Ethical Decision Making in Robots: Autonomy, Trust and Responsibility

Autonomy Trust and Responsibility

Fahad Alaieri[2,3] and André Vellino[1(✉)]

[1] School of Information Studies, University of Ottawa, Ottawa, Canada
avellino@uottawa.ca
[2] Electronic Business Technologies, University of Ottawa, Ottawa, Canada
falai055@uottawa.ca
[3] Management Information Systems, Qassim University, Buraydah, Saudi Arabia

Abstract. Autonomous robots such as self-driving cars are already able to make decisions that have ethical consequences. As such machines make increasingly complex and important decisions, we will need to know that their decisions are trustworthy and ethically justified. Hence we will need them to be able to explain the reasons for these decisions: ethical decision-making requires that decisions be explainable with reasons. We argue that for people to trust autonomous robots we need to know which ethical principles they are applying and that their application is deterministic and predictable. If a robot is a self-improving, self-learning type of robot whose choices and decisions are based on past experience, which decision it makes in any given situation may not be entirely predictable ahead of time or explainable after the fact. This combination of non-predictability and autonomy may confer a greater degree of responsibility to the machine but it also makes them harder to trust.

Keywords: Robot ethics · Autonomy · Trust · Responsibility

1 Introduction

Many aspects of robot behavior are ethically relevant. Robots may be either ethical patients, i.e. the subject of ethical behaviour by others (people or other robots), or ethical agents whose actions have ethical consequences. There are also questions of ethics in the design and the use of robots, such as whether and how robots are deployed, either for military purposes or to save human lives [6].

Our focus in this paper is robots viewed as ethical agents, i.e. moral reasoners with a sufficiently high degree of autonomy that enables them to make choices with ethical implications. We aim to provide both a framework for understanding the concept of autonomy in robotic devices and to analyze the process of choice, i.e. the making of decisions that characterizes the ethical dimension of their actions. We also indicate how these features of autonomy and choice and especially how these choices are made, could have implications for ascribing moral responsibility to robots.

A. Agah et al. (Eds.): ICSR 2016, LNAI 9979, pp. 159–168, 2016.
DOI: 10.1007/978-3-319-47437-3_16

2 Autonomy

Our working definition of a robot is a task-oriented device that has sensors and other information input interfaces, which is able to physically alter its environment, move, and have both the energy and ability to make decisions about how to accomplish its tasks. A key feature in a robot is whether its ability to make decisions is autonomous, i.e. whether it has the ability to operate without external intervention. From the point of view of its ethical decisions, autonomy is important because it is a necessary condition for ethical agency. While some argue that an autonomous robot cannot be considered truly autonomous unless it makes *all* its decisions without any human intervention, we prefer to say that such robots are not only autonomous but also *independent*.

One key characteristic of an autonomous robot is whether it is able to respond appropriately to a wide variety of situations. A machine that requires no external input to make a decision but is only ever able to make one decision could not be said to have a meaningful degree of autonomy. For example, a collaborative robot like Baxter, which is used to repeatedly perform only very specific tasks, exhibits some degree of autonomy but does not have the ability to make complex decisions that depend on highly variable environmental conditions and is unable to handle unpredictable situations.

2.1 Ethical Decision Making

Our discussion of autonomy relies on a model of the steps that a robot undergoes in the process of committing an action. Following the sense-plan-act robotics paradigm (see [8]) we propose a 5-stage model of the information processing in a robot: (i) obtaining the information (e.g. from sensors or telemetry); (ii) analyzing the information (e.g. by categorizing and integrating data); (iii) generating alternative courses of action (e.g. computing outcomes for a set of candidate decisions); (iv) selecting from among the alternatives (e.g. making a choice from among the candidates), and (v) performing an action that corresponds to this choice (e.g. activating an actuator) [1].

From the point of view of the ethical agency of a robot, the key stages are those that involve the generation of alternatives (iii) and the selection of a decision (iv). The hallmark of an ethical agent is that it has autonomy of choice in the decisions it makes: given a set of alternative courses of action from which the agent can choose, it has a method to select one. For a robot to be considered ethical, its actions need to not only conform to ethical norms but to perform these actions as a result of some process that morally obligates it to perform those actions. Thus a critical element in the decision selection, step (iv), is the ethical theory that is used to evaluate each alternative course of action. Although there are many ethical theories that can be adopted for the design of an ethical robot, here we consider only two: utilitarianism and deontological rules.

From the point of view of utilitarianism the value of an action is determined by the overall benefit of its consequences. Hence robots that have the ability to calculate the consequences of their actions and to evaluate the benefits they

bring about must be considered ethical. For instance, an autonomous car that has the capacity to detect pedestrians on the streets and to avoid them or stop driving in order to not harm them, behaves ethically from this point of view.

If, instead, a robot were to use a deontological framework that expresses its moral and ethical duties, it would have to act according to ethical principles that are independent of the consequences of its actions. For example, a deontologically ethical robot could be instructed not kill or lie or cheat, or cause harm, no matter what the circumstances or consequences.

Bernard Gert proposed ten such deontological rules that determine which of its actions are permitted, obligatory or prohibited, independently of their consequences: do not kill, do not cause pain, do not disable, do not deprive of freedom, do not deprive of pleasure, do not deceive, keep your promises, do not cheat, obey the law, and do your duty [13]. This could lead to situations in which some rules are at odds [28] with one another. For example, a robot's obligation to keep its promises (such as the promise to keep a secret) may be at odds with the obligation do not deceive. In other words ethical robots could experience much the same kinds of dilemmas as humans do and would also require mechanisms to resolve them.

Thus the choice of moral theory that governs a robot's behaviour is determined by the robot's decision-processing capabilities. If it has the ability to look ahead, plan, and evaluate the "goodness" of outcomes, then it could be designed to implement utilitarian principles. If it is only able to obey rules, then it may be that a purely deontological approach is more suitable, notwithstanding the need for a method to resolve rule-conflicts.

2.2 Top-Down and Bottom-Up Decision Making

A slightly different but complementary characterization of the ethical decision making process in a machine has to do with *how* the machine arrives at its ethical conclusions. Allen et al. refer to these alternative decision-processes as the 'top-down' and 'bottom-up' methods [2]. In the top-down approach the robot programmer installs decision making algorithms that produce predictable outcomes: in essence, it embeds in the machine what a human being considers to be ethical behaviour, which then needs only to determine *when* it is appropriate to apply them.

In the bottom-up approach, the programmer builds an open-ended system that is able to collect information from its environment, to predict the outcomes of its actions, to select from among alternatives and, most importantly, has the capacity learn from its experience. Such a machine can be described as having the ability to learn what is right and what is wrong because it is capable of learning from its choices and mistakes: it has the ability to self-modify its decision-making system through the acquisition of experiences. As Allen et al. put it "Top-down approaches ... involve turning explicit theories of moral behavior into algorithms. Bottom-up approaches involve attempts to train or evolve agents whose behavior emulates morally praiseworthy human behavior" [2].

A bottom-up approach can manifest in at least three ways: the robot could develop its own ethical decision selection methods by a process of trial and error (unsupervised learning); the machine's engineers could train the robot to learn pre-established moral rules (supervised learning); or the robot could adopt a hybrid learning method, which would allow it to keep learning from its experience and surroundings, but be grounded in pre-established principles.

For instance, a supervised learning method similar in nature to the neural networks in the Go playing program AlphaGo could be trained to learn to behave ethically by example with instances of situation-response pairs. In AlphaGo, this training step enables the computer to prune the space of possible Go moves from which it can choose (the so-called 'value networks' used to evaluate board positions) and then make a choice (using so-called 'policy networks') to evaluate which from among them is the best [11]. Both the 'value networks' and the 'policy networks' are trained from a large number of human games and the design methodology for such a game-player is a plausible model for how a bottom-up, learning, ethical reasoner might be trained.

3 Trust and Predictability in a Robot's Ethical Decisions

Robots in the future will have a greater capacity to perform even more tasks and an increasing number of these tasks will be related to people's safety, health, and even their lives. Hence people will have to develop confidence that robots are correctly obeying ethical principles if there is a risk that not following them could cause harm.

Two elements will contribute to this trust: humans will have to have repeated positive experiences with high-quality robot decisions and the decisions they make that obey ethical principles will have to be predictable and, retroactively, explainable. Without a coherent explanation for a robot's actions, a human would not be able to assess the validity of a robot's decision and therefore not have grounds for trusting it.

Yet, a robot that has the ability to modify the method by which it generates choice alternatives and calculates the consequences of its possible future actions may not be entirely predictable: its behaviour may become non-deterministic and how it came to make a choice may be complex, and hard for it or a human to explain. Tay, the Microsoft AI Twitter Chatbot, is an early example [7] of a hard-to-trust adaptable machine. It was designed to learn and adapt its (verbal) behaviour as a function of the input it received from its Twitter followers but its behaviour was not predictable by its programmers and it was easily 'vandalized' by people into uttering sexist and racist remarks.

It was not possible for Tay to conform to norms of ethical verbal behaviour because natural language understanding in machines has not yet reached the maturity required to deduce the consequences of verbal actions (such as the offense that can be caused by making racist remarks, which can be uttered in an infinite number of ways) let alone solved the problem of recognizing whether a remark would be considered racist or otherwise offensive.

But even if some of Tay's behaviour could have been moderated with (verbal) ethical norms, its unpredictability still poses a problem. It may be difficult for anyone, even programmers, to provide explanations for the behaviour of any machine whose behaviour is programmed 'bottom-up'. Consider for example, the choices made by decision-making algorithms such as those in AlphaGo. These are very hard to both predict and explain. When something goes wrong (or very well), such as when AlphaGo made some errors (or brilliant moves) in its recent games against the world champion Lee Se-dol, it was difficult, even for its developers, to know why it made those mistakes (or how it made some brilliant moves) [23]. This is a significant impediment to building human trust in a machine's ability to either generate an appropriate set of candidate-actions or to select the best from among them.

Suppose a machine had to make a choice in a complex utilitarian decision problem (e.g. a complicated version of the Trolley Problem [3]), in which a lot of options, choices and consequences had to be calculated. Suppose also that it functions perfectly and it makes the "right" decision and picks the morally correct course of action. It is conceivable that a human might not immediately recognize that this decision is optimal from the consequentialist point of view. A human (with limited computational ability) might conclude, incorrectly, that the robot's decision was unethical. But, without a coherent explanation for the robot's actions, a human would not be able to assess the validity of the robot's decision and therefore have no grounds for trusting it. On the other hand if the robot can explain its decision process in a way that a human can understand, that explanation could be the foundation for inducing human trust. Indeed, such explanations could be quite impressive to humans and eventually convince them of the robots' superior ability to make ethical decisions.

Such attempts at mapping machine-decision making into human-understandable accounts of their actions has been attempted with conventional robot planners [20] and noted to be necessary components for ethical robots [26]. However, as Colombo and Hartmann [9] remark about Bayesian models of cognitive phenomena, "[they do] not reveal ... the causal structure of a mechanism". Thus a deontological, rule-based ethical framework for controlling a robot's ethical decisions could generate clear human-understandable explanations for its actions whereas it may be very hard to do the same for decisions made by adaptive machine-learning algorithms such as the neural networks in AlphaGo.

4 The Moral Responsibility of Robots

The questions of choice and autonomy have an important role to play in determining whether robots can be held morally responsible. As Stahl observes, the traditional debate about whether computers can be responsible hinges on the question of whether they satisfy the conditions for agency and person-hood [27]. Hence the question of whether a robot is making its own decisions and how those decisions are being made would determine, at least in part, whether or not it could or should be held responsible for its actions.

The question of a machine's moral responsibility has been addressed using two approaches: the classical approach and the pragmatic approach. The classical approach views machines as not responsible for their actions under any circumstance — because they are mechanical instruments or slaves. In the pragmatic approach, 'artificial morality' envisages some situations under which machines can be viewed as responsible for their choices [12]. In this view responsibility in artificially ethical agents is a "social regulatory mechanism".

Others have focused on how to enable responsibility in artificial agents by embedding ethical codes of conduct in them. If these codes of conduct are formulated by the robot's designers, then the responsibility for those rules lies squarely with the robots' designers and owners (assuming that the owner has been apprised of theses rules). However, if these rules of conduct — whether or not they can be formulated in human-intelligible terms — are arrived at from experience (i.e., 'bottom-up'), the burden of responsibility for mistakes is more evidently on the machine's shoulders. In the case of Tay, Microsoft assumed a kind of "meta-responsibility" for not having predicted the possibility that it could be crowd-hacked by malicious users who would coax it into verbal misbehaviour. But many people saw its actual verbal mis-behaviours as being its fault—*it* was viewed as responsible for its racist utterances, not its manufacturer.

Jarvik's philosophical analysis divides human moral responsibility into three types: causal responsibility, role responsibility, and liability responsibility [16]. In causal responsibility, a person is responsible for everything that she has caused to happen: she is the cause of her actions. In role responsibility, a person's role in a certain area of society or community obligates them to perform a task, meaning that the task simply *is* the responsibility. The final form is liability responsibility, which identifies who is to be "praised or blamed" for certain actions or outcomes. Dodig-Crnkovic and Persson add one more critical element of moral responsibility besides causal responsibility: intention [12]. Causal responsibility may be assigned to non-humans, but, according to them, only humans have intentions. Insofar as malicious users intentionally fooled Tay into mis-behaving, the responsibility for its inappropriate comments also lies with them. As we noted above, we cannot say that Tay ever had the intention to offend and hence it is blameless.

Which of these types of responsibility can or cannot be assigned to robots? One consideration is that robots come in different varieties and not all types have the ability to shoulder responsibilities. For example, a highly autonomous robot could be said to have some causal responsibility because it is capable of making decisions that cause actions in a broad range of environments whereas the actions of a robot with low autonomy, typically caused by the human that controls it, would not.

Computers may be superior to humans in terms of the accuracy and quality of their decisions [4] because they have a greater ability to calculate all the consequences of an action that may be performed in a certain situation. So, with these innately superior capacities they might perform their social tasks [27] both perfectly and accurately and therefore be able to be *more* responsible than

humans — at least in the sense of role responsibility — in so far as they are better able to perform tasks effectively. For example, Japan is experimenting with 'urban surveillance robots' that are responsible for identifying criminals and detecting unusual behaviours [25]. Bank fraud detection systems that are responsible for blocking customers' credit cards when they detect unusual purchase patterns are another example of role-responsible machines: their responsibility is to protect both the card owner and the bank's financial assets. Robotic decision systems are therefore assuming responsibilities because of their ability to calculate, detect, inspect, and track.

Autonomous robots that have the capacity to interact with their environment, make decisions, perform tasks, and calculate the consequences of their actions can be thought to be responsible for their actions if they are also able learn how they ought to behave as a result of their experiences ('bottom-up'). Robots are morally responsible for performing actions that lead to ethical consequences in those cases where the action-choice is determined by the selection of one from among several alternatives [12] and that choice is not deterministically programmed by humans.

Perhaps most importantly, it is the capacity that robots have to learn from their mistakes that that allows humans to assign responsibility to them [14]. A robot that learns from its experience and is able to improve its own decision-making system is more capable of being afforded responsibility. Asaro predicts that, in the future, autonomous robots will have a greater ability to come up with their own moral rules, goals, and reasoning methods, and that they will thus be equipped to make moral decisions that fulfill the moral responsibility which has been assigned to them. We believe that this is a sound prediction and firmly based on the evolution of autonomy in the robot industry [6].

4.1 Can Robots Be Responsible?

What would happen if, in the future, autonomous robots were given full responsibility for their actions and outcomes [21,24]? Some researchers including Deborah Johnson believe that it is dangerous to give robots full responsibility for their actions because they might go beyond the programmers' control [17]. They might be autonomous because they perform tasks without human control, but, according to this view, it is the humans — including the manufacturers, designers, programmers, and users who must take responsibility if anything goes wrong. In this case, the mistake that caused the harm is human and the robots cannot be held responsible for their actions.

According to Kuflik, responsibility does not rest with robots, because they are just machines running programs that are manifestations of (human) intentions [19]. Responsibility for any action performed by a robot may be divided amongst different people such as the robot manufacturer and the user, and each group will shoulder part of this responsibility [5,28]. Hew claims that, in the foreseeable future of technology, "robots will carry zero responsibility" for their actions, and that this responsibility should remain with humans. This is because

"its rules for behaviour and the mechanisms for supplying those rules must not be supplied entirely by external humans" [15].

But also, abrogating responsibility by the robots' users and creators could encourage some people to create dangerous autonomous robots that may harm people or perform dangerous or unwanted tasks. Therefore, as Wallach argues, people and corporations should be held responsible for all harm that is caused by technology [29]. Kuflik agrees, concluding that the responsibility of robots' outcomes rests with the people who design them and who program their systems [19].

In some situations, users should shoulder all the responsibility if they use their robots intentionally for the purpose of harming others [10]. If a driver configures the autopilot system in an autonomous car to cause a collision, then the driver must take full responsibility for the consequences of the car's behaviour. Hence, if autonomous cars were given full responsibility for their actions it could be possible for evil people to fool them into harming others but fail to take responsibility for doing so. Hence, according to this argument, for every action performed by an autonomous robot, there should be a human agent who is held responsible in when something goes wrong.

One consideration when attributing liability responsibility to a machine is to ask who is responsible if the machine makes a mistake. Do we blame the machine, the manufacturer, the designer, the programmer, or the owner? Johnson argues that since artificial agents have become more autonomous and that nobody can fully predict their decisions, no one person can be held responsible for their actions [18]. This is all the more true if their actions cannot be fully explained.

The Ad Hoc Committee on Responsible Computing takes a different position. This committee crafted "The Rules", which were intended as ethical guidelines for computer professionals and state that people are answerable for their behavior when they produce or use computing artifacts, and that their actions reflect on their character. The first of these five rules states that "The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact." The third rule states "People who knowingly use a particular computing artefact are morally responsible for that use." [22]. Thus the people who design, develop, program, create, deploy, and use artificial agents are responsible for their agents according to their role in the action, decision, result, or harmful effects, at least to the extent to which these effects are "foreseeable".

5 Conclusions

If what constitutes an ethical choice in humans is either deliberating about the precedence of deontological rules amongst themselves in a given situation (e.g. what duty over-rides another) or analysing the consequences of a potential set of candidate options in a given situation and picking the one that optimizes a well-being function, then, in either case, these processes have a counterpart in the choice-behaviour of autonomous robots.

Therefore, autonomous robots can and will be ethical agents that are able to make ethical decisions. A key question is: will we be able to trust them if the methods by which they make decisions are opaque to humans? If their learning-by-experience algorithms have unpredictable consequences, will humans be able to trust them? Their unpredictability also means, symmetrically, that their actions may not be (easily) explainable—at least not in human terms.

Another key question is: who will be held responsible for the actions committed by autonomous ethical robots? If their actions are entirely predictable, then they are machines that are doing what they are programmed to do and the responsibility for the consequences of their actions must lie with the manufacturers and users. If, however, they are more like children who eventually learn to make their own decisions on the basis of experience and who induce their own deontological or utilitarian principles from a series of unsupervised learning processes, then they should be considered responsible for their actions.

The design of ethical robots that give them some degree of responsibility but also a sufficient degree of predictability to remain trustworthy might best be achieved with a hybrid strategy or a method that combines the 'bottom up' and the 'top down' approaches. An ethical robot built using a hybrid approach would have well-defined rules that predictably prevent catastrophic ethical failures, but also have the ability to learn new ethical principles from its experiences. A robot that is able to learn from its mistakes and perform utilitarian calculations to select one from among a set of alternative actions could thus be constrained by deontological rules that forbid it from considering some alternatives or oblige it to consider others, yet also perform consequentialist calculations more effectively than humans.

References

1. Alaieri, F., Vellino, A.: The Ethical Characteristics of Autonomous Robots. We Robot Poster, April 2016. http://hdl.handle.net/10393/34809
2. Allen, C., Smit, I., Wallach, W.: Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf. Technol. **7**(3), 149–155 (2005)
3. Allen, C., Wallach, W., Smit, I.: Why machine ethics? IEEE Intell. Syst. **21**(4), 12–17 (2006)
4. Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent. AI Mag. **28**(4), 15–26 (2007)
5. Asaro, P.: Robots and responsibility from a legal perspective. In: IEEE International Conference on Robotics and Automation, Rome, Italy (2007)
6. Asaro, P.: What should we want from a robot ethic? Int. Rev. Inform. Ethics **6**, 9–16 (2006)
7. Bass, D.: Clippy's back: the future of Microsoft is Chatbots. Bloomberg Businessweek, March 2016. http://www.bloomberg.com/features/2016-microsoft-future-ai-chatbots/
8. Beer, J.M., Fisk, A.D.: Toward a framework for levels of robot autonomy in human-robot interaction. J. Hum.-Robot Interac. **3**(2), 74–99 (2014)
9. Colombo, M., Hartmann, S.: Bayesian cognitive science, unification, and explanation. Br. J. Philos. Sci., axv036 (2015)

10. Crabb, P.B., Stern, S.E.: Technology traps. In: Luppicini, R. (ed.) Ethical Impact of Technological Advancements and Applications in Society, pp. 39–46. IGI Global, April 2012
11. Silver, D., Huang, A., Maddison, C.J., et al.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
12. Dodig-Crnkovic, G., Persson, D.: Sharing moral responsibility with robots: a pragmatic approach. In: Tenth Scandinavian Conference on Artificial Intelligence, SCAI 2008, pp. 165–168 (2008)
13. Gert, B.: Morality: Its Nature and Justification. Oxford University Press, USA (1998)
14. Hellström, T.: On the moral responsibility of military robots. Ethics Inf. Technol. **15**(2), 99–107 (2012)
15. Hew, P.C.: Artificial moral agents are infeasible with foreseeable technologies. Ethics Inform. Technol. **16**(3), 197–206 (2014)
16. Jarvik, M.: How to understand moral responsibility? TRAMES: J. Humanit. Soc. Sci. **7**(3), 147–163 (2003)
17. Johnson, D.G.: Computer systems: moral entities but not moral agents. Ethics Inf. Technol. **8**(4), 195–204 (2006)
18. Johnson, D.G.: Technology with no human responsibility? J. Bus. Ethics **127**(4), 707–715 (2014)
19. Kuflik, A.: Computers in control: rational transfer of authority or irresponsible abdication of autonomy? Ethics Inf. Technol. **1**(3), 173–184 (1999)
20. Lomas, M., Chevalier, R., Vincent Cross II, E., Garrett, R.C., Hoare, J., Kopack, M.: Explaining robot actions. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 187–188. ACM (2012)
21. Malle, B.F.: Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics Inf. Technol. **18**(70), 1–14 (2015)
22. Miller, K.W.: Moral responsibility for computing artifacts: "The Rules". IT Prof. **13**(3), 57–59 (2011)
23. Moyer, C.: How Google's AlphaGo Beat a Go World Champion. The Atlantic, March 2016. http://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/
24. Noorman, M., Johnson, D.G.: Negotiating autonomy and responsibility in military robots. Ethics Inf. Technol. **16**(1), 51–62 (2014)
25. Royakkers, L.: A literature review on new robotics: automation from love to war. Int. J. Soc. Robot. **7**(5), 1–22 (2015)
26. Scheutz, M., Malle, B.F.: think and do the right thinga plea for morally competent autonomous robots. In: 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, pp. 1–4. IEEE (2014)
27. Stahl, B.C.: Responsible computers? a case for ascribing quasi-responsibility to computers independent of personhood or agency. Ethics Inf. Technol. **8**(4), 205–213 (2006)
28. Tzafestas, S.G.: Roboethics. A Navigating Overview, vol. 79. Springer, Heidelberg (2016)
29. Wallach, W.: A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control. Basic Books, New York (2015)