Artificial moral agents are infeasible with foreseeable technologies

Abstract: For an artificial agent to be morally praiseworthy, its rules for behaviour and the mechanisms for supplying those rules must not be supplied entirely by external humans. Such systems are a substantial departure from current technologies and theory, and are a low prospect. With foreseeable technologies, an artificial agent will carry zero responsibility for its behavior and humans will retain full responsibility.

Keywords: artificial agent; moral agent; ethical agent; moral responsibility; automation; robotics.

Introduction

The emergence of robotics and automation has spawned an enormous literature on their ethical issues (Krishnan, 2009; Lichocki, Kahn, & Billard, 2011; Lin, Bekey, & Abney, 2008; Tonkens, 2012). A central issue is of *responsibility* (or *accountability*), prompting a multitude of questions: Can an artificial system be a moral agent (Allen, Smit, & Wallach, 2005; Allen, Varner, & Zinser, 2000; Dodig-Crnkovic & Çürüklü, 2012; Floridi & Sanders, 2004; Himma, 2009; Johansson, 2010; Stahl, 2004; Sullins, 2006; Torrance, 2008; Wallach, 2008)? Can an artificial system be responsible (Sparrow, 2007; Stahl, 2006)? When (if ever) does accountability transfer from human designer(s) to robot (Grodzinsky, Miller, & Wolf, 2008; Hanson, 2009; Johnson & Miller, 2008; Kuflik, 1999)? What is the distinction between autonomy and accountability (Swiatek, 2012)? How do we preserve the intuition that humans are somehow special with regard to morality (Coeckelbergh, 2009)?

This paper establishes that artificial moral agents are infeasible with foreseeable technologies, as its contribution to the debate on responsibility for robots and automated systems. For an artificial agent to be morally praiseworthy, its rules for behaviour and the mechanisms for supplying those rules must not be supplied entirely by external humans. Such systems are a substantial departure from current technologies and theory, and are a low prospect. For the foreseeable future, an artificial agent will carry zero responsibility for its behavior and humans will retain full responsibility. The result supports the central premises of (Billings, 1991; Kuflik, 1999), confirms the claims of (Grodzinsky et al., 2008), and distinguishes actual or potential real-world systems from those imagined (and feared) in speculative fiction.

The paper proceeds in three parts. We first propose a criterion for denying moral agency. We then establish that systems of foreseeable construction are trapped by the criterion. We finish with implications for the debate on responsibility vis-à-vis robotics and automation. For clarity, this article is concerned with the technologically feasibility of creating an artificial moral agent. We leave aside the question of whether they ought to be created (see (Johnson & Miller, 2008) for example).

Criterion for denying moral agency

Definitions

We require a definition of moral responsibility that is precise enough to argue a position, and consistent with earlier thinking in moral philosophy. We start by accepting the concept of an *agent* as something that can act in the world. We then declare that an agent is *morally praiseworthy*, and can be held *morally responsible* for an action, if it is worthy of praise for having performed the action. We call such an agent a *moral agent*. (For brevity, 'praise' will cover both praise or blame throughout this article.)

The definition frames the issue of responsibility for artificial systems in terms of worthiness, as in, 'By what criteria is an agent *worthy* of moral praise?' The ascertaining of sufficient conditions for moral praise is perhaps *the* question of moral philosophy, and we will not attempt it here. Given the state of robotics and automation technology, we instead seek a necessary condition, as in 'By what criteria is an agent *unworthy* of moral praise?'

Our definition should otherwise be uncontentious. We are working within what (Himma, 2009) dubbed the *standard view* of moral agency, citing as reference works (Eshleman, 1999; Haksar, 1998; Williams, 2014). Neither he nor we claim that the standard view is correct. Our interest is merely to proceed from a starting point of relevance to moral philosophy in the large. At reviewer request, we follow (Eshleman, 1999) into Aristotle's *Nicomachean Ethics* Book III.1-5 (Aristotle (translated by W. D. Ross)):

> *Since virtue is concerned with passions and actions, and on voluntary passions and actions praise and blame are bestowed, on those that are involuntary pardon, and sometimes also pity*, to distinguish the voluntary and the involuntary is presumably necessary for those who are studying the nature of virtue, and useful also for legislators with a view to the assigning both of honours and of punishments. … . [Book III.1 *emphasis added*]

Aristotle (apparently) makes two proposals. First, that 'virtue' (moral responsibility) is about the bestowing of praise. Second, that to qualify for such praise, an agent must be capable of 'voluntary' action. We accept the first proposal. It due course, we will see a resonance between this article's contribution and Aristotle's second proposal.

More recently, (Allen et al., 2000) proposed that 'the ultimate objective of building an AMA [Artificial Moral Agent] should be to build a morally praiseworthy agent'. Again, we do not wish to invoke authority, only to establish that our definitions allow us to participate in the debate as per Allen et al and similar commentaries. A parallel discussion has also evolved on the possibility of artificial ethical agents (Moor, 2006). We focus on moral agency as defined, though our methods and findings may also apply to ethical agency where the two concepts intersect.

Given our focus on robotics and automation, we also take an *intelligent agent* as anything that can *close a loop* from sensors to effectors without human intervention (Russell & Norvig, 2003). 'Without human intervention' is evidently a very low threshold for 'intelligent', but the term is accepted as jargon. Mousetraps, toilet tank-fill valves, thermostats and automobile cruise controls are all examples of intelligent agents, albeit they are *simple-reflex intelligent agents*, the simplest type. Of course, electronics and digital computing have enabled the construction of sophisticated intelligent agents that can outperform humans, at least in some circumstances.

Position

We assert that a mousetrap is *not* morally praiseworthy. It closes its loop from trigger to trap entirely at the volition of one or more humans: those who emplaced and armed it, designed it, constructed it and so on. Responsibility for the mousetrap is held in total by those humans. We defer the question of how responsibility is apportioned across the humans; our focus is on the mousetrap's responsibility being zero. Our assertion is unlikely to be controversial – if a mousetrap is morally praiseworthy, then any intelligent agent can be (recall that a mousetrap is of the simplest type).

In studies of moral agency it is conventional, convenient and sufficient to talk of systems that follow rules. Thus a mousetrap may be described as following the rules, 'Continuously monitor the pressure on the trigger. If the pressure exceeds threshold, then activate the trap.' When we declared that the mousetrap acts entirely at the volition of one or more humans, we may equally say that the mousetrap follows rules were supplied entirely by one or more humans.

We generalize from the mousetrap to a *criterion for denying moral agency* as follows:

> **Definition**: For any system S, let $\mathfrak{R}(S)=\{S'_i\}_{i\in I}$ denote the systems that supplied S with its rules, where I is an index set.
>
> **Proposed criterion for denying moral agency**: Given system S, construct the sets $\mathfrak{R}(S)=\{S'_i\}_{i\in I}$, $\mathfrak{R}(S'_i)=\{S''_{i'_i}\}_{i'_i\in I'_i}$, $\mathfrak{R}(S''_{i'_i})=\{S'''_{i''_{i'}}\}_{i''_{i'}\in I''_{i'}}$, …. Extract the sequences S, S', S'', S''', … (by iterating across the index sets I, I'_i, $I''_{i'}$, …). Then S is *not a moral agent* if for all such sequences there exists n finite such that $S^{(n+1)}$ is a human(s) who is external to systems $S \ldots S^{(n)}$.

If applied to S as a mousetrap, we would find that the $S^{(n+1)}$ were the aforementioned humans at who's volition the mousetrap operates. The caveat that the humans be 'external to systems $S \ldots S^{(n)}$' is for logical consistency, for systems where humans close loops from sensors to effectors. A human, equipped and waiting to ambush a mouse, *may* be morally praiseworthy, if acting under their own volition or volition shared with others. Said human might *not* be morally praiseworthy, if acting under coercion (as one example).

As shorthand, we say that 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. ‘System’s rules must not be supplied entirely by external humans’ corresponds to the case for $n = 0$. ‘Mechanisms for supplying those rules must not be supplied entirely by external humans’ restates the case for n general.

In denying moral agency to S, we imply a procedure for ascribing responsibility for S which we now make explicit:

> **Proposed procedure for ascribing responsibility**: Suppose that S has been denied moral agency by our proposed criterion. Then responsibility for S is held by the humans identified as $S^{(n+1)}$ therein.

Location of position within the history of moral philosophy

Our position is hardly new, although our descriptions, reasoning and formalisms appear to be novel. As foreshadowed and at reviewer request, we return to *Nicomachean Ethics* Book III.1-5, for Aristotle’s ponderings on the nature of ‘voluntary’ or ‘compulsory’ action:

> Since virtue is concerned with passions and actions, and on voluntary passions and actions praise and blame are bestowed, on those that are involuntary pardon, and sometimes also pity, to distinguish the voluntary and the involuntary is presumably necessary for those who are studying the nature of virtue, and useful also for legislators with a view to the assigning both of honours and of punishments. Those things, then, are thought involuntary, which take place under compulsion or owing to ignorance; *and that is compulsory of which the moving principle is outside, being a principle in which nothing is contributed by the person who is acting or is feeling the passion*, e.g. if he were to be carried somewhere by a wind, or by men who had him in their power. [Book III.1 *emphasis added*]
>
> *What sort of acts, then, should be called compulsory? We answer that without qualification actions are so when the cause is in the external circumstances and the agent contributes nothing*. But the things that in themselves are involuntary, but now and in return for these gains are

> worthy of choice, and whose moving principle is in the agent, are in themselves involuntary, but now and in return for these gains voluntary. They are more like voluntary acts; for actions are in the class of particulars, and the particular acts here are voluntary. What sort of things are to be chosen, and in return for what, it is not easy to state; for there are many differences in the particular cases. [Book III.4 *emphasis added*]
> But if someone were to say that pleasant and noble objects have a compelling power, forcing us from without, all acts would be for him compulsory; for it is for these objects that all men do everything they do. And those who act under compulsion and unwillingly act with pain, but those who do acts for their pleasantness and nobility do them with pleasure; it is absurd to make external circumstances responsible, and not oneself, as being easily caught by such attractions, and to make oneself responsible for noble acts but the pleasant objects responsible for base acts. *The compulsory, then, seems to be that whose moving principle is outside, the person compelled contributing nothing*. [Book III.5 *emphasis added*]

The emphasized text coincides with the existence of n finite such that $S^{(n+1)}$ is a human(s) external to $S \dots S^{(n)}$. What Aristotle called the 'moving principle' we would call the rules followed by a system.

Systems of foreseeable construction are trapped by the criterion for denying moral agency

The problem with foreseeable artificial systems is that for all sequences S, S', S'', S''', … considered under our criterion, there exists n finite such that $S^{(n+1)}$ is a human(s) who is external to $S \dots S^{(n)}$. Indeed, the following candidate technologies are trapped:

- *Self-replicating programs*. A self-replicating program can write an exact copy of itself. Said programs have been written by human programmers (as textbook exercises).

- *Self-modifying code*. Self-modifying code modifies itself in the course of being executed. In examples to date, the section of code that performs modification itself remains static, in the form supplied by its human programmer.
- *Machine learning* systems are said to have the ability to learn from experience. The computer has programs for transforming inputs to outputs, and can modify those programs on the basis of experience. However, the modifications are made by a program that itself remains static, in the form supplied by its human programmer.
- *Self-regulating adaptive systems and meta-adaptive systems*. Such systems have a capacity to adapt, and a capacity to assess and modify this adaptive behaviour (Paramythis, 2004, 2006; Trevellyan & Browne, 1987). Instead of adapting under some fixed logic, the logic of adaptation can evolve over time. However on foreseeable technologies (and for the examples cited here), the logic of adaptation is itself fixed, under a human-supplied program.
- *Self-organizing systems*. An artificial system that supplies its own rules would indeed be a 'self-organizing system', but the converse does not hold. The term 'self-organizing system' has, unfortunately, become ambiguous on this very point. De Wolf and Holvoet supply a working definition (Serugendo, Gleizes, & Karageorgos, 2006; Wolf & Holvoet, 2004, 2005), 'Self-organization is a dynamical and adaptive process where systems acquire and maintain structure themselves, without external control.' They then cite as an example, 'Plugging in a PnP [plug and play] device in a computer can be considered as normal data input. A self-organizing behavior could be the autonomous configuration of drivers by the computer system. If a user has to install the drivers himself then there is no self-organization.' The ambiguity is that the computer's operating system embeds an algorithm that specifies the driver for the plug and play device. Said algorithm was supplied by a human programmer.
- *Evolutionary computing*. Evolutionary computing sets up a population of candidate solutions to problem. The population is iteratively grown into new candidate solutions, and culled against selection criteria. In current

implementations, the mechanisms for growth and selection are specified by a human programmer.

- *Hypercomputation* refers to the computability of functions beyond the reach of Turing Machines (Ord, 2006). The rules of computation are still supplied by a human programmer.

How to avoid being trapped by the criterion

For S to have any possibility of being a moral agent, we want at least one sequence where there is no n finite such that $S^{(n+1)}$ is a human who is external to $S \dots S^{(n)}$. Perhaps the most promising way is to contrive for systems $S \dots S^{(n)}$ with $S^{(m)} \in \Re\left(S^{(n)}\right)$ for some $m \leq n$. When applying our criterion, we would then have a sequence S, S', S'', S''', … $S^{(n)}$, $S^{(m)}$, $S^{(m+1)}$, … $S^{(n)}$, $S^{(m)}$, $S^{(m+1)}$, … $S^{(n)}$, $S^{(m)}$, $S^{(m+1)}$, … , and continuing in an endless loop. Moreover, at no point in the sequence will we find a system $S^{(N)}$ such that $S^{(N+1)}$ is a human who is external to $S \dots S^{(N)}$.

Our claim is that such a sequence is beyond the horizon of foreseeable technologies. We do *not* claim that the requisite setup is impossible, but note that speculations about plausibility are not the same as a demonstration of existence. We are looking for a system $S^{(n)}$ that can provide rules to $S^{(m)}$ and thereby rewrite its own rules. We may equivocate on the magnitude of changes that need to be made to $S^{(n)}$ – the pertinent point is that system(s) $S^{(n)}$ are static under foreseeable technologies. (The idea of a system rewriting its rules so that its origins are no longer recognizable was suggested by (Turing, 1947 (1986)). The structure of $S^{(n)}$ to $S^{(m)}$ and then back to $S^{(n)}$ is reminiscent of ‘twisted hierarchies’ and ‘strange loops’ from (Hofstadter, 1999, 2007).)

Connectionist approaches (such as neural networks) may offer a path to the requisite capability. We understand that connectionist systems are characterized by units interacting via weighted connections, where a unit’s state is determined by inputs received from other units (Hinton, 1989). The opportunity is for unit states to define the rules used by other units. In this way, the connectionist system as a whole could come to supply its own rules. If the states and weights are determined by training applied by humans, then the system no longer qualifies – its rules were supplied by an external human.

Responses to commentary

The reviewers raised two further questions that are best addressed here, in the spirit of open commentary.

What about an adult human in her 'right mind'?

An adult human in her 'right mind' would typically be regarded as a moral agent. Our criterion should accord with this position.

We propose that if S is an adult human in her 'right mind' then $S \in \Re(S)$. That is, if we describe S as a system that follows rules, then we include S in the suppliers of those rules. Under our criterion, S retains the possibility of being a moral agent; we do *not* deny moral agency to S. In terms of the earlier discussion of avoiding the criterion, S is a sequence with $m = n = 0$.

Emphasizing for clarity, our criterion gives correct results in *not* denying moral agency to an adult human in her 'right mind'. This is obviously not the same as affirming that said human is a moral agent, but is within the goals of our project.

Perhaps moral responsibility dwindles away?

Suppose a human builds a machine that in turn builds a more sophisticated machine, and so on to a huge (but finite) number of generations n. Would we hold the human responsible for the actions of the nth machine? Or does moral responsibility 'dwindle away'?

In the terminology of our criterion, we have the following counter-proposition:

> **Counter-proposition on moral responsibility 'dwindling away'**: Let S be such that for all sequences S, S', S'', S''', … considered under the criterion for denying moral agency, there exists n finite such that $S^{(n+1)}$ is a human(s) who is external to systems $S \ldots S^{(n)}$. If n is sufficiently large then said humans need not be held responsible for S.

Our initial response is to reject the counter-proposition, at least in situations where the end product is known. In particular, choose S as a mousetrap. Our position is that a mousetrap is *not* morally praiseworthy, and that responsibility for the mousetrap is held in total by the humans that built it. The existence of machines $S' \ldots S^{(n)}$ does not change the mousetrap, and neither therefore should its praiseworthiness change. We can, furthermore, envisage the machines $S' \ldots S^{(n)}$ being very simple, as in $S^{(k+1)}$ builds $S^{(k)}$ by assembling parts according to a fixed algorithm. Hence $S' \ldots S^{(n)}$ will not morally praiseworthy (otherwise anything can be morally praiseworthy). But if we accept the counter-proposition, then for n sufficiently large we can absolve (or partially absolve) the humans from responsibility for the mousetrap – a contradiction.

Hence our initial response is that it is impossible to set a threshold n of the kind envisaged by the counter-proposition. But there are at least two further possibilities that we can examine here, that change the character of the counter-proposition.

The first is that we have worked backwards from a known end product (such as a mousetrap). What about setting off from some $S^{(n)}$ which eventuates at some S that was not predicted? One counter is to consider a machine $\tilde{S}$ that builds machines, where we cannot predict what $\tilde{S}$ will build. Suppose in particular that we activate $\tilde{S}$ and it builds some S. As we chose to use $\tilde{S}$, it would seem reasonable (and our procedure for ascribing responsibility decrees) that we gain total responsibility for S. Then suppose that we then learn that $\tilde{S}$ is actually $S^{(n)}$ in disguise, such that it actually constructed a sequence of machines $S' \ldots S^{(n)}$ but we did not see them. Under the counter-proposition, we would lose some of the responsibility for S. That is, if we know that S is constructed from $S' \ldots S^{(n)}$, we can declaim responsibility for S, but if we know only that S is constructed from $S^{(n)}$ (disguised as $\tilde{S}$) then we are responsible S. Responsibility for S therefore rests on the *appearance* of n generations from $S^{(n)}$ to S. We do not have an outright contradiction of the counter-proposition, but the situation is unsatisfactory nonetheless.

The second possibility is in the commentator’s proposal that the machines become more sophisticated. Suppose we start a machine $\tilde{S}$ that eventually builds a machine $\tilde{\tilde{S}}$, where $\tilde{\tilde{S}}$ modifies $\tilde{S}$ such that rules governing $\tilde{S}$ bear no resemblance to those that were originally supplied. Now suppose that such a combination eventually builds a machine S. Such a combination avoids being trapped by our criterion. Moral responsibility has not ‘dwindled away’ – it has been redirected into an infinite loop.

Implications for studies of responsibility for robots and automated systems

For applied ethicists

The findings bring clarity to contemporary studies of responsibility for robotics and automation. If we accept the proposed criterion for denying moral agency to an agent, and the consequent appraisal that artificial moral agents are technologically infeasible (for the foreseeable future), then we revert to apportioning responsibility for an artificial system to one or more humans.

Just how responsibility ought to be apportioned across those humans is a separate matter. However, while the question isn’t trivial, it can be pursued with the confidence that the robot holds no portion of the responsibility. For example, in his influential examination of military robots, Sparrow asks whether the programmer, commanding officer or robot should be held responsible if the robot makes an unauthorised attack (Sparrow, 2007). For the foreseeable future, the question reduces to whether the programmer or commanding officer is responsible, in the mixture of faulty-machine-used-correctly to working-machine-used-incorrectly.

Correspondingly, ethicists should we should monitor developments towards artificial systems that can supply their own rules. The indicator is where the system rewrites its components to the point that they cannot be attributed to a human.

Our findings have an application in cognitive ergonomics. Billings assumed that the human operator carried ultimate responsibility for automation, and noted that this assumption is axiomatic in civil aviation (Billings, 1991). Billings’s assumption underpins his concept of human-centered automation, which has featured strongly in the design of human-machine systems. We have replaced the axiom with a chain of logic as to why current machines cannot be held responsible.

For future studies

Two further points may be made about the methods of study. The first relates to language. In studying the ethics of robots and automated systems, we are straddling across two sets of jargon – each internally consistent, but with vastly different meanings outside of their domain. Expanding from the earlier distinction of an *intelligent agent* from a *moral agent*: for technologists, an intelligent agent is *autonomous* from being able to close a loop from sensors to effectors without human intervention (Russell & Norvig, 2003). In customary approaches to moral philosophy (and especially in the Kantian school), a moral agent is *autonomous* from being able to impose the moral law on itself (Christman, 2011; Denis, 2012). The same word labels different concepts (see (Stensson & Jansson, 2013) for complaint and a discussion of consequences). We made progress here by codifying how 'autonomy' in the moral sense requires a capability that is beyond the reach of foreseeable technological systems.

The difference in jargon leads to our second point, on exploring further conditions for a moral agent. The key is to express the conditions in terms that are independent of implementation to avoid circular reasoning (setting a condition that requires a human and then concluding that the condition can only be met by a human). 'Close a loop from sensors to effectors' can be met by human or machine. 'Supply its own rule-supplying mechanisms' can be satisfied by a human, but not by a machine under foreseeable technologies.

Comparison with earlier proposals

Artificial moral agents proposed as being impossible

We acknowledge and contrast our findings with proposals from the literature. Our closest precedent is (Grodzinsky et al., 2008), who considered finite state machines. If the machine had a fixed state transition table, it could not be a moral agent. If the machine could modify its table, then the designer still retained some moral responsibility. Grodinsky et al's position rests on their assertion that if a system's behavior can be ascribed explicitly to its designer, then the system is not a moral agent. We concur with their position and recognize its precursors; notably (Turing, 1947 (1986)) proposed to look at machines that modified their instructions out of all recognition. We make a further claim – for machines built from foreseeable technologies, we may find that the machine modifies its own table *but* we would then find that portion of the table that performs modification remains intact. Consequently, the machine can be decomposed into two machines S' and S, where the table for S is modified by S' but the table for S' is fixed. By Grodinsky et al (and our) criteria, machine S' is not a moral agent and so too for the original machine.

Our second-nearest precedent is (Kuflik, 1999). He asserted that humans must bear the ultimate moral responsibility for a computer's decisions, as it is humans who design the computers and write their programs. He further proposed that humans can never relinquish oversight over computers. On the possibility of a programmed computer that evolves beyond its original program, Kuflik believed that if it is humans who programmed the self-reprogramming, then those humans will bear responsibility ('in some sense'. Commentaries on Kuflik need care, as he explores six possible definitions for 'responsibility'). We concur with Kuflik on this point. On the proposal that humans can never relinquish oversight, Kuflik does not appear to provide a defense. We offer that humans will be unable to relinquish oversight for computers built from foreseeable technologies.

Stahl denied that artificial systems are morally praiseworthy. Two of his reasons are of interest here: that artificial systems lack consciousness in the human sense and therefore cannot be said to have intentions, and that they lack freedom of will or action from being determined by hardware or software (Stahl, 2006). Friedman and Kahn Jr had earlier stated that intentionality was necessary for moral praise, and that computer systems as conceivable today in material and structure cannot have intentionality (Friedman & Kahn Jr, 1992). Johnson asserted that computers lack mental states, or the intendings to act (Johnson, 2006); Tonkens similarly asserted that artificial systems would be programmed to act according to rules installed by the programmer, and otherwise not act (Tonkens, 2009); Bryson likewise stated that responsibility for actions executed by an artefact lie with the humans (Bryson, 2012); Ruffo asserted that robots follow the program as supplied to it (Ruffo, 2012); Stensson and Jansson asserted that technological artefacts do not have a life of their own and therefore cannot know the real meaning of fundamental human values (Stensson & Jansson, 2013). The positions need to be justified, against proposals that artificial consciousness is possible (see (Chalmers, 1993) for example), and that the brain gives rise to intention yet it is a deterministic system (Chalmers, 1992). We supply a justification that applies to artificial systems that we currently know how to build, namely those that follow rules supplied by a human.

Himma argued that artificial systems will need to be conscious if they are to be moral agents (Himma, 2009). Himma's position faces ambiguities in definitions – for example, he describes cats as being conscious beings. His strategy otherwise resembles ours, in seeking a capability that humans exhibit but that is missing in foreseeable technologies. Commenting on Himma, Gunkel noted that consciousness is a difficult condition to assess (Gunkel, 2012). We concur, and posit that we may first assess whether behaviour is entirely at the volition of an external human. If we were to encounter an artificial system that drove its own behaviour, then new thresholds may apply.

Matheson proposed that an artificial system is worthy of holding responsibility if it is has *weak programming*. *Weak* programming was contrasted with *strong programming* – an agent that is strongly programmed cannot overcome the effects of the programming because it will always cause the agent to reason and behave in the manner the programming dictates (Matheson, 2012). Matheson's notion of weak programming might be viewed as the agent supplying its own program, as per this article. Our further contribution is to appraise the prospects for such systems under foreseeable technologies.

Hellström proposed that humans' tendency to assign moral responsibility to a robot increases with the robot's degree of autonomous power (Hellström, 2013). While the proposal is an hypothesis about how humans think, we might equally take it as a proposal for whether robots should be held responsible. Hellström left vague the notion of *autonomous power*, and in particular, the question of what it means for a robot to act 'on its own'; for example, he noted that a landmine could be said to act 'on its own' when it reacts to its trigger, but that on the other hand the landmine explodes only as a result of its constructor's agenda. Asaro similarly considered a continuum from amorality to (what he called) 'fully autonomous morality', and noted that children are not treated as full moral agents (Asaro, 2009). Asaro was also vague on the conditions for full moral agency. In our view, the distinction between an intelligent agent versus a moral agent (where both terms are as defined for this paper) is key to the debate on whether robots are worthy of being held responsible.

Parthemore and Whitby railed against the tendency to assign responsibility to robots of present and immediately foreseeable construction, stating that they lacked the requisite decision-making machinery (Parthemore & Whitby, 2012). We concur with their conclusion, from a tightened chain of logic. The important step was in describing an element of the requisite decision-making machinery, independent of implementation (be it by human or artifact).

Champagne and Tonkens proposed to assign a ‘blank cheque’ responsibility to a person of sufficiently high standing, in return for social prestige (Champagne & Tonkens, 2012). The person needed to be able to choose the responsibility of their own volition. For our purposes, the important point is that Champagne and Tonkens were responding to requirements for moral agency that are different to, and stronger than, those which we have used. Addressing (Sparrow, 2007), they accept Sparrow’s position that a moral agent must be capable of suffering (in addition to its actions originating from the agent). The same position was taken by (Torrance, 2008)). On the presumption that it is impossible for machines to suffer, a human needs to be substituted. We do not contest Champagne and Tonkens’ proposal; we and they have a common starting point, and then they have reasoned from a further premise to an additional conclusion.

Indeed, Champagne and Tonkens observed that Sparrow’s concern (and thus their proposal) applies to a very restricted set of robots – those with sophistication such that it is difficult to say that a human is responsible, but also difficult to say that the robot is praiseworthy. Sparrow’s concerns rest on an appraisal of technology by proponents of artificial mind (citing Brooks, Dyson, Kurzweil and Moravec), ambiguities in the nature of ‘autonomy’ (a question that he leaves open) and the inability of machines to suffer. We have clarified an aspect of autonomy, to establish a threshold with respect to technology as can be foreseen at time of writing.

Artificial moral agents proposed as feasible

Allen et al proposed that a robot may be regarded as a moral agent if its behaviours are functionally indistinguishable from a moral person – a Moral Turing Test (Allen et al., 2000). The Moral Turing Test was subsequently disavowed as a criterion for genuine moral agency, in recognition of controversies surrounding the Turing Test (Allen et al., 2005). Indeed, the Moral Turing Test inherits the Turing Test's (potential) vulnerabilities to the Chinese Room Argument. Our position echoes (Bringsjord, Bello, & Ferrucci, 2001) points about the Turing Test, wherein proponents must articulate why the robot holds moral responsibility, and not its programmer.

Coeckelbergh proposed that humans are justified in ascribing a virtual moral responsibility to those non-humans that appear similar to themselves (Coeckelbergh, 2009). Coeckelbergh avoided the question of whether artificial systems were morally praiseworthy. He argued that assessment of moral praiseworthiness required the ability to establish whether a given machine is a free and conscious agent, but that doing so was impossible. We differ by working from the contrapositive – we concur that to be morally praiseworthy entails a certain capability, and argue that current artificial systems are unable to supply this capability.

Matthias presented a number of cases in which a machine's behaviour ought to be attributed to the machine and not its designer or operators (Matthias, 2004). To hold the humans responsible would be an injustice, but to hold the machine responsible would challenge 'traditional' ways of ascription. He dubbed the result a 'responsibility gap'. Matthias's reasons coalesce into three propositions. The first proposition is that modern machines are inherently unpredictable (to some degree), but they perform tasks that need to be performed yet cannot be handled by simpler means. We accept Matthias's position that a machine may be unpredictable, but reject the conclusion of attributing responsibility to the machine. Rather, we would apportion some of the responsibility to those humans who choose to construct and/or use the machine. In footnotes, Matthias indicates that the choice to use a risky technology is made by society as a whole. We disagree, and argue that the choices are made by individual humans. A given human may choose to follow one or more other humans ('society'). If choice is coerced then we attribute responsibility to the coercers.

Matthias's second proposition is that there are increasing 'layers of obscurity' between manufacturer and system, as handcoded programs are replaced with more sophisticated means. Matthias's point is immensely important when apportioning responsibility across the multiplicity of people who contribute to making a modern automated system, across its constituent components. We reject, however, the conclusion that responsibility ought to be attributed to the machine.

Matthias's third proposition concerns machine learning systems, namely that the rules by which they act are not fixed during the production process, but can be changed during the operation of the machine. We agree, with the caveat that the rules by which they learn are fixed, at least on foreseeable technology. We therefore have a variant on Matthias's first proposition, about unpredictable machines. We may otherwise accept that the system may come to a portion of responsibility, if it has or gains the ability to supply its own rules – a possibility beyond foreseeable technology. It would be an error, however, to hold that a system will be able to supply its own rules as a result of being increasingly sophisticated. Matthias does not make this error, but it can be seen in (Human Rights Watch, 2012) for example.

In a similar vein, we dispute the claim by (Dodig-Crnkovic & Çürüklü, 2012) that as technology improves, there will be no problem in ascribing to artificial systems the capacities by which they can be regarded as worthy of holding responsibility.

Floridi and Sanders proposed that an artificial agent is moral if it met criteria designated as interactivity, autonomy and adaptability (Floridi & Sanders, 2004). Floridi and Sanders framed their proposal in their language of Levels of Abstraction (LoA), effectively a mirror image to talking of systems meeting or failing to meet criteria. Where they say that a system constitutes an agent at one LoA but fails to be an agent at another LoA, we would say that a system qualifies as a particular type of agent but fails the criteria for another type. The criteria (interactivity, autonomy and adaptability) were defined in a manner specific to their paper, but were claimed to be consistent with (Allen et al., 2000). For the current author, the Floridi and Sanders criteria are difficult to work with, and to apply reliably.

We also support the observation by (Johnson & Miller, 2008) that if Floridi and Sanders wish to establish that computer systems can be autonomous moral agents, they would have to establish a level of abstraction in which the notion of being moral is delineated. In effect, they would have to establish a set of sufficient criteria for being moral. We have taken a weaker (though still pertinent) position, in establishing a necessary criterion for being moral, and then showing that foreseeable systems fail to achieve the criterion.

Proposals that embed human ethics into an artificial agent

A number of articles are best characterized as proposals to embed human ethics into an artificial agent. We have the thought-leading work by (Anderson, Anderson, & Armen, 2005) on creating an ethically-sensitive machine. Their machines execute algorithms that are said to implement an ethical principle. As the algorithms are imposed on the machine, the machine is itself *not* an artificial moral agent, regardless of how well it performs. The same observation applies to the machines of (Arkoudas, Bringsjord, & Bello, 2005).

Powers explored how a computer could be programmed to be (or at least simulate) a Kantian moral agent (Powers, 2006). He explicitly recognized that if we stipulate the class of universal moral laws to the machine, then we would have human ethics operating through a tool and not machine ethics. In broad terms, his system would generate its own maxims and test them for consistency. As it built up maxims, it would need a set of rules for accepting each additional maxim. Said rules undermine a claim for moral agency, in relying on an external programmer.

Arkin proposed to embed ethics into battlefield robots (Arkin, 2007).

The proposed architecture assigns ultimate responsibility to a human operator/commander, and provides a Responsibility Advisor to this end. Responsibility might otherwise be apportioned to the robots' builders (Lucas, 2011).

Conclusion

With foreseeable technologies, an artificial agent will carry zero responsibility for its behavior and humans will retain full responsibility. Ethicists should attend the question of how responsibility for an artificial system apportions to one or more humans. They should also monitor for systems that can modify their own rules and rule-supplying mechanisms, such that the modifications erase the rules and mechanisms that were originally supplied.

Acknowledgments: The authors thank the anonymous reviewers for their comments and feedback. This article is UNCLASSIFIED and approved for public release. Any opinions in this document are those of the author alone, and do not necessarily represent those of the Australian Department of Defence. This article was originally published in Ethics and Information Technology. The final publication is available at Springer via http://dx.doi.org/10.1007/s10676-014-9345-6 .

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology, 7*(3), 149-155. doi: 10.1007/s10676-006-0004-4

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251-261. doi: 10.1080/09528130050111428

Anderson, M., Anderson, S. L., & Armen, C. (2005). Towards Machine Ethics: Implementing Two Action-Based Ethical Theories. In M. Anderson, S. L. Anderson & C. Armen (Eds.), 2005 Association for the Advancement of Artificial Intelligence Fall Symposium (Vol. FS-05-06). Menlo Park, California: The AAAI Press. Retrieved from http://www.aaai.org/Library/Symposia/Fall/fs05-06.

Aristotle (translated by W. D. Ross). ( ). *Nicomachean Ethics*

Arkin, R. C. (2007). Governing Lethal Behaviour: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture: Georgia Institute of Technology.

Arkoudas, K., Bringsjord, S., & Bello, P. (2005). Toward Ethical Robots via Mechanized Deontic Logic. In M. Anderson, S. L. Anderson & C. Armen (Eds.), 2005 Association for the Advancement of Artificial Intelligence Fall Symposium (Vol. FS-05-06). Menlo Park, California: The AAAI Press. Retrieved from http://www.aaai.org/Library/Symposia/Fall/fs05-06.

Asaro, P. (2009). What Should We Want from a Robot Ethic? In R. Capurro & M. Nagenborg (Eds.), *Ethics and Robotics*. Amsterdam: IOS Press.

Billings, C. E. (1991). Human-Centered Automation: A Concept and Guidelines: National Aeronautics and Space Administration.

Bringsjord, S., Bello, P., & Ferrucci, D. (2001). Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines, 11*(1), 3-27.

Bryson, J. J. (2012). *Patiency Is Not a Virtue: Suggestions for Co-Constructing an Ethical Framework Including Intelligent Artefacts*. Paper presented at the AISB/IACAP World Congress 2012, Birmingham, UK.

Chalmers, D. J. (1992). Subsymbolic Computation and the Chinese Room. In J. Dinsmore (Ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap*: Lawrence Erlbaum.

Chalmers, D. J. (1993). *A Computational Foundation for the Study of Cognition*.

Champagne, M., & Tonkens, R. (2012). *Bridging the Responsibility Gap in Automated Warfare*. Paper presented at the AISB/IACAP World Congress 2012, Birmingham, UK.

Christman, J. (2011). Autonomy in Moral and Political Philosophy. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2011 Edition ed.). Retrieved from http://plato.stanford.edu/archives/spr2011/entries/autonomy-moral/.

Coeckelbergh, M. (2009). 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. doi: 10.1007/s00146-009-0208-3

Denis, L. (2012). Kant and Hume on Morality. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2012 Edition ed.). Retrieved from http://plato.stanford.edu/archives/fall2012/entries/kant-hume-morality/.

Dodig-Crnkovic, G., & Çürüklü, B. (2012). Robots: ethical by design. *Ethics and Information Technology, 14*(1), 61-71. doi: 10.1007/s10676-011-9278-2

Eshleman, A. (1999). Moral Responsibility. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2009 Edition ed.). Retrieved from http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/.

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines, 14*(3), 349-379. doi: 10.1023/B:MIND.0000035461.63578.9d

Friedman, B., & Kahn Jr, P. H. (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software, 17*(1), 7-14. doi: 10.1016/0164-1212(92)90075-u

Grodzinsky, F., Miller, K., & Wolf, M. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*(2), 115-121. doi: 10.1007/s10676-008-9163-9

Gunkel, D. J. (2012). *A Vindication of the Rights of Machines*. Paper presented at the AISB/IACAP World Congress 2012, Birmingham, UK.

Haksar, V. (1998). Moral agents. *Concise Routledge Encyclopaedia of Philosophy*.

Hanson, F. (2009). Beyond the skin bag: on the moral responsibility of extended agencies. *Ethics and Information Technology, 11*(1), 91-99. doi: 10.1007/s10676-009-9184-z

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99-107. doi: 10.1007/s10676-012-9301-2

Himma, K. (2009). 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. doi: 10.1007/s10676-008-9167-5

Hinton, G. E. (1989). Connectionist Learning Procedures. *Artificial Intelligence, 40*, 185-234.

Hofstadter, D. R. (1999). *Gödel, Escher, Bach: An Eternal Golden Braid* (20th Anniversary Edition ed.). New York: Basic Books.

Hofstadter, D. R. (2007). *I am a Strange Loop*. New York: Basic Books.

Human Rights Watch. (2012). *Losing Humanity: The Case against Killer Robots*: International Human Rights Clinic.

Johansson, L. (2010). The Functional Morality of Robots (pp. 65-73): IGI Global.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195-204. doi: 10.1007/s10676-006-9111-5

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2), 123-133. doi: 10.1007/s10676-008-9174-6

Krishnan, A. (2009). *Killer Robots: Legality and Ethicality of Autonomous Weapons*. Surrey, England: Ashgate.

Kuflik, A. (1999). Computers in control: Rational transfer of authority or irresponsible abdication of autonomy? *Ethics and Information Technology, 1*(3), 173-184. doi: 10.1023/a:1010087500508

Lichocki, P., Kahn, P. H., & Billard, A. (2011). The Ethical Landscape of Robotics. *Robotics & Automation Magazine, IEEE, 18*(1), 39-50. doi: 10.1109/mra.2011.940275

Lin, P., Bekey, G., & Abney, K. (2008). Autonomous Military Robotics: Risk, Ethics, and Design: California State Polytechnic University.

Lucas, G. R. (2011). Industrial Challenges of Military Robotics. *Journal of Military Ethics, 10*(4), 274-295. doi: 10.1080/15027570.2011.639164

Matheson, B. (2012). *Manipulation, Moral Responsibility, and Machines*. Paper presented at the AISB/IACAP World Congress 2012, Birmingham, UK.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183. doi: 10.1007/s10676-004-3422-1

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics*, 21,* 18-21.

Ord, T. (2006). The many forms of hypercomputation. *Applied Mathematics and Computation, 178*(1), 143-153. doi: 10.1016/j.amc.2005.09.076

Paramythis, A. (2004). *Towards Self-Regulating Adaptive Systems*. Paper presented at the Proceedings of the Annual Workshop of the SIG Adaptivity and User Modeling in Interactive Systems of the German Informatics Society (ABIS04), Berlin.

Paramythis, A. (2006). Can Adaptive Systems Participate in Their Design? Meta-adaptivity and the Evolution of Adaptive Behavior *Adaptive Hypermedia and Adaptive Web-Based Systems* (Vol. 4018/2006). Berlin / Heidelberg: Springer.

Parthemore, J., & Whitby, B. (2012). *Moral Agency, Moral Responsibility, and Artefacts*. Paper presented at the AISB/IACAP World Congress 2012, Birmingham, UK.

Powers, T. M. (2006). Prospects for a Kantian Machine. *Intelligent Systems, IEEE, 21*(4), 46-51. doi: 10.1109/mis.2006.77

Ruffo, M.-d.-N. (2012). *The robot, a stranger to ethics*. Paper presented at the AISB/IACAP World Congress 2012, Birmingham, UK.

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd Edition ed.). Upper Saddle River, NJ: Prentice Hall.

Serugendo, G. D. M., Gleizes, M.-P., & Karageorgos, A. (2006). Self-Organisation and Emergence in MAS: An Overview. *Informatica, 30*, 45-54.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy, 24*(1), 62-77.

Stahl, B. C. (2004). Information, Ethics, and Computers: The Problem of Autonomous Moral Agents. *Minds and Machines, 14*(1), 67-83. doi: 10.1023/b:mind.0000005136.61217.93

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology, 8*(1), 205-213. doi: DOI 10.1007/s10676-006-9112-4

Stensson, P., & Jansson, A. (2013). Autonomous technology – sources of confusion: a model for explanation and prediction of conceptual shifts. *Ergonomics, 57*(3), 455-470. doi: 10.1080/00140139.2013.858777

Sullins, J. P. (2006). When Is a Robot a Moral Agent? *International Review of Information Ethics, 6*(12), 23-30.

Swiatek, M. (2012). Intending to err: the ethical challenge of lethal, autonomous systems. *Ethics and Information Technology, 14*(4), 241-254. doi: 10.1007/s10676-012-9302-1

Tonkens, R. (2009). A Challenge for Machine Ethics. *Minds and Machines, 19*(3), 421-438. doi: 10.1007/s11023-009-9159-1

Tonkens, R. (2012). Should autonomous robots be pacifists? *Ethics and Information Technology*, 1-15. doi: 10.1007/s10676-012-9292-z

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & Society, 22*(4), 495-521. doi: 10.1007/s00146-007-0091-8

Trevellyan, R., & Browne, D. P. (1987). A self-regulating adaptive system. *ACM SIGCHI Bulletin, 18*(4), 103-107.

Turing, A. M. (1947 (1986)). Lecture to the London Mathematical Society on 20 February 1947. In B. E. Carpenter & B. W. Doran (Eds.), *Charles Babbages Reprint Series for the History of Computing*: The MIT Press.

Wallach, W. (2008). Implementing moral decision making faculties in computers and robots. *AI & Society, 22*(4), 463-475. doi: 10.1007/s00146-007-0093-6

Williams, G. (2014). Responsibility. In J. Fieser & B. Dowden (Eds.), The Internet Encyclopedia of Philosophy. Retrieved from http://www.iep.utm.edu/.

Wolf, T. D., & Holvoet, T. (2004). *Emergence and Self-Organisation: a statement of similarities and differences*. Paper presented at the 2nd International Workshop on Engineering Self-Organising Applications.

Wolf, T. D., & Holvoet, T. (2005). Towards a Methodology for Engineering Self-Organising Emergent Systems. In H. Czap & R. Unland (Eds.), *Self-Organization and Autonomic Informatics* (pp. 18-34): IOS Press.