour and the mechanisms for supplying those rules must not be supplied entirely by external humans"[34].

These arguments reinforce the reasons why we have typically denied the possibility that a non-human entity could be a moral agent, something which has traditionally been reserved exclusively for humans [17, 48, 51].

Another issue with the development of moral machines is the actual practicalities of development. Building an Artificial Moral Agent is by its nature a practical goal. It is a distinctly different objective to that of moral philosophers, and as such determining a clear specification of what constitutes 'moral' from a technical perspective can be a challenge, as a comprehensive model of moral decision making does not yet exist [61].

This is well summarized by Allen et al. [2] who notes that the development of Artificial Moral Agents (AMAs) is hindered by two areas of disagreement from theorists. The first is that philosophers disagree about what behavioural standards constitute morality. The other is ontological, a question of what it means to be moral. For example, an act which is ethical from an egoist's perspective may conflict or be incompatible with the act a utilitarian would select given the same circumstance.

Until these practical and philosophical issues are resolved, we propose that research should focus instead on Artificial Ethical Agents. This change in focus allows us to sidestep the philosophical objections entirely, by designing agents that *simulate* moral behaviour. As previously mentioned, an agent that simply exhibits moral behaviour may (in some ways) be preferential to a genuine moral agent anyway.

4 Defining Artificial Ethical and Moral Agents

In this section, we will define differences between Artificial Moral Agents (AMAs) and Artificial Ethical Agents (AEAs). This is for two reasons: (i) the terms are often used interchangeably, without clear definition; and (ii) research and theories in one of these domains does not automatically presume application to the other. We do not wish to argue that the AEA theories are applicable to the larger questions of morality, and moral agency.

The Oxford English dictionary defines morals as:

> *"The principles of right and wrong behaviour."*

and Ethics as:

> *"A set of moral principles, especially ones relating to or affirming a specified group, field, or form of conduct."*

We can interpret these definitions to mean that *morals* are the core concepts which allow us to define the difference between right and wrong. By contrast *ethics* relates to how these rules are applied. With this understanding, and the insights provided by [49] on Moral agency and moral subjecthood (mentioned in section 2), we can define an Artificial Moral Agent (AMA) as:

> *An autonomous entity that is capable of making, and being held accountable for, unsupervised decisions with reference to an understanding of right and wrong.*

Whereas we would define an Artificial Ethical Agent (AEA) as:

> *An autonomous entity that is capable of acting according to a set of morally defined considerations.*

The principal difference between these two is that the Artificial Moral Agent must be able to derive its own concepts and understanding of what constitutes moral behaviour. Consequently, it must also be accountable for its actions. On the other hand, the principles (or rules) for an Artificial Ethical Agent may come from an external source.

An alternate definition of Ethical Agents is provided by Moor [44]. Moor's taxonomy specifically defines four different types of ethical agent:

Ethical Impact Agents are any agents whose actions could have ethical consequences. Almost any agent has the potential to be an Ethical Impact Agent if it has the potential to cause harm (or benefit) to humans.

Implicit Ethical Agents are agents that are designed with ethical considerations in mind. These generally refer to safety considerations, including warning systems.

Explicit Ethical Agents are agents that can identify a variety of ethical considerations, and process situational information to determine a course of action.

Full Ethical Agents are similar to Explicit ethical agents but also have metaphysical features that we typically attribute to human ethical agents, including free will.

We would argue that the definitions for Ethical Impact Agents and Implicit Ethical Agents are too broad to be practically useful, although they do serve to highlight the scope of the ethical nature of machines. Furthermore, the difference between Explicit Ethical Agents and Full Ethical Agents is purely philosophical. There is no evidence that a 'Full Ethical Agent' would be substantially different to a 'Explicit Ethical Agent', especially considering that neuroscientists are beginning to question the concept of metaphysical qualities such as 'free will' [20]. Regardless of whether qualities such as free will exist or not, unless we can provide a model of what defines them, their inclusion in a definition is not helpful from an engineering perspective.

That considered, our definition purposefully avoids this philosophical distinction. Instead, the central concept of our two definitions is a matter of responsibility, specifically, can a machine be ever considered accountable for its actions [52], and can moral responsibility ever be transferred from a human (the designer or operator) to the machine [31]? If yes, then the machine is an AMA; if not, but the machine still operated within ethical considerations, then it is defined as an AEA.

Before an AMA (according to our definition) could be developed (or recognized) as a Moral Agent, there is an abundance of legal and technological hurdles which must first be addressed. Considering the debate over morality in non-human animals (biological entities), it is unlikely these issues will be resolved conclusively for synthetic agents in the near future. An unfortunate reality of this situation is that even if we were to develop a genuine moral agent, we might not recognize it as such, due to our inability to agree on a standard criteria.

We would argue that an AMA is not possible within the current technological generation, as many core AI issues must first be solved. Also, as detailed in the previous section, there are many who would argue that an AMA will never be possible.

However, as the rules for an Artificial Ethical Agent can come from an external source, this makes their development a practical engineering problem immune to the objections defined in the previous section. Agents that would pass our definition of an AEA are not only possible, but have already become a reality. In the following subsection, we will discuss some AEA models defined in the literature, and discuss current applications.

5 Practical Artificial Ethical Agents

Arkina [6] describes three possible ways that ethical decision making could be implemented into an autonomous agent. The first system referred is to as a *Governor*, which would halt the agent from proceeding with actions deemed unethical. This is analagous to the governors that are release valves which prevent steam engines from running too hot or too fast. The second is *Behaviour Control*, which monitors the behaviours a robot is engaged with and ensures that actions fall within a set of constraints. The final system is the *Adapter*, which modifies the first two systems if somehow an unethical action occurred despite their intervention [7].

Winfield et al. [65] describe an alternative approach referred to as a consequence engine. In their approach, a simulator is embedded within a robot, providing the agent with a pseudo-imagination that allows it to try actions before executing them in the real world. Through this process, the robot would be able to find the sequence of actions which best achieves its goal, and 'ponder' hypothetical situations which may arise. The authors of this work argue that this capacity can be applied to ethical decision making. For example, consider a robot that has rules which prevent it from colliding with a human to avoid causing injury. The robot observes a human about to step into the path of a car. Its internal simulator determines that the action that causes the least harm to the human is for the robot to collide with the human, preventing them being hit by the vehicle[8]. The Winfield 'What-if' engine is shown to work on simple situated robots in a limited world, with one robot changing its behaviour to prevent another from being harmed.

A similar model is proposed by [9] for a general purpose ethical engine. They prescribe a number of requirements for ethical robotics (best summarized by [55]) which are:

1. estimate with high detail and accuracy the immediate state of the world around the agent;
2. predict the likely future states given the current possible candidate actions.

The problem with this model is that it could be used to describe any decision making process in AI. As such it falls prey to some of the common problems surrounding the field of artificial intelligence, such as estimating the state of the world in dynamic environments, and the computational complexity in evaluating every possible future state. We could also argue that current algorithms exist which meet this criteria (for example, MiniMax [50]), but only work in worlds of constrained complexity with a finite number of possible actions.

Early attempts at building machines capable of ethical reasoning have focused on decision support systems [60]. Anderson argues that the best way to start tackling the challenge of ethical decision making is to build machines that act as advisers to humans, in a select community in a finite number of circumstances [3]. To use Moor's definitions, this would involve creating an explicit ethical agent that is not autonomous. This approach also allows the community to postpone the philosophical debate regarding the moral status of machines.

Anderson et al. [5] developed MedEthEx, the earliest example of a prototype artificial ethical adviser based on Bio-medical Ethical principles. In their approach, machine learning techniques are used to abstract decision principles from a library of cases with conflicting prima facie duties (duties that suggest conflicting courses of action in specific ethical dilemmas). This is trained based on a 'correct action' determined by an agreement of ethicists. This is in an attempt to capture the complexities of ethical decision making, and to codify a decision process for determining the ethically correct course of action when conflicts between duties, values or principles arise. They argue their their method is a "useful first step" towards an artificial ethical agent that behaves in a certifiable ethical way.

Another example is euro-transplant, a software tool which generates priority lists of organ recipients based on various factors (age, waiting time, distance between donor and recipient) and must follow the medical ethical criteria [31]. What makes this example particularly interesting is that it is generally believed that the software is capable of making these judgements better than previous (human controlled) systems [25]. There are also systems in use which use machine learning to resolve biomedical ethical dilemmas [4].

The US Army have also funded research into autonomous ethical advisers, using a utility system referred to as the 'Metric of Evil' [45]. The authors note that the intention of this system is to generate results that resemble human decision making, rather than attempt to replicate the human moral reasoning. The metric of evil is calculated by adding together an evil value for each consequence of a specific action. The weightings for each consequence are defined by a panel of experts based on a series of test cases. Guarini [30] also explored this area, and trained a neural network on a number of cases regarding the acceptability of killing in certain situations (such as self-defence). After training, the system was capable of providing acceptable responses to a variety of new cases.

While the majority of this work has been examples of top-down systems, an alternate route which could be considered is the bottom-up approach (Wallach et al. [64] provide an excellent discussion on the subject). In our research, we have taken inspiration from studies of cognition, and built ethical value-systems based on Braitenberg Vehicles [13], demonstrating that the bottom-up, emergent approach can be a viable model [32]. Reactive approaches such as these have the advantage of being fast enough to operate in real time without requiring an internal model of the world. This is a particularly important consideration when working in dynamic environments. Wallach and Allen even argue that the Alife field has contributions to make to research into artificial moral agents, by "helping to understand the bottom-up emergence of dynamic and flexible moral behaviour" [62]. However, although Wallach [64] provides an excellent discussion on the subject, we are not aware of many practical examples of true bottom-up, or behaviour based AEAs. Furthermore, in their seminal book, *Moral Machines: Teaching Robots Right from Wrong*, Wallach and Allen conclude that a hybrid approach between top-down and bottom-up may be necessary [63].

In reality, the practical research in this area is in its infancy, and there is no clearly accepted route. At this stage in the field, all avenues will need to be considered, and this will undoubtedly require close collaboration between the fields of moral philosophy and arti-

[8] The authors [65] note that this example is remarkably similar to the first law of robotics described by Asimov: "A robot may not injure a human being or, through inaction, allow a human being to come to harm"

ficial intelligence.

6 Discussion

By imbuing a sex robot with the ability to identify the difference between ethical and unethical behaviour, we may overcome some of the concerns over their ongoing development. However, research into artificial ethics is severely outpaced by the speed of mechanical development in this area. At the time of writing, the majority of research into artificial ethical agents has been constrained to 'toy problems'. While there have been a number of promising studies, the research needs to progress significantly. Furthermore a number of technical challenges need to be addressed before a real world application in artificial sexuality can be considered.

Providing robots with the ability to make ethical judgements may alleviate some of societies' growing fear regarding their use. However, what an artificial ethical agent is, whether it is possible and how we should build them, remain open questions. This paper provides a short review of the literature surrounding artificial ethical agents and seeks to address these questions. We began by first introducing some arguments for moral agency in non humans, before discussing Artificial Moral Agents specifically. Through this we described a number of the key practical and philosophical objections to artificial moral agency. With these objections considered, we explored definitions for Artificial Moral Agents, and Artificial Ethical Agents, and concluded that while AMAs may not be possible, AEAs certainly are. Furthermore, according to an alternative definition provided by Moor [44], AEAs are already a reality (in some limited sense). In our penultimate section, we discussed some practical approaches towards the development of AEAs, and concluded that while the majority of research in this area has been top-down, bottom-up approaches warrant further exploration.

REFERENCES

[1] Colin Allen, Iva Smit, and Wendell Wallach, 'Artificial morality: Top-down, bottom-up, and hybrid approaches', *Ethics and Information Technology*, **7**(3), 149–155, (2005).

[2] Colin Allen, Gary Varner, and Jason Zinser, 'Prolegomena to any future artificial moral agent', *Journal of Experimental & Theoretical Artificial Intelligence*, **12**(3), 251–261, (2000).

[3] Michael Anderson and Susan Leigh Anderson, 'The status of machine ethics: a report from the aaai symposium', *Minds and Machines*, **17**(1), 1–10, (2007).

[4] Michael Anderson, Susan Leigh Anderson, and Chris Armen, 'An approach to computing ethics', *Intelligent Systems, IEEE*, **21**(4), 56–63, (2006).

[5] Michael Anderson, Susan Leigh Anderson, and Chris Armen, 'Medethex: a prototype medical ethics advisor', in *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 21, p. 1759. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, (2006).

[6] Ronald C Arkina, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture', in *Artificial General Intelligence, 2008: Proceedings of the First AGI Conference*, volume 171, p. 51. IOS Press, (2008).

[7] Ronald C Arkina, *Governing lethal behavior in autonomous robots*, Taylor and Francis, 2009.

[8] Isaac Asimov, 'Runaround', *Astounding Science Fiction*, **29**(1), 94–103, (1942).

[9] Yoseph Bar-Cohen and David Hanson, *The coming robot revolution: Expectations and fears about emerging intelligent, humanlike machines*, Springer Science & Business Media, 2009.

[10] Anthony F Beavers, 'Between angels and animals: The question of robot ethics, or is Kantian moral agency desirable', in *Annual meeting of the association of practical and professional ethics, Cincinnati, OH*. Association for Practical and Professional Ethics, (2009).

[11] Thomas H Bivins, 'Responsibility and accountability', *Ethics in public relations: Responsible advocacy*, 19–38, (2006).

[12] Nick Bostrom, 'Ethical issues in advanced artificial intelligence', *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277–286, (2003).

[13] Valentino Braitenberg. Vehicles, 1984.

[14] Karel Čapek, *RUR (Rossum's Universal Robots)*, Penguin, 2004.

[15] Russell M Church, 'Emotional reactions of rats to the pain of others.', *Journal of comparative and physiological psychology*, **52**(2), 132, (1959).

[16] Grace Clement, 'Animals and moral agency: The recent debate and its implications', *Journal of Animal Ethics*, **3**(1), 1–14, (2013).

[17] Mark Coeckelbergh, 'Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents', *AI & society*, **24**(2), 181–189, (2009).

[18] Mark Coeckelbergh, 'Moral appearances: emotions, robots, and human morality', *Ethics and Information Technology*, **12**(3), 235–241, (2010).

[19] Mark Coeckelbergh, 'Drones, information technology, and distance: mapping the moral epistemology of remote fighting', *Ethics and information technology*, **15**(2), 87–98, (2013).

[20] Jerry A Coyne, 'You dont have free will', *The Chronicle of Higher Education*, **18**, (2012).

[21] Peter Danielson, 'Designing a machine to learn about the ethics of robotics: the n-reasons platform', *Ethics and information technology*, **12**(3), 251–261, (2010).

[22] Hugo De Garis, *The Artilect War: Cosmists vs. Terrans: A Bitter Controversy Concerning Whether Humanity Should Build Goldlike Massively Intelligent Machines*, ECT Publications, 2005.

[23] Etienne De Sevin and Daniel Thalmann, 'An affective model of action selection for virtual humans', in *Proceedings of Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action symposium at the Artificial Intelligence and Social Behaviors 2005 Conference (AISB" 05)*, number VRLAB-CONF-2005-014, (2005).

[24] Frans de Waal, 'The animal roots of human morality', *New Scientist*, **192**(2573), 60–61, (2006).

[25] Frans BM de Waal, 'Do animals feel empathy?', *Scientific American Mind*, **18**(6), 28–35, (2007).

[26] Iain Douglas-Hamilton, Shivani Bhalla, George Wittemyer, and Fritz Vollrath, 'Behavioural reactions of elephants towards a dying and deceased matriarch', *Applied Animal Behaviour Science*, **100**(1), 87–102, (2006).

[27] Jessica C Flack and Frans BM de Waal, 'any animal whatever'. darwinian building blocks of morality in monkeys and apes', *Journal of Consciousness Studies*, **7**(1-2), 1–29, (2000).

[28] Luciano Floridi and Jeff W Sanders, 'Artificial evil and the foundation of computer ethics', *Ethics and Information Technology*, **3**(1), 55–66, (2001).

[29] Keith Grint and Steve Woolgar, *The machine at work: Technology, work and organization*, John Wiley & Sons, 2013.

[30] Marcello Guarini, 'Particularism and generalism: how ai can help us to better understand moral cognition', in *Machine ethics: Papers from the 2005 AAAI fall symposium*, (2005).

[31] F Allan Hanson, 'Beyond the skin bag: on the moral responsibility of extended agencies', *Ethics and information technology*, **11**(1), 91–99, (2009).

[32] Christopher J Headleand, Llyr Ap Cynedd, and William J Teahan, 'Towards ethical robots:revisiting braitenberg's vehicles', in *Science and Information Conference SAI, 2016*. IEEE, (2016).

[33] Thomas Hellström, 'On the moral responsibility of military robots', *Ethics and information technology*, **15**(2), 99–107, (2013).

[34] Patrick Chisan Hew, 'Artificial moral agents are infeasible with foreseeable technologies', *Ethics and Information Technology*, **16**(3), 197–206, (2014).

[35] Kenneth Einar Himma, 'Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?', *Ethics and Information Technology*, **11**(1), 19–29, (2009).

[36] Deborah G Johnson, 'Computer systems: Moral entities but not moral agents', *Ethics and information technology*, **8**(4), 195–204, (2006).

[37] Deborah G Johnson and Keith W Miller, 'Un-making artificial moral agents', *Ethics and Information Technology*, **10**(2-3), 123–133, (2008).

[38] Melanie Killen and Frans BM de Waal, 'The evolution and development of morality', *Natural conflict resolution*, 352–372, (2000).

[39] Felicitas Kraemer, Kees van Overveld, and Martin Peterson, 'Is there

an ethics of algorithms?', *Ethics and Information Technology*, **13**(3), 251–260, (2011).
[40] Armin Krishnan, *Killer robots: legality and ethicality of autonomous weapons*, Ashgate Publishing, Ltd., 2009.
[41] Pawel Lichocki, Aude Billard, and Peter H Kahn, 'The ethical landscape of robotics', *Robotics & Automation Magazine, IEEE*, **18**(1), 39–50, (2011).
[42] Lee McCauley, 'Ai armageddon and the three laws of robotics', *Ethics and Information Technology*, **9**(2), 153–164, (2007).
[43] Jeff McMahan and Bradley Jay Strawser, *Killing by remote control: the ethics of an unmanned military*, Oxford University Press, 2013.
[44] James H Moor, 'The nature, importance, and difficulty of machine ethics', *Intelligent Systems, IEEE*, **21**(4), 18–21, (2006).
[45] Gregory S Reed and Nicholaos Jones, 'Toward modeling and automating ethical decision making: design, implementation, limitations, and responsibilities', *Topoi*, **32**(2), 237–250, (2013).
[46] George E Rice and Priscilla Gainer, '"altruism" in the albino rat.', *Journal of comparative and physiological psychology*, **55**(1), 123, (1962).
[47] Kathleen Richardson, 'The asymmetrical'relationship': parallels between prostitution and the development of sex robots', *ACM SIGCAS Computers and Society*, **45**(3), 290–293, (2016).
[48] Mark Rowlands, *Can animals be moral?*, Oxford University Press, 2012.
[49] Mark Rowlands, 'Animals and moral motivation: A response to clement', *Journal of Animal Ethics*, **3**(1), (2013).
[50] Stuart Russell, Peter Norvig, and Artificial Intelligence, 'Artificial intelligence: A modern approach', *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, **25**, (1995).
[51] Bernd Carsten Stahl, 'Information, ethics, and computers: The problem of autonomous moral agents', *Minds and Machines*, **14**(1), 67–83, (2004).
[52] Bernd Carsten Stahl, 'Responsible computers? a case for ascribing quasi-responsibility to computers independent of personhood or agency', *Ethics and Information Technology*, **8**(4), 205–213, (2006).
[53] John P Sullins, 'Ethics and artificial life: From modeling to moral agents', *Ethics and Information technology*, **7**(3), 139–148, (2005).
[54] John P Sullins, 'When is a robot a moral agent?', (2006).
[55] John P Sullins, 'Robowarfare: can robots be more ethical than humans on the battlefield?', *Ethics and Information technology*, **12**(3), 263–275, (2010).
[56] Surgeon General's Office, '(MHAT) IV Operation Iraqi Freedom 05–07, Final Report', Advisory Report, Mental Health Advisory Team, (2006).
[57] Ryan Tonkens, 'Should autonomous robots be pacifists?', *Ethics and information technology*, **15**(2), 109–123, (2013).
[58] Ryan S Tonkens, 'Ethical implementation: A challenge for machine ethics', in *2nd Symposium on Computing and Philosophy*, pp. 38–45. AISB, (2009).
[59] Sam Vincent and Jonathan Brackley. Humans (eppisode 2), 2015.
[60] Wendell Wallach, 'Implementing moral decision making faculties in computers and robots', *AI & Society*, **22**(4), 463–475, (2008).
[61] Wendell Wallach, 'Robot minds and human ethics: the need for a comprehensive model of moral decision making', *Ethics and Information Technology*, **12**(3), 243–250, (2010).
[62] Wendell Wallach and Colin Allen, 'Ethicalife: A new field of inquiry', in *AnAlifeX workshop, USA*. Citeseer, (2006).
[63] Wendell Wallach and Colin Allen, *Moral machines: Teaching robots right from wrong*, Oxford University Press, 2008.
[64] Wendell Wallach, Colin Allen, and Iva Smit, 'Machine morality: Bottom-up and top-down approaches for modeling human moral faculties', *Ai & Society*, **22**(4), 565–582, (2008).
[65] Alan FT Winfield, Christian Blum, and Wenguo Liu, 'Towards an ethical robot: internal models, consequences and ethical action selection', in *Advances in Autonomous Robotics Systems*, 85–96, Springer, (2014).