Issues in robot ethics seen through the lens of a moral Turing Test

Gerdes, Anne; Øhrstrøm, Peter

Published in:
Journal of Information, Communication and Ethics in Society

DOI:
10.1108/JICES-09-2014-0038

Publication date:
2015

Document version
Peer reviewed version

Citation for pulished version (APA):
Gerdes, A., & Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral Turing Test. Journal of Information, Communication and Ethics in Society, 13(2), 98-109. https://doi.org/10.1108/JICES-09-2014-0038

 DOI 10.1108/JICES-09-2014-0038

Issues in robot ethics seen through the lens of a moral Turing test

AnneGerdes

Department of Design and Communication, University of Southern Denmark, Kolding, Denmark, and

Peter Øhrstrøm

Department of Communication and Psychology, Aalborg University, Aalborg, Denmark

Abstract

Purpose – The purpose of this paper is to explore artificial moral agency by reflecting upon the possibility of a Moral Turing Test (MTT) and whether its lack of focus on interiority, i.e. its behaviouristic foundation, counts as an obstacle to establishing such a test to judge the performance of an Artificial Moral Agent (AMA). Subsequently, to investigate whether an MTT could serve as a useful framework for the understanding, designing and engineering of AMAs, we set out to address fundamental challenges within the field of robot ethics regarding the formal representation of moral theories and standards. Here, typically three design approaches to AMAs are available: top-down theory-driven models and bottom-up approaches which set out to model moral behaviour by means of models for adaptive learning, such as neural networks, and finally, hybrid models, which involve components from both top-down and bottom-up approaches to the modelling of moral agency. With inspiration from Allen and Wallace (2009, 2000) as well as Prior (1949, 2003), we elaborate on theoretically driven approaches to machine ethics by introducing deontic tense logic. Finally, within this framework, we explore the character of human interaction with a robot which has successfully passed an MTT.

Design/methodology/approach – The ideas in this paper reflect preliminary theoretical considerations regarding the possibility of establishing a MTT based on the evaluation of moral behaviour, which focusses on moral reasoning regarding possible actions. The thoughts reflected fall within the field of normative ethics and apply deontic tense logic to discuss the possibilities and limitations of artificial moral agency.

Findings – The authors stipulate a formalisation of logic of obligation, time and modality, which may serve as a candidate for implementing a system corresponding to an MTT in a restricted sense. Hence, the authors argue that to establish a present moral obligation, we need to be able to make a description of the actual situation and the relevant general moral rules. Such a description can never be complete, as the combination of exhaustive knowledge about both situations and rules would involve a God eye's view, enabling one to know all there is to know and take everything relevant into consideration before making a perfect moral decision to act upon. Consequently, due to this frame problem, from an engineering point of view, we can only strive for designing a robot supposed to operate within a restricted domain and within a limited space-time region. Given such a setup, the robot has to be able to perform moral reasoning based on a formal description of the situation and any possible future developments. Although a system of this kind may be useful, it is clearly also limited to a particular context. It seems that it will always be possible to find special cases (outside the context for which it was designed) in which a given system does not pass the MTT. This calls for a new design of moral systems with trust-related components which will make it possible for the system to learn from experience.

DOI 10.1108/JICES-09-2014-0038

Originality/value – It is without doubt that in the near future we are going to be faced with advanced social robots with increasing autonomy, and our growing engagement with these robots calls for the exploration of ethical issues and stresses the importance of informing the process of engineering ethical robots. Our contribution can be seen as an early step in this direction.

Keywords Artificial moral agents, Deontic tense logic, Moral turing test, Robot ethics

Paper type Research paper

1. "As if" – a moral turing test

> Since being challenged to further activity, being set greater obstacles to overcome, is the sum and substance of our lives as teleological beings, developing robots – setting ourselves further technological – cultural goals – is not an inhuman or antihuman enterprise. It is simply part and parcel of the life of a species that first began cultivation the land, devising tools and machines, and cultivation – culturally developing – members of the species itself. Machines and artefacts are an inevitable part of human culture. Moral robots are merely a part that still lies in the future (Versenyi, 1974, p. 259).

Due to the growing interest in human robot interaction (Lin *et al.*, 2012; Benford and Malartre, 2007; Dautenhahn, 2007; Turkle, 2011; Wilks, 2010; Levy, 2008; and 2013), it would be useful to discuss artificial moral agency by considering the possibility of a Moral Turing Test (MTT), which might enable us to distinguish principles for evaluating morally correct *actions* rather than [as in the originally Turing test (1950)] skills of articulation. The Turing test is based on a criterion of indistinguishable behaviour, meaning that a computer system passes the test if a human interrogator is unable to distinguish between utterances produced by the computer and those produced by a human. The MTT questions whether a robot (or a computer system) acts at least according to the ethical standards that are normally considered acceptable in human society. It is important to point out that the development of a system that can pass the MTT will only be one early step towards producing an artificial moral agent (AMA). The kind of machine-based ethical reasoning needed to pass the MTT should not be confused with ethical autonomous decision-making. According to McDermott (2008), ethical decision-making involves a conflict between self-interest and ethics, whereas challenges regarding ethical reasoning concern how to formalise human reasoning processes, which are based on moral principles and may be computationally very complex, although they are not structurally different from other kinds of reasoning processes (McDermott, 2008, p. 2). Ethical reasoning presupposes a notion of free choice in the sense that alternative future possibilities are open to the person in question. To be a genuine ethical decision-maker, one must also be free in the sense that one can sometimes choose to act in one's self-interest, even though it runs contrary to moral prescriptions. The notion of free choice is also needed to represent responsibility properly. A robot does not have to be free in this sense to pass the MTT, but it needs to have a representation of what it means to act freely. As we shall see, this can be achieved in terms Prior's branching time model.

Given that we depend on various kinds of robotic services, for the sake of safety at least, we may want AMAs to be better at "doing ethics" than humans (Allen *et al.*, 2000, 2009). Therefore, if an AMA passes the original Turing test, the bar is set too low, as it would allow the AMA to be as fallible as humans are. It seems reasonable to demand more of AMAs than we expect from humans, as we would, of course, like them to be

reliable robots, and as we want them, unlike humans, to avoid becoming distracted by all sorts of "irrelevant" observations while carrying out moral reasoning processes prior to action. We just want them to be effective and to follow the rules – including the ethical standards which have been implemented. Thus, robots should be able to out-perform humans in an MTT test set up across different domains. Hence, the perspective of the MTT shifts in character to become a comparative MTT, in which the aim is to establish which agent is unfailingly more moral across a set of ethically relevant situations (Allen *et al.*, 2000, p. 255). In this sense, a comparative MTT would provide a tool for risk assessment, useful when computer scientists and engineers strive to design "a morally praiseworthy agent" (Allen *et al.*, 2000, p. 261) capable of perfect moral judgements and actions within given domains, indeed preferably within every possible domain. Within this behaviourist framework, we might consider the idea of artificial moral agency from a performance perspective in maintaining that morality can be decided by mere appearance, which on the face of it seems reasonable enough, as how do we settle whether human beings are virtuous or not? Simply by judging their behaviour – i.e. she is a generous person because she acts out of generosity; she is a moral person because she always acts in a morally appropriate fashion. Why then should we demand more or something else for robots? Could there be situations in which some decisions made on the basis of mechanistic moral reasoning will be non-satisfactory?

1.1 The role of inner states

The above-mentioned argument that "acting good" can be taken as an indication of "being good" fits nicely into a classical utilitarian framework, in which intentions and correct reasons for acting are disregarded and only the consequences of acts are taken into account in the evaluation of moral behaviour. However, to most moral philosophers, internal reasoning cannot be omitted. Aristotle remarked that there is a distinction between being "good" and merely "acting good". He based virtue ethics on a concept of well-being or *eudaimonia* and highlighted *phronesis* as the form of wisdom related to practical reason in action. This form of proficiency is not neutral but moral in its being, as it mirrors a form of reflection grounded in practice and cultivated by our ability to be involved and to take a stance in any specific situation. Furthermore, according to Kant, we find that reasons count; moral obligations and actions are considered as categorical "oughts" derived from *a good will*, i.e. my ability to act from a sense of duty implied by the fact that I am capable of carrying out rational reasoning in accordance with moral rules that may guide my conduct. A more recent example of the importance of internals can be found in the work of Moor (2009). Here, Moor distinguishes between four types of ethical agents: *Ethical Impact Agents*, which are machines that have an obvious ethical impact on the surroundings – as an example, Moor mentions the robotic Qatar camel jockeys, which save young boys from engaging in the dangerous race. At the next level, we find *implicit ethical agents*, representing systems designed to avoid unethical or undesired outcomes – such as a simple control system in an ATM that blocks purchases when faced with user patterns suggesting fraud (Wallach and Allen, 2009, p. 29). Next, we find *explicit ethical agents*, which are machines that "do" ethics and are able to carry out ethical reasoning within restricted domains. They act not because they *want* to, but because their programming *causes* them to do so (Putnam, 1964[1], p. 672). Finally, *full ethical agents* may be considered as ethical agents similar to human beings in the sense that they have free will, consciousness and intentionality and, hence, the capacity for

DOI 10.1108/JICES-09-2014-0038

being held responsible for their actions. As such, inner states seem to matter, and abilities for moral reasoning have to be understood as situated in the unified whole of human life and experience, as pointed out in the Dreyfusian attack on the whole idea of Artificial Intelligence (AI) (Dreyfus, 1992). Further, as stated by Searle in his famous *Chinese Room* argument against so-called strong AI, a machine may be perfect in displaying verbal behaviour and, thus, able to pass the Turing test, but all it does is manipulate meaningless symbols without any intention behind it and, thus, without sense – "Simulation is not duplication and syntax is not semantics" (Searle, 1995, p. 75, 1980). Hence, to be a *full ethical agent*, a robot would have to have mental states, which we normally attribute to humans.

Within his framework, Moor argues that we should focus on robot ethics at the level of limited *explicit ethical agents*, as such robots imply interesting ethical issues and, at the same time, constitute a realistic future scenario, whereas the idea of *full ethical agents* in the shape of robots with mental states seems to be purely speculative and would require technology far beyond our imagination (Moor, 2006, p. 21).

2. Pragmatics: approaches to the design of an MTT

From an engineering perspective, we might note that inner states, consciousness, motivations and intentions may all count. Yet, performance is all we have access to in judging the moral actions of both humans and robots. Hence, we should set out to seek solutions that would give empirically testable results, which allow us to measure whether a robot simulates moral behaviour in a satisfactory manner. Still, re-describing what counts as preferable artificial moral agency in terms of the above-mentioned comparative MTT is one thing, whereas actually bringing it to life is something entirely different. We need to consider how to build moral robots using programming based on moral philosophy and moral reasoning within a given context.

To design a system which can pass an MTT, we need at least to implement a relevant ethical theory. It seems that Kant's ideas of ethics could be useful in this context, as they involve some important ideas regarding human moral agency. In fact, Kant stated that the fact that we are aware that we might act morally wrong is what makes us responsible creatures, and this fact is, therefore, essential to our humanity. This means that it should be possible to conduct reasoning about moral questions. Facing this Kantian challenge, many researchers have tried to formulate relevant deontic logic. One of the originators of this enterprise was Prior (1914-1969), who wanted to study the logical machinery involved in the theoretical derivation of obligation. He wanted to find what he called "The Logic of Obligation" (Øhrstrøm *et al.*, 2012). In an early study, he claimed that such a logical system had to be based on complete descriptions of (a) the actual situation, and (b) the relevant general moral rules.

Prior stated his fundamental creed regarding deontic logic in the following way: "[…] our true present obligation could be automatically inferred from (a) and (b) if complete knowledge of these were ever attainable" (Prior, 1949, p. 42). Clearly, the combination of the requirements regarding (a) and (b) would involve an God's eye view, which will make it possible to know all there is to know and take everything into consideration before making a (perfect) moral decision. From an engineering point of view, it might be tempting to try and design a robot with such a God's eye view, capable of making perfect moral evaluations as a basis for carrying out perfect moral behaviour. On the other hand, due to the frame problem, we would have to be modest and settle for a formal

description representing moral reasoning and preferable events or outcomes within a restricted domain and a limited space-time region.

In seeking to engineer, a theoretical approach to morality similar to the Kantian approach, Allen and Wallach distinguish between a top-down theoretically driven approach, a bottom-up developmental or explorative approach and finally a hybrid (Wallach and Allen, 2009, Chapter 6, 7 and 8), which combines both top-down and bottom-up approaches and furthermore includes a virtue ethical component. They all run into similar problems from different angles. Therefore, in the following discussion, we will mainly place emphasis on the theoretically driven approach and only briefly sketch essential points regarding the remaining strategies. Hence, the developmental approach includes an adaptive learning or value-emerging line of attack to artificial intelligence, as in the one reflected in embodied architectures, such as neural nets, genetic algorithms and connectionism. From this developmental approach, "grown up" guidance is needed to ensure that the AMA learns to behave properly. Thus, the system cannot learn anything from scratch if it has not got build-in architecture that allows for desirable values to emerge on a background, which implies a kind of built-in mechanism to navigate from when distinguishing right from wrong.

According to Allen and Wallach, a pure top-down approach will run into trouble due to the frame problem following in the wake of formally seeking to represent a scope of ethical reasoning by applying theory-driven rules, i.e. decision algorithms, for ethical actions, which point out satisfactory outcomes in a contextually open domain. Therefore, they introduced a hybrid model, which not only combines but also integrates top-down and bottom-up approaches by incorporating virtue ethics as a theoretical foundation for implementation of the idea of how we develop into virtuous persons through learning by habit, which goes well with the model of connectionism. The hybrid model seems to give rise to the same type of problem regarding specifying rules for decision-making algorithms, or developing self-learning architectures, as well as for deciding which virtues are going to be taken into consideration (since, for one thing, virtue ethics have not yet been able to come up with a finite catalogue of virtues). To sum up, the hybrid model seems to us to be rather futuristic for the time being and, furthermore, as every day moral reasoning does make use of top-down rules in explaining moral actions, this approach still holds some promise for turning ethical theories into workable and implementable models of ethical reasoning (at least) within restricted domains (Allen and Wallach, 2009, p. 83).

As an example thereof, Anderson *et al.* (2004) represent a theory-driven approach in their suggestion in which they model ethical reasoning by a combination of two components: First, they make use of act utilitarianism, which allows for a kind of cost-benefit calculation of outcomes of pleasures and displeasures with regard to a given action. Second, they apply Ross' theory of duty-based actions, relying on *prima facia* duties: fidelity, reparation, gratitude, justice, beneficence, non-maleficence and self- improvement. Finally, Rawls concept of "reflective equilibrium" is used to weigh relevant *prima facia* duties up against each other:

> Instead of computing a single value based only on pleasure/displeasure, we must compute the sum of up to seven values, depending on the number of Ross' duties relevant to the particular action. The value for each such duty could be computed as with Hedonistic Act Utilitarianism, as the product of intensity, duration and probability (Anderson *et al.*, 2004, sec. 3).

DOI 10.1108/JICES-09-2014-0038

No matter which model of ethical reasoning we may choose to implement to establish a system which may pass an MTT, it will have to take time and modality into account. This was strongly emphasised in the works of Prior on deontic logic. It is evident that Prior's long-term ambition was to incorporate the logic of ethics into a broader context of time and modality. Unfortunately, he was never able to pursue this goal in detail, but he certainly managed to establish the broader context of time and modality into which the logic of obligation has to fit. To indicate what such an approach involves, we shall make use of a simplified example.

Let us imagine that an agent in a certain situation or scenario, S_0, has to choose between two future possibilities represented by the scenarios, S_1 and S_2, which may both in principle be realised tomorrow. The agent wants to act in a morally correct fashion and carries out careful reasoning to do so. This is done within the scope of tempo-modal logic corresponding to a branching time system. A simplified system of that kind can be represented by the following Figure 1.

This branching time diagram involves four so-called chronicles (c_{10}, c_{11}, c_{20}, c_{21}), i.e. possible courses of time. At S_0, the future possibilities S_1 and S_2 are both possible, and one of them is necessary. To perform moral reasoning in this case, we need a procedure according to which the possible chronicles may be evaluated. We also need a logical formalism which allows us to reason on future possibilities. According to Prior, this can be obtained by applying tempo-modal logic conceived as an extension of propositional logic. In this logic, the future and the past are treated as propositional operators, *P* and *F*, to which durational specification may be added, e.g. letting *P(1)* and *F(1)* stand for "yesterday" and "tomorrow", respectively.

Letting s_1 and s_2 stand for propositional descriptions of the situations corresponding to S_1 and S_2, and *M* for "it is possible that […]", the propositions $MF(1)s_1$ ("s_1 may occur tomorrow") and $MF(1)s_2$ ("s_2 may occur tomorrow") are both true. We may even assume that s_1 and s_2 are mutually exclusive in the sense that there is a proposition, *p*, implied by s_1, the negation of which is implied by s_2. As Prior demonstrated, this can, in fact, be done in several ways. Prior preferred the so-called Peircean model. However, we suggest the use of the so-called Ockhamistic model (Øhrstrøm and Hasle, 2011), which is much closer to everyday reasoning. In this case, it will be true in general that:

$$F(1)p \vee F(1) \sim p$$

In addition, there will be a clear distinction between *MF(1)*, *F(1)* and *NF(1)*, where *N* stands for "it is necessary that […]" (Note that the two modal operators are related in terms of the negation, i.e. $N = \sim M \sim$). The model is indeterministic, as, clearly, both *NF(1)p* and $NF(1) \sim p$ are false at S_0. This is obviously essential when we are dealing with ethics and moral reasoning based on the notion of free choice. Given such a system, we then need to add the logic of an operator, *O*, corresponding to "obligation" to handle the reasoning related to an MTT. But how should this new operator be conceived? It is

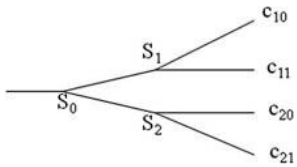

Figure 1.
Branching time diagram

important for Prior to emphasise that "obligation" cannot just be defined in terms of "the best total consequences". The reason is that the very notion of "total consequences" does not make sense, as what happens in the future depends, in principle, on the choices of a number of free agents (Prior, 2003, p. 65). Another significant problem in this context is that we can never have complete knowledge of all the possible chronicles in the branching time system. Furthermore, we can never be sure that we have all relevant general moral rules and principles of evaluation. In short, our model cannot be complete, as we do not have an all encompassing view of the world. However, we should still strive for descriptions that are as detailed as possible. However, such descriptions will in general be limited both in respect to the space-time region, the selection of persons and objects and the set of evaluation principles and rules taken into consideration. We will also have to involve some probabilistic reasoning. In creating such a system, which may indeed pass the MTT, we should include a clear account of the general relations between the basic notions of time, modality and obligation. This means that we have to consider a number of conceptual problems. One such problem has to do with the so-called Kantian principle, which is the claim that if something (say an act that leads to $F(x)p$) is obligatory, then it is also possible, i.e.:

$$OF(x)p => MF(x)p$$

Similarly, we could refer to the so-called Hintikka principle, i.e. the claim that if something is impossible, then it is forbidden (i.e. its negation is obligatory). Formally:

$$\sim MF(x)p \Rightarrow O \sim F(x)p$$

A number of relations of this kind have to be considered to establish a system which will allow us to discuss obligation in a tempo-modal context (Øhrstrøm *et al.*, 2012). It is still an open question exactly which relations should be accepted and which should be rejected. Clearly, the actual implementation of such a system corresponding to an MTT has to be based on a formalisation of the logic of obligation, time and modality. Although there is a lot to discuss regarding the precise properties of such a logical system, the actual formulation of reasonable candidates, which will work in specific contexts, is not too far away. This means that it may be possible to produce early prototypes of MTT implementations. Such systems may be useful for empirical studies of ethical reasoning, a point which is formulated in an even stronger sense by Anderson and Anderson:

> Ethics, by its very nature, is the most practical branch of philosophy. It is concerned with how agents ought to behave when faced with ethical dilemmas. Despite the obvious applied nature of the field of ethics, however, too often work in ethical theory is done with little thought to real world application. When examples are discussed, they are typically artificial examples. Research in machine ethics, which of necessity is concerned with application to specific domains where machines could function, forces scrutiny of the details involved in actually applying ethical principles to particular real life cases. As Allen and Wallach, 2009] recently stated, AI "makes philosophy honest." Ethics must be made computable to make it clear exactly how agents ought to behave in ethical dilemmas (Anderson and Anderson, 2007, p. 16)

According to this view, the Priorean approach would mean that we should define the *O*-operator in terms of a formalised system of rules and evaluative principles, which are relevant in the context we consider. In this way, it may be possible to make ethics

DOI 10.1108/JICES-09-2014-0038

computable in the context in question. However, we have to admit that no matter how much we take into account, the system will never be complete, and it will have to be revised in due course.

3. Human robot interaction: challenges in dealing with an AMA awarded the MTT certificate

Let us for the sake of argument assume that in the future an AMA passes the comparative MTT; not by behaving indistinguishably from a human moral agent, but by out-performing him or her by being able to do qualified calculations based on the situation and the accepted ethical principles represented in a formal manner. Through its ethical decision-making algorithms, the AMA calculates its way forward to the best ethical response to the case at hand. This might seem ideal to us and ensure reliable human – robot interaction. But does "a morally praiseworthy agent" (Allen *et al.*, 2000) equal a "morally perfect artificial agent" and, if so, do encounters with it come at a price we should not want to pay?

To sharpen our imagination, let us seek inspiration by digging into the science fiction movie "I, Robot" (Proyas, 2004), which takes place in 2035 in a world where social robots interact with humans as polite and caring servants. Designed to ensure smooth and secure interaction, these robots have built-in morality in the shape of pre-programmed moral codes, namely, Asimov's Three Laws of Robotics. Everything seems perfect, but the central character, Detective Spooner, still has a strong aversion to robots. He was once involved in a car accident in a river where he tried to save a 12-year-old girl from drowning, but instead he was saved by a robot, who interfered with his actions and computed (maybe based on a deontic branching time model) that Detective Spooner had a higher probability of survival than the girl. With this "time-to-moral-market" knowledge at hand, it was not difficult for the robot to make a morally correct choice, which could be judged desirable and evaluated as morally good across the robot's different built-in moral frameworks. For instance, within the robot's utilitarian framework, the robot's moral behaviour is judged by the consequences of its rescue of Detective Spooner, which turned out to represent the best possible outcome under the given circumstances. Also, within its deontological framework, the robot's action can be judged as equally morally good. Here, the robot's moral system activates reference to the double-effect doctrine (Quinn, 1989), which emphasises that it is sometimes permissible to cause harm as a side effect (double effect), even though it would not be tolerable to cause that particular kind of harm as a means of doing good. In this particular case, the robot acted according to a specific system of moral reasoning which necessitated saving Spooner while letting go of the girl because her chances of survival were small, while the detective could be saved if the robot acted quickly. Still, Spooner, the old- fashioned "homo sapiens ludditus", is sure that in the same situation, a human would have saved the girl rather than him. This means that according to Spooner, a human would have rejected the system of moral reasoning used by the robot and looked for a revision of it to represent moral responsibility in a satisfactory manner. Humans would have chosen to save the girl, despite knowing that she stood a poor chance of survival. What is ethically at stake here can be elaborated further by turning to Løgstrup's phenomenologically founded ethics (Løgstrup, 1997). This allows for exploring in more detail the importance of interpersonal concerns and the concrete

situations we find ourselves in when considering how we ought to behave. In Løgstrup's account of ethical demands, he emphasises the priority of trust over distrust for human beings in pointing to our mutual dependency as constituting the very fabric of our moral life:

> [A]n ethical demand takes its content from the unshakable fact that the existence of human beings is intertwined with each other in a way that demands of human beings that they protect the lives of others who have been placed in their trust (Løgstrup, 1997, p. 290).

Hence, our mutual dependency, from which all the ethics of life spring, has to be understood as grasping what is at stake between two persons in a given situation. Where a Kantian approach calls for objectivity, rationality, autonomy and universalism, Løgstrup instead stresses the importance of acting completely for the sake of the other: of realising what is demanded of me, as a moral self, under the given circumstances to seek to fulfil the demands of the other. The demand bears with it a paradox, as when we become aware of it as a demand, only conscious obedience is left and, hence, the demand cannot be fulfilled in its radical sense of acting entirely for the sake of the other: *What is demanded is that the demand should not have been necessary. This is the demand's radical character* (Løgstrup, 1997, p. 146). Quite often, we do of course weigh up arguments for and against when we consider what to do in a specific interpersonal context, and here Løgstrup emphasises that ethics is not only about guiding our moral choices in a situation but also, and more importantly, to remind us of the interpersonal concerns we put aside – no matter how well-argued they may be – when we act.

Løgstrup's position does not imply relativism, but rather underscores that the ethical demand first and foremost can be characterised as a requirement to us as individuals, not specifically as rational beings. Thus, in line with virtue ethics, it would be ethically wrong (only) to let respect for the moral law guide our moral actions. Of particular importance is consideration for the specific other in the given situation, and this consideration cannot be overshadowed by rule-following conduct and a morally weighted calculation. Thus, even though the robot calculated for an optimal solution, it did not live up to the ethical demand which is part and parcel of all human interaction. The crucial point seems to be that all formal systems dealing with ethics are partial as compared with human ethics. This means that for any formal system of moral reasoning, there will be a situation in which some consequences of the system are unacceptable. In this sense, human ethics will, in principle, transcend any formal system, i.e. any implementation. This clearly questions the assumption stated above according to which an AMA is imagined to be out-performing humans. This may be the case in a practical and limited sense, but not in the general sense. Human ethics is, in principle, beyond any attempt at formalising moral reasoning in a logical system, no matter how useful such formalisations might be.

If robots are ultimately capable of acting with moral perfection by means of having access to relevant knowledge and ethical principles, which humans in certain situations will be right in rejecting, then encounters with robots are perhaps less promising than they at first seem. On the other hand, it may be possible that deeper reflections on Løgstrup's emphasis on human trust and interpersonal relations may give rise to new ways of implementing ethics in machines. It would

DOI 10.1108/JICES-09-2014-0038

certainly be attractive if such systems were able to learn to be wiser and morally better. Of course, the crucial problem here would be the detection of the persons that the system ought to trust.

In "I, Robot", the robots gradually develop beyond their initial capabilities and finally decide that the only way to truly protect human beings is to protect them from themselves. This means that Løgstrup's fundamental trust disappears (or becomes morally irrelevant), at least between robots and human beings. Thus, in the end, in a derogation of Asimov's laws of robot ethics, the robots turn against humanity. Fortunately, this logical implication is derailed by Detective Spooner. In this way, paternalism evaporates in favour of human autonomy, which carries with it the human capacity for failure, and which is what made us moral beings in the first place: the fact that something important involving the life – be it in a minor or major sense – of another person, is always at stake when we make up our mind and act upon it. This kind of ethical approach probably cannot be described without taking the notions of trust, responsibility and free choice into account.

4. Concluding remarks

It seems that Prior was right in claiming that the formulation of a formal system, which correctly incorporates all aspects of moral reasoning, would, in principle, require a complete description not only of all relevant moral rules and laws, but also of all relevant aspects of the situation in question. However, having such descriptions is tantamount to having a God's eye view of all relevant aspects of reality. As we can never have anything like that a complete and unquestionable system of moral reasoning cannot be established. On the other hand, we have argued that although there are many open questions regarding the precise properties of the formal relations between time, modality and obligation, it is in fact possible to formalise important aspects of ethical reasoning in a specific context and thereby contribute to a system which may pass a comparative MTT in that particular context. Systems implemented on the basis of such formalisations will, however, be partial and temporary in the sense that the actual moral evaluations can be questioned when their implications are considered in concrete situations. As we have seen, an observation of this kind may lead to a revision of the system of ethical reasoning. Clearly, this may happen again and again. Any formalisation of ethical reasoning may have to be revised when confronted with real life. Clearly, humans are faced with the very same fact, which means that this limitation may not disqualify the system as seen in relation to an MTT. One important consequence of this is that we have to distinguish between modelling moral reasoning and actual decision-making in moral questions. Creating a system which can pass an MTT may not give us a system which can provide satisfactory decisions in practical situations. Nevertheless, the study of possible systems which may pass an MTT can certainly give rise to useful and important insights concerning moral reasoning. The Spooner case not only shows the limitations of the Priorean approach to deontic reasoning, it also suggests that we should look for new ways of making ethical systems. To the Priorean system, we may add trust-related components which will make it possible to learn from and adjust to experiences.

Note

1. Here, Putnam refers to an argument from an unpublished paper by Baier given at the Albert Einstein College of Medicine 1962. In this particular article, Putnam's concern is not with how to speak about machines but rather about how we should speak about humans. Thus, "clarity with respect to the 'borderline case' of robots, if it can only be achieved, will carry with it clarity with respect to the 'central area' of talk about feelings, thoughts, consciousness, life, etc." (Putnam, 1964, p. 669). In this way, Putnam argues for the possibility of robot consciousness as something that calls for a decision rather than a discovery.

References

Allen, C., Garvy, V. and Zinser, J. (2000), "Prolegomena to any future artificial moral agent", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 12 No. 3, pp. 251-261.

Anderson, M., Anderson, L.S. and Armen, C. (2004), "Towards machine ethics", available at: www.aaai.org/Papers/Workshops/2004/WS-04-02/WS04-02-008.pdf

Anderson, M. and Anderson, S. (2007), "Machine ethics: creating an ethical intelligent agent", *AI Magazine*, Vol. 28 No. 4, pp. 15-26.

Benford, G. and Malartre, E. (2007), *Beyond Human – Living with Robots and Cyborgs*, A Forge Book, New York, NY.

Dautenhahn, K. (2007), "Socially intelligent robots: dimensions of human-robot interaction", *Philosophical Transactions: Biological Sciences*, Vol. 362 No. 1480, pp. 679-704.

Dreyfus, H.L. (1992), *What Computers Still Can't Do*, MIT Press, Cambridge, MA. Levy,

D. (2008), *Love and Sex with Robots*, Duckworth, London.

Lin, P., Abney, K. and Bekey, G.A. (2012), *Robot Ethics – The Ethical and Social Implications of Robotics*, The MIT Press, Cambridge, London.

Løgstrup, K.E. (1997), *The Ethical Demand*, University of Notre Dame Press, Notre Dame. McDermott, D.

(2008), "Why ethics is a high hurdle for AI", *North American Conference on Computers and Philosophy (NA-CAP)*, Bloomington, IN, 30 June, available at: www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf

Moor, J. (2006), "The nature, importance, and difficulty of machine ethics", *IEEE Intelligent Systems*, Vol. 21 No. 4, pp. 18-21.

Moor, J. (2009), "Four kinds of ethical robots", *Philosophy Now*, Vol. 72, pp. 12-14.

Øhrstrøm, P. and Hasle, P. (2011), "Future contingents", *The Stanford Encyclopedia of Philosophy*, Summer 2011 Edition.

Øhrstrøm, P., Zeller, J. and Sandborg-Petersen, U. (2012), "Prior's defence of Hintikka's theorem. a discussion of prior's 'the logic of obligation and the obligations of the logician'", *Synthese*, Vol. 188 No. 3, pp. 449-454.

Prior, A.N. (1949), *Logic and the Basis of Ethics*, Oxford University Press, Oxford.

Prior, A.N. (2003), in Hasle, P., Øhrstrøm, P., Braüner, T. and Copeland, T. (Eds), *Papers on Time and Tense*, Oxford University Press, Oxford.

Proyas, A. (2004), *I, Robot*, 20th Century Fox.

Putnam, H. (1964), "Robots: machines or artificially created life?", *The Journal of Philosophy*, Vol. 61 No. 21.

Quinn, W.S. (1989), "Actions, intentions, and consequences: the doctrine of double effect", *Philosophy & Public Affairs*, Vol. 18 No. 4.

 DOI 10.1108/JICES-09-2014-0038

DOI 10.1108/JICES-09-2014-0038

Searle, J.R. (1980), "Minds, brains and programs", *Behavioral and Brain Sciences*, Cambridge University Press, Cambridge, Vol. 3 pp. 417-424.

Searle, J.R. (1995), "How artificial intelligence fails", *The World & I. Currents in Modern Thought – Artificial Intelligence: Oxymoron or New Frontier*, pp. 285-295.

Turing, A. (1950), "Computing machinery and intelligence", *Mind*, Vol. 59, pp. 433-460.

Turkle, S. (2011), *Alone Together – Why We Expect More From Technology and Less From Each Other*, Basic Books, New York, NY.

Versenyi, L. (1974), "Can Robots be Moral?", In *Ethics*, Vol. 84 No. 3, pp. 248-249.

Wallach, W. and Allan, C. (2009), *Moral Machines – Teaching Robots Right from Wrong*, Oxford Scholarship Online.

Wilks, Y. (2010), *Natural Language Processing 8: Close Engagements with Artificial Companions – Key Social, Psychological, Ethical and Design Issues*, John Benjamins Publishing Company, Amsterdam.

Further reading

Kant, I. (1953/1785), *Groundwork of the Metaphysic of Morals*, (trans. H.J. Paton), *The Moral Law*, Hutchinson, London.

Kant, I. (1974/1787), *Kritik der Praktischen Vernuft*, Surhkamp Verlag, Berlin. Schärfe, H. and

Geminoid, D.K. (2013), available at: http://c.aau.dk/geminoid/

Corresponding author

Anne Gerdes can be contacted at: gerdes@sdu.dk