Limitations and Risks of Machine Ethics

Miles Brundage

Consortium for Science, Policy, and Outcomes, Arizona State University, Tempe, United States

987 W. Washington St. Apt. E303

Tempe, AZ 85281

Miles.brundage@asu.edu

Limitations and Risks of Machine Ethics

Many authors have proposed constraining the behaviour of intelligent systems with “machine ethics” to ensure positive social outcomes from the development of such systems. This article critically analyses the prospects for machine ethics, identifying several inherent limitations. While machine ethics may increase the probability of ethical behaviour in some situations, it cannot guarantee it due to the nature of ethics, the computational limitations of computational agents, and the complexity of the world. Additionally, machine ethics, even if it were to be “solved” at a technical level, would be insufficient to ensure positive social outcomes from intelligent systems.

Keywords: machine ethics; risk; artificial general intelligence

Introduction

In response to the increasing sophistication and social impact of artificial intelligence-based technologies, some authors have proposed that software agents and robots be imbued with explicit ethical principles to govern their behaviour (Allen & Wallach, 2010). This area of research, known as "machine ethics," incorporates insights from normative ethics, AI, and other disciplines and seeks to instantiate models of ethical reasoning in machines in order to ensure ethical machine behaviour as well as to explore the nature and computability of human ethics. In addition, other authors concerned with the risks posed by future greater-than-human intelligence machines have characterized machine ethics as one among multiple strategies for ensuring positive social outcomes from the creation of such machines. Yudkowsky (2001), for example, argues that failing to develop a sufficiently well-formulated computational instantiation of morality will result in catastrophic outcomes from developing superintelligent machines, since the instrumental goals of self-improving systems seeking to maximize some arbitrary utility function will at some point come into conflict with human interests.

What these literatures (referred to here as "traditional machine ethics" on the one hand, and "artificial general intelligence (AGI) ethics" on the other, though the two have cross-pollinated to some extent) have in common is the apparent belief that the ethical behaviour "problem" can be "solved" to some extent through philosophical and technological innovation. While this article does not seek to undermine the desirability of such a solution, in principle, it points to a number of reasons to believe that such a project will necessarily fail to guarantee ethical behaviour of a given AI system across all possible domains. The intrinsic imperfectability of machine ethics has been suggested by several authors, such as Allen et al. (2000); the present paper seeks to synthesize, deepen, and extend these concerns, and draw out some of their implications for the project of machine ethics viz-a-viz the possible social outcomes from AGI. While such inevitable imperfection may be acceptable to traditional machine ethicists, who have often explicitly acknowledged and accepted that machines, like humans, will inevitably make some mistakes, it presents a fundamental problem for the usefulness of machine ethics as a tool in the toolbox for ensuring positive outcomes from powerful computational agents (Muehlhauser et al., unpublished).

The paper will proceed as follows: first, I will identify a diverse set of inherent limitations on the machine ethics project which come from the nature of ethics, the nature of computational agents in general, and the nature of the world. Next, I will review some specific categories of machine ethics and AGI ethics proposals, finding that they are susceptible to the limitations identified in the first section of the paper. A number of possible failure modes for machine ethics will summarized, i.e. ways in which instantiating ethical principles in a powerful AGI system could result in outcomes that are unintended and possibly even catastrophic. Finally, I will characterize machine ethics as an insufficient (albeit potentially helpful) tool for ensuring positive social outcomes from AGI, since other factors (such as cybersecurity, human decisions, and systemic risks) will also have to be reckoned with in the creation of human-level or greater-than-human intelligence machines.

Limitations from the Nature of Ethics

This section will survey some of the literature in normative ethics and moral psychology which suggest that morality does not lend itself to an algorithmic solution. This is not to say that humans and machines cannot improve their moral behaviour in many situations by following the prescriptions of something akin to an algorithm – indeed, there is widespread agreement on at least some moral rules (Gert, 2007). However, these rules are often ambiguous and should sometimes be broken, and there is persistent disagreement about the conditions in which such exceptions should be made, as well as broad agreement that some specific ethical domains are still problematic despite the best efforts of philosophers. Importantly for the present discussion, this "unsolved" nature of ethics may not be a transient condition owing to insufficient rational analysis, but rather a reflection of the fact that the

intuitions on which our ethical theories are based are unsystematic at their core, which creates difficulties for the feasibility of machine ethics.

Researchers dating back at least to Darwin in *The Descent of Man* (1871) have attempted to explain human moral judgments as deriving, at least in part, from evolutionary processes. Recently, Boeme (2012) has distilled much anthropological and evolutionary psychological evidence to argue that many of our intuitions about morality have their origins in pre-historic humans' and pre-human primates' attempts to suppress anti-social behaviour. Under this reading of the scientific evidence, virtue, altruism, shame, and other morally-salient concepts and emotions exist because those who are alive today descended from humans that adapted, genetically and culturally, to the demands of social cooperation. Findings in the social psychology and neuroscience of morality bear out the deep connection between ethical judgments and evolutionary requirements.

Cushman et al. (2010) articulate an emerging consensus in moral psychology work that much of our moral judgments can be conceptualized through a "dual-process" model. Under this framework, humans rely to varying extents, depending on the situation at hand and factors such as cognitive load, on an intuitive moral system and a more deliberate moral system. Moreover, this dual-process model of moral psychological processes roughly maps onto two major camps in moral philosophy: deontological (means-based) and consequentialist (ends-based) decision-making. Given these distinct ways by which humans arrive at moral judgments, and the fact that ethics is, at least in part, a project aimed at systematizing moral judgments into a rational framework, dual-process moral psychology is of great importance to machine ethicists. These findings are recent and have been contested on grounds ranging from their empirical bases (Klein 2011) to their normative implications (Berker, 2009), but at the moment, they seem to offer a plausible explanation for the persistence of ethical quandaries and the difficulty of articulating an exceptionless and well-specified moral theory.

Having assessed some empirically-based reasons to suspect that human moral judgments may not lend themselves to a consistent computational specification, the remainder of this section will consider barriers to reliable machine ethics that are discussed in the literature on ethics proper without commenting on the underlying moral psychology involved in these issues. Comprehensive moral theories that have been put forward so far are inadequate in varying ways and to varying extents. An exhaustive survey of the state of the debate in normative ethics is beyond the scope of this paper, but a few brief points on the two leading branches of ethical theories (deontological and consequentialist) will illuminate some of the challenges facing a hypothetical machine ethicist seeking to instantiate an "off-the-shelf" ethical theory in a machine (issues facing "bottom-up" approaches based on, ex. case-based based reasoning and machine learning will be discussed later).

Consequentialist theories of ethics have been criticized as inadequate for a variety of reasons (Pojman, 2005). These include the claim that they cannot sufficiently account for the moral significance of an individual's personal social commitments (ex. to friends and family) and life projects; that they impose excessive demands on individuals to contribute to others' welfare (Wolf, 1982); that it seems to arrive at unacceptable conclusions in certain situations; and that it fails to give sufficient consideration to the separateness of persons, distributional justice considerations, and individual rights (Williams, 1973). Some of these objections have been leveled at consequentialism in general, and others at specific variations thereof; additionally, a wide variety of consequentialist theories have been developed to attempt to grapple with the concerns that have been raised, but none have achieved anything remotely resembling a consensus in philosophy.

Deontological theories of ethics also have their limitations. Some notable critiques include the argument that deontological theories give the wrong answers in situations involving extreme trade-offs between the interests of the few and the interests of the many, and can thus produce catastrophic moral results; that deontological theories cannot adequately resolve conflicts between duties; and that deontology collapses into consequentialism since an actor who opposes producing harm X is rationally committed to reducing the amount of X in the world.

Each of the theories thus described has articulate defenders, as do other ones that attempt to synthesize the insights of each (Parfit, 2011). However, for the purposes of this discussion, I seek merely to highlight that a machine ethicist seeking for input from normative ethicists will receive a wide range of (often heavily qualified) responses, and to the knowledge of the author, each proposed solution is likely vulnerable to at least one of the objections above. This pervasive disagreement about and unease with comprehensive moral theories has been noted by many authors (Gert, 2007) – indeed, even the defenders of specific theories mentioned above typically note the difficulty of their respective frameworks in addressing certain issues and the need for further analysis. An additional overarching issue facing comprehensive moral theories is ambiguity. While moral theories have varying degrees of specificity in terms of both their criteria of moral rightness and proposed decision-procedures, they must be augmented by knowledge of and sensitivity to the relevant domains in which an individual finds him or herself. Ethics on the battlefield are very different from ethics in the shopping mall – thus, while applied ethical decisions can be loosely inspired by overarching moral theories, much work remains to be done in order to flesh out what, ex., pleasure, welfare, beneficence, or non-malevolence actually mean in a particular domain, and this derivation and specification introduces new challenges (some of which will be discussed in the next section).

Another challenge to ethical systematization is posed by the plurality of values which arguably should (and empirically do) motivate people. This issue is related to, but can also be distinguished from, the objections to comprehensive moral theories summarized above. Berlin notes the cultural and historical variation in the values that people do, in fact, hold (Berlin, 1991), and argues that the move to encompass all of what ought to be done under the umbrella of a single value has historically been associated with oppression and violence. Ross (1988), whose ethical views (drawn on by some machine ethicists) do not fit cleanly into either the deontological or consequentialist categories, argues based on an analysis of ethical intuitions that multiple prima facie duties exist, none of which is reducible to the others. The challenge of value pluralism within comprehensive moral theories has also been noted: ex. Shaw (1999) surveys the literature on utilitarianism (a variant of consequentialism) and finds no clear answer to the question of what it is that ought to be maximized by a utilitarian, i.e. pleasure, welfare, the satisfaction of preferences, etc. Shulman et al. (2009) arrive at a similar conclusion. Haidt (2012) argues based on consequentialist grounds that in the public domain, people should invoke a plurality of moral "foundations," some oriented more towards individual autonomy and others more oriented towards group cohesion and welfare, since each on their own adds something important to normative discourse. Similar arguments have been made for thousands of years, in both secular and religious traditions (see, ex., the 10 Commandments, Aristotle's virtues, etc.) and incorporating these insights into a machine ethics system would seem to entail unpredictable behaviour or paralysis in the case of value trade-offs.

Next, regardless of whether a comprehensive moral theory is sufficient to capture the "right" answers to a wide range of ethical, there remain some unresolved issues in ethics that must be grappled with if ethical machines are to be deployed in a flexible way across many domains. These issues inspire many of the counter-examples often used against comprehensive moral theories, and are also investigated in and of themselves by ethicists, without clear resolution to date. Crouch (2012) highlights several of these "unsolved problems" in ethics, including population ethics, small probabilities of huge amounts of value, issues associated with the possibility of infinite value, moral uncertainty, the relationship between intuitions and theoretical virtues, prevention of wrongs versus alleviating suffering, and the assignment of moral value to various entities. While many philosophers have preferred answers to these questions, these issues remain deeply controversial and a machine making decisions in an open, complex environment over the long-term will need to grapple with them and may arrive at conclusions that, based on the present understanding of ethicists, will be vulnerable to compelling objections.

Finally, the (possible) existence of genuine moral dilemmas is problematic for the machine

ethicist. The theoretical and actual possibility of genuine moral dilemmas is a matter of disagreement among ethicists (Gowans, 1987). On the one hand, some argue for the existence of such dilemmas based on some of the difficulties discussed earlier (such as value pluralism and conflicts within and between ethical theories (Sinnott-Armstrong, 1988) and on an argument for rational regret. Williams (1973) argues that even when one is confident that one has made the right ethical decision, doing so often requires making trade-offs between fundamentally distinct considerations, leaving an ethical "remainder" that can rationally be regretted since it was a different sort of thing than that which was ultimately prioritized. In contrast, others argue based on the principles of deontic logic and other considerations that moral dilemmas are not logically possible within a rational theory of ethics. Yet whether such dilemmas can be considered "genuine" or merely perceived as such from an internal or external perspective, conflicts between duties, values, and interests are commonplace in real-life situations that ethical machines will find themselves in, and as such will pose difficult challenges for creating systems that will be *perceived as* having acted morally after the fact.

Limitations Arising from Bounded Agents and Complex Environments

Given an arbitrary ethical theory or utility function, a machine (like a human) will be limited in its ability to act successfully based on it in complex environments. This is in part due to the fact that ethical decision-making requires an agent to estimate the wider consequences (or logical implications) of his or her actions, and possess relevant knowledge of the situation at hand. Yet humans and machines are limited in their perceptual and computational abilities, and often only have some of the potentially relevant information for a given decision, and these limitations will create possible failure modes of machine ethics.

Computational and knowledge limitations apply to ethics in different ways depending on the ethical theory involved. Consequentialist theories, for example, are quite explicitly dependent on the ability to know how one's actions will affect others, though there is still much heterogeneity here, such as the distinction between objective utilitarianism (which prescribes acting in a way that, in fact, maximizes good outcomes) and subjective utilitarianism (which emphasizes the expected outcomes of one's actions). Deontological theories, in contrast, are not explicitly about foreseeing the outcomes, but computational limitations are related to deontology in at least two ways. First, to even know in the first place that a given action is, for example, consistent with a given deontological duty may require some knowledge and analysis of the situation and actors involved, not all of which will necessarily be apparent to the actor. Second, some deontological theories and rule consequentialist theories (which prescribe acting on the set of rules that, if universally accepted or adopted, would lead to the best outcomes) require judgments about the logical implications of a given decision if it were to be performed by many or all actors, i.e. the universalizability of a moral judgment. As such, the knowledge and information processing limitations of an artificial agent will be relevant (though perhaps to varying degrees) regardless of the specific ethical theory invoked.

There is wide recognition in AI, cognitive science, and other fields that humans and machines are not able to act in a fully rational manner in non-trivial domains, where this is defined as maximizing a given utility function by processing information from the information and acting in one's environment based on it, although there are efforts to move closer and closer to approximating such rationality. This inherent limitation is due to the potentially infinite courses of action available to an agent, and the consequent inability to exhaustively analyse all of them; the inability to obtain and store all potentially relevant facts; and the intrinsically uncertain relationship between chosen actions and consequences when the environment has complex dynamics. Consequently, some have suggested that concepts like bounded rationality (acting more or less rationally given one's computational limitations) and satisficing (acting in a way that leads to satisfactory or "good enough" outcomes) are better models for thinking about decision-making by agents such as humans and machines. Indeed, Wang argues that

the ability to act on the basis of insufficient knowledge and resources is the essence of intelligence (Wang, 2006). Gigerenzer (2010) and others note that heuristics, while deviating from normative theories of rationality, in fact work quite well in many situations in which we find ourselves, i.e. they have adaptive value. The literature on human cognitive heuristics may seem removed from machine ethics at first glance - after all, as Yudkowsky (2007) and others note, the goal of AI is not merely to replicate human behavioural patterns but to improve upon them. Bu this is not the case. Insofar as humans serve as the "proof of concept" for intelligence and some AI theorists seek to develop cognitive architectures that are inspired by human cognition, we may ultimately see ethical machines that exhibit similar cognitive heuristics to humans. Additionally, since machines will learn and evolve in similar environments (namely, complex social ones) to those in which humans evolved, even if AI does not explicitly seek to model human behaviour, some of the same heuristics that benefitted humans may ultimately creep in. Regardless, the points made here do not depend on the existence of any particular heuristic being instantiated in a machine, but rather it suffices to note that machines are and will remain limited in their possession and processing of knowledge.

Computational and knowledge limitations enter into ethical decision-making in several ways. First, one can think of the drawing up of a list of possible actions as a search problem. Under this framing, it is impossible to exhaustively search the space of all possible actions in any non-trivial domain that could lead to better and better outcomes with regards to a specific ethical theory, though this does not preclude responding to binary ethical decisions. Heuristics will be required to select a set of actions to evaluate, and while these may be useful in many situations (as they are for humans), important options for navigating moral dilemmas may consequently be overlooked (Gigerenzer, 2010). Gigerenzer argues that satisficing is the norm in real-life ethical decision-making. Second, given a particular problem presented to an agent, the material or logical implications must be computed, and this can be computationally intractable if the number of agents, the time horizon, or the actions being evaluated are too great in number (this limitation will be quantified later and discussed in more detail later in the section). Specifically, Reynolds (2005, p. 6) develops a simple model of the computation involved in evaluating the ethical implications of a set of actions, in which N is the number of agents, M is the number of actions available, and L is the time horizon. He finds:

> It appears that consequentialists and deontologists have ethical strategies that are roughly equivalent, namely $O(MN^L)$. This is a "computationally hard" task that an agent with limited resources will have difficulty performing. It is of the complexity task of NP or more specifically EXPTIME. Furthermore, as the horizon for casual ramifications moves towards infinity the satisfaction function for both consequentialism and deontologism become intractable.

While looking infinitely to the future is an unreasonable expectation, this estimate suggests that even a much shorter time horizon would quickly become unfeasible for an evaluation of a set of agents on the order of magnitude of those in the real world, and as previously noted, a potentially infinite number of actions is always available to an agent. This argument has also been made qualitatively in (Allen et al., 2000) and (Allen and Wallach, 2010). Goertzel has also outlined a similar argument to Reynolds's (2006).

Lenman (2000) has spelled out the implications of such limitations for consequentialism, arguing that a person is only aware of a small number of the possible consequences of a given decision, and that the longer-term consequences may in fact dominate the ultimate ethical impact, making consequentialism impossible to fully adhere to – for example, a seemingly trivial decision made while driving that slightly affects traffic patterns could have implications for who ends up meeting whom, when children are conceived and their consequent genetic make-up, etc. which are far more significant than the driver's immediate impacts, yet are utterly impossible to foresee and act upon. Third, an agent

could lack knowledge that is of great ethical significance for the decision. This is related to the "frame problem" in AI, and has been extended to the domain of ethics by Horgan and Timmons, who note that small changes in the context of a given decision could have great ethical import yet cannot be captured in a single ethical principle (Horgan and Timmons, 2009).

Assuming that an agent had a well-specified utility function for ethics and exhibited normative rationality in dealing with ethical dilemmas, the agent would nevertheless run into problems in complex task domains. Social systems exhibit many of the characteristics of complex systems in general, such as non-linear interactions, heavy-tail distributions of risks, systemic risks introduced by failure cascades in coupled systems, self-organization, chaotic effects, and unpredictability (Helbing, 2010). These are not avoidable features of an agent making assessments of the systems, but rather are fundamental to the complex world we inhabit. In addition, the adoption of technologies adds additional elements of complexity to social analysis, since technologies influence social outcomes at multiple scales that are not reducible to a simple predictive model (Allenby and Sarewitz, 2011). These complexities pose challenges for ethical decision-making by making the system being analysed beyond the knowledge and computational scope of any agent making the decision – that is, any assessment by an artificial agent such as, ex,. that buying stock in a particular company would be beneficial for its owner, or that diverting a train to save the lives of several people standing on a train track at the expense of one person, are necessarily no more than educated guesses that could be wrong in the face of "black swan" events that fly in the face of predictive models that made sense beforehand (Taleb, 2007). Additionally, an agent could be ignorant of the fact that he or she will have a non-linear effect on a complex system – such as, ex., that selling a certain stock at a certain volume, combined with other trading activities, will induce a "tipping point" and runaway sales of that stock, with ethical ramifications dwarfing that of the originally intended sale. Lack of knowledge about the intentions of other agents, too, poses a challenge, but even given such knowledge, a superintelligent agent could not solve the predictability problem: for example, using a relatively simple model of a multi-agent system (compared to the real ones we inhabit), researchers (Tosic and Agha, 2005) have found that:

> counting the number of possible evolutions of a particular class of CFSMs [communicating finite state machine] is computationally intractable, even when those CFSMs are very severely restricted both in terms of an individual agent's behaviour (that is, the local update rules), and the inter-agent interaction pattern (that is, the underlying communication network topology).

Discussing such complexities, Allenby notes that "CASs [complex adaptive systems] thus pose a challenge to the existing ethical approaches most familiar to scientists and technologists" (Allenby, 2009, p. 2). While these points suggest difficulties for any ethical agent, they particularly pose a challenge for an artificial agent seeking to act across multiple domains and to intervene at a large scale in complex systems – for example, by seeking to maximize some global utility function such as human welfare, rather than more limited action at a small scale.

Two final points can be made about the limitations of machine ethics viz-a-viz action in a social environment. First, as Müller notes (2012), to improve their effectiveness, artificial agents will need to learn and evolve over time, which requires more autonomy and, correspondingly, less human control of the agent in question. Likewise, Wallach and Allen (2010) and others have noted the importance of integrating top-down and bottom-up approaches to machine ethics, the latter requiring learning from specific cases and experiences. Representative training data and experiences of an artificial agent will thus be essential for the "successful" development of machine ethics. Yet it is impossible to expose an agent to an infinite array of possible ethical situations, and care will be needed to ensure that the lessons being learned are those intended by the designers of the system. Additionally, since the ethical principles appropriate to a given domain are often unique, it is unclear when and how one could determine that an ethical training process were ever "complete." Second, the complexity of artificial agents may pose challenges to humans' ability to understand, predict, and manage them. Already, there are technological systems that elude human understanding, and the development of human-level or greater-than-human intelligence systems would exacerbate this problem and pose difficulties for the allotment of trust to such systems.

Current Approaches to Machine/AGI Ethics

This section will summarize some elements of the machine and AGI ethics literatures and note how current approaches are vulnerable to the limitations outlined above. Notably, few of these approaches have been suggested by researchers as the be-all-end-all for machine ethics – rather, the field is at an early stage, and these approaches are mostly being presented (by the researchers behind them, and by me) as illustrative of the issues involved in developing a computational account and implementation of ethics.

Following the distinction made by Wallach and Allen (2010), I will first discuss some "top-down" approaches to machine ethics which involve programming a specific theory of morality in a machine, which is then applied to specific cases (as opposed to "bottom-up" approaches, which includes case-based reasoning, artificial neural networks, reinforcement learning, and other tools to progressively build up an ethical framework). Besides top-down and bottom-up approaches, I will also cover some hybrid approaches, psychological approaches (which seek to directly model the cognitive processes involved in human ethical decision-making), and AGI ethics proposals, which unlike the other methods discussed, are specifically aimed at constraining the behaviour of human-level or greater-than-human-level intelligence systems.

Some representative top-down machine ethics approaches that have been theorized and/or built in prototypes are Cloos's Utilibot (Cloos, 2005), Powers' Kantian approach (Powers, 2011), Bringsjord et al.'s category theory approach (Bringsjord et al., 2011), Mackworth's constraint satisfaction approach (Mackworth, 2011), and Arkin's Ethical Governor for lethal autonomous robots (Arkin, 2009). Beginning with Cloos's Utilibot - utilitarian approaches to ethics in general were discussed in an earlier section of this paper, but to reiterate, there are many reasons to think that utilitarianism does not capture the whole of our ethical intuitions. Furthermore, it is highly vulnerable to problematic conclusions on unsolved problems in ethics such as population ethics and questions of moral status, and thus a utilitarian robot making decisions outside of narrowly scoped domains (such as those which Cloos discusses, like healthcare) would be potentially dangerous. Grau (2006) notes that in the movie

based on Asimov's "I, Robot," the robots arguably treated humans in a utilitarian fashion, and that this "cold" rationality was used as the basis for humanity's loss of control over machine systems. While that particular outcome seems unlikely (at least in part because so many people are so familiar with such scenarios in science fiction), this example illuminates a general concern with utilitarian approaches to machine ethics – that the maximizing, monistic nature of such ethical frameworks may justify dangerous actions on a large scale.

Different but related issues apply to Powers' Kantian approach (2011), which inherits the problems associated with deontological ethical theories, including most notably the potential for catastrophic outcomes when deontological rules should, in fact, be broken (Bringsjord, 2009). Bringsjord also discusses a category theory approach, which would allow a machine to reason over (not merely within) particular logics and formally guarantee ethical behaviour (Bringsjord, 2011). While this approach may be a suitable representational system, it is agnostic as to the ethical system being implemented, which is the present concern of this paper. Additionally, Bringsjord notes a significant limitation of the approach – the absence of a psychological account of other's intentions, which Bringsjord and Bello (2012) argue is necessary, and will be discussed later in this section. Likewise, Mackworth's constraint satisfaction approach may capture some of what is involved in moral cognition (namely, making sure that various moral constraints on action are satisfied) but the devil is in the details (Mackworth, 2011). Attempts to spell out such constraints, or to suitably prioritize them, may lead to perverse outcomes in certain situations and fall victim to the "one wrong number" problem discussed by Yudkowsky (2011), i.e. how the absence of one particular consideration in a moral system could lead to dangerous results even if the rest of the system is well-developed, analogous to the way that one wrong digit in a phone number makes the resulting number useless.

Finally, Arkin's Ethical Governor approach would constrain the behaviour of lethal autonomous robots by representing the Laws of War (LOW) and Rules of Engagement (ROE). While this approach is specifically for military combat situations, not ethical action in general, it is arguably the most rigorously developed and explicated approach to constraining robot behaviour with ethical considerations and merits discussion. While philosophical arguments have been raised against the Ethical Governor approach (Matthias, 2011), here I will focus on pragmatic considerations. As noted by roboticist Noel Sharkey (Human Rights Watch, 2012), today (and, he argues, for the foreseeable future) it is not possible for machines to reliably discriminate between combatants and non-combatants, thus making the LOW and ROE impossible to apply in a fool-proof way by robots. This is not merely a quibble with the state of the art that may someday change; rather, it is well-known that even humans make mistakes in conflict situations, and this may be a reflection of the knowledge and computational limitations of finite agents rather than a solvable problem. A combat version of the "frame problem" may apply here: in addition to combatant/non-combatant distinctions, features of situations such as whether a building is a church or a hospital, whether a woman is pregnant, etc. all bear on the consistency of an action with the LOW and ROE yet are not necessarily amenable to algorithmic resolution by humans or machines, and there are many possible reasons why a rule should be broken in a given situation, such as for consequentialist reasons (as humans sometimes do in similar situations). None of this should be seen as arguing that autonomous lethal robots are necessarily unethical to develop – though some have made such an argument (Human Rights Watch, 2012). However, the point to be emphasized here is that even in a situation with relatively well-specified ethical constraints, there does not appear to be a possibility of computationally solving the problem of ethical behaviour in a fool proof way, which should give one pause regarding the prospects for generally intelligent, reliably moral agents.

Next, bottom-up and hybrid approaches will be analysed. Some examples of strictly bottom-up

approaches to machine ethics include Guarini's artificial neural networks (Guarini, 2011) and McLaren's SIROCCO and Truth-Teller systems (McLaren, 2011). In Guarini's experiments, he has trained an artificial neural network on specific ethical judgments in order to model them with that network. Likewise, McLaren's SIROCCO and Truth-Teller systems use case-based reasoning in order to learn about the morally relevant features of such cases and make future judgments. Notably, bottom-up approaches to machine ethics are highly dependent on the training data used, and thus the ethical values of those humans training such systems cannot be separated from the computational framework being trained. Nevertheless, a few things can be said about the limitations of these approaches. As Wallach and Allen note (2010), bottom-up ethical approaches appear to be less "safe" than top-down or hybrid approaches in that they lack assurances that any particular principle will be followed. Moreover, the process of training such a system (particularly an embodied robot that is able to take morally significant actions) may introduce risks that are unacceptable, which would presumably increase as the cross-domain flexibility and autonomy of the system increases. Computational limitations may pose problems for bottom-up approaches, since there could be an infinite number of morally relevant features of situations, yet developing tractable representations will require a reduction in this dimensionality. There is thus no firm guarantee that a given neural network of case-based reasoning system, even if suitably trained, will make the right decision in all future cases, since a morally relevant feature that didn't make a difference in distinguishing earlier data sets could one day be important. Finally, the presumption behind bottom-up approaches, namely that learning based on human judgments will lead to a useful framework for making future decisions, may be fundamentally flawed if there exists no cluster of principles consistent with human intuitions that is not self-contradictory, inconsistent, or arbitrary in some way.

Hybrid systems for machine ethics have also been proposed. Anderson and Anderson (2011) have developed systems incorporating ideas from W.D. Ross's prima facie duty approach to ethics. In this framework, several duties can bear on a given situation, and by default a decision-making should seek to avoid the violation of any duties, but sometimes moral conflicts can occur and a judgment must still be made. Using machine learning, Anderson and Anderson have discovered a decision procedure for prioritizing prima facie duties in a particular domain (medical ethics) that previously had not been articulated, yet conforms to expert opinion. This hybrid approach (incorporating explicit normative principles form the philosophical literature, while using machine learning to aid in the prioritization of these principles) has significant promise in narrow domains, but it has some of the same limitations discussed earlier in the section on bottom-up approaches. For example, there is no reason to assume a priori that any suitable hierarchy of principles will necessarily be found across a wide range of situations – Anderson & Anderson seem to agree, noting that autonomous systems should only be deployed in situations where expert opinion has reached a consensus on the relevant ethical issues involved. Likewise, Pontier and Hoorn (2012) developed the Moral Coppelia, an approach that integrates concepts from connectionism, utilitarianism, and deontological approaches to ethics. Pontier and Hoorn present intriguing results and have moved towards systems that take dual-process theories of moral judgment seriously, but there is again no guarantee that the resulting amalgam of principles will end up being coherent.

Next, Gomila (2009), Deghani et al. (2011), and Bringsjord and Bello (2012) describe machine ethics approaches involving the direct modelling of human cognitive and emotional processes. While it was argued earlier that moral views depend in part on intuitive/emotional processes, and thus all machine ethics approaches could, in some sense, be seen as indirectly modelling human psychology, these approaches take a more direct route by drawing specifically on cognitive science and neuroscience findings rather than starting from an ethical theory. Gomila focuses on the role of emotions in moral cognition and ways this can be instantiated computationally. Deghani et al.

developed MoralDM, which possesses some features similar to Pontier and Hoorn's approach (specifically, it integrates both utilitarian calculations and more deontological ones such as sacred values). Finally, Bringjsord and Bello focus on the role of theory of mind (ToM) in moral psychology, and argue that an appropriate computational account of ToM is necessary in order to reason about moral implications of one's actions. Additionally, they make an argument for the importance of taking folk moral psychology seriously in machine ethics, as opposed to waiting in vain for a solution to ethics by philosophers. While these efforts are important steps in the direction of scientific accounts of moral judgment, there does not appear to be any reason to believe the resulting systems will be able to act reliably ethically. The problem of human ethics is not merely that we don't always follow our own ethical judgments - the problem is also that we don't have a clear understanding of what such ethical perfection could entail, and dual-process theories of moral judgment and the current state of the normative ethics literature suggest the pursuit of a deterministic algorithm may be fruitless.

Moreover, Savulescu and Persson (2012) note many characteristics of folk morality that appear unsuitable for the modern techno-human condition, such as a high discount rate, little regard for distant and unrelated people, and distinctions based on (arguably) morally irrelevant features such as inflicting harm with or without direct physical contact. One could respond that with a computational account of folk psychology, we could then seek to address these deficiencies by tweaking the program. However, this would seem to bring one back to where we started, which is that solving that problem requires a well-specified understanding of what ethical behaviour requires in the first place and how to improve on our intuitions, and the psychological approach does not appear to offer a way around this problem or the knowledge and computational limitations discussed earlier.

Finally, several approaches categorized here as AGI ethics have been put forward. I will discuss three in particular that are characteristic of the literature: Coherent Extrapolated Volition (Tarleton, 2012), Compassion/Respect (Freeman, 2009), and Rational Universal Benevolence (Waser, 2011). Coherent Extrapolated Volition (CEV) is an approach based on the argument that human values are not well-specified as-is, but that upon substantial reflection by a superintelligence, our values may cohere to some extent and that these convergent extrapolated values are what an AGI should seek to realize in the world. This approach has many desirable features – one is that it can be seen as broadly democratic, in that it takes into consideration all human desires (though not all of them may fit into a coherent whole, so there are potentially important considerations related to how minority rights would fit into such a framework). Additionally, the specification that we should realize the values that we would have "if we knew more, thought faster, were more the people we wished we were, had grown up farther together" seems to get around concerns that could be levelled at other approaches, such as utilitarianism, that could lead to outcomes that we would not in fact want.

Nevertheless, CEV has various limitations. First, it makes a number of normative assumptions that make it vulnerable to objections from within the normative literature, such as rejecting objective theories of morality, in which some things are good regardless of whether or not they're desired, such as that put forward by Parfit (2011). It is also implicitly monistic in the sense that it proposes CEV as the sole criterion of moral rightness, and makes no room for, ex, prima facie duties or individual rights, or even the possibility of moral dilemmas more generally. Also, the notion that what we would want upon reflection should be prioritized over what we, in fact, want is a controversial claim in normative ethics. This is not to suggest that all of these assumptions are false, but simply that CEV does not avoid, but rather is deeply bound up with, existing controversial debates in normative ethics. Note that the Machine Intelligence Research Institute does not characterize CEV as necessarily being the one right ethical approach, but rather, an attempt at ensuring a safe and desirable outcome for humanity. Nevertheless, the concerns raised here apply both to CEV as a criterion of moral rightness and as a

decision procedure for selecting safe actions.

Second, CEV appears to be computationally intractable. As noted earlier, Reynolds' analysis finds that ever larger numbers of agents and decision options, as well as ever longer time horizons, make ethical decision-making exponentially more difficult. CEV seems to be an unsolvable problem both in that it has an unspecified time horizon of the events it considers, and in the sense that it is not clear how much "further" the modelled humans will need to think in the simulation before their morals will be considered sufficiently extrapolated.

Third, even if extrapolation of ethical judgments is normatively plausible in some sense, it may be unacceptable for other reasons. For example, consider the broadly accepted consideration that people should have some say in their lives. If our volitions are extrapolated into the future, our extrapolated selves may come to conclusions that are unacceptable to our current selves, and particularly to disenfranchised groups whose views will not "win out" in the process of making our volitions coherent. Consider again the "I, Robot" scenario, which would hopefully be avoided by CEV as opposed to utilitarianism. But what if it is actually the case that humans, upon sustained critical reflection, will conclude that the robots, in fact, should be allowed to take over the world and make decisions for humans? Muehlhauser et al. (unpublished) also note that some humans exhibit beliefs that could lead to what most people consider catastrophic outcomes, like human extinction (an example being philosopher David Benatar, who argues, without any glaring logical contradiction, that humans should allow themselves to die out gradually in light of the suffering associated with existence). Maybe we would accept such views upon significant reflection, and maybe we wouldn't, and the actual accuracy of such views is not being judged here, but in practice, this uncertainty regarding our ideally extrapolated beliefs may be enough to prevent public acceptance of any such regime (even if it had a strong normative argument in its favour).

Next, there is no reason to assume that any given person, let alone all people, will be coherent under reflection. As emphasized throughout this paper, ethical disagreements and incoherencies abound, and the extent to which humans lack clearly ordered and self-consistent preferences has been well-documented. Given this, the notion of a coherent extrapolated volition of humanity, while desirable in theory, may not actually be possible. The proponents of CEV have noted the possibility of conflicting volitions leading to difficulty of decision-making, and that bringing humans into the loop is an appropriate solution to this problem. While this may help, the arguments put forward in this paper suggest that such conflicts could be so crippling as to not let CEV be useful in making very many decisions at all. Finally, the very process of attempting to extrapolate the volition of humankind could entail acquiring a massive amount of computational resources and/or invading the privacy of humans as instrumental goals along the way to CEV (Omohundro, 2008). While an omniscient CEV-based system would be able to optimally allocate its time and resources between calculating CEV and attaining more computational resources, before such a superintelligent final state, early AGI systems could paradoxically perform unethical actions along the way to refining their understanding of ethics.

Next, Freeman has articulated a relatively simple (only 1000 lines of code in Python) theory of how an AGI could be motivated based on something akin to "compassion and respect." Under this framework, a machine would be trained on past observations of the world and would learn the preferences of the people inhabiting it, and would seek to maximize the sum of the utility functions of people. Like CEV, this theory is not neutral viz-a-viz ethical theory, but instead reflects utilitarian assumptions and as such is vulnerable to arguments against utilitarianism. Nevertheless, a few specific comments can be made about this proposal. First, as noted by Freeman, the Compassion/Respect approach is specifically designed to work with infinite computing resources, and as such it is unclear

whether any suitable approximation could be developed for actual systems. Second, inferring utility functions from behaviour is problematic. While extrapolation-based approaches like CEV also have their problems, it is also the case that taking people's desires at face value is an insufficient basis for a moral system used by an extremely powerful agent. For example, imagine that the majority of the world exhibits some prejudicial tendencies directed towards a certain group of people. Such a system, based solely on maximizing global utility, could exterminate those people in attempt to appease the prejudiced majority. Third, it is not clear that the rhetorical move made by Freeman from maximizing people's utility functions to characterizing this as akin to "compassion and respect" is appropriate. As noted in the section on ethics, utilitarianism has been judged by some as not taking sufficiently seriously the integrity and separateness of persons. A strictly maximizing theory such as Freeman's Compassion/Respect approach would seem to lead to many counter-intuitive decisions such as the sacrificing of an innocent person to provide organs to five others who need organ transplants (in a common counter-example to consequentialism), and Freeman's approach gives no clear way to reason about the trade-offs between such utility maximizing decisions and the broader utilitarian implications of acting on such "cold" logic, which also may be computationally intractable.

Finally, Waser defends an approach entitled Rational Universal Benevolence (RUB), which he argues is grounded in evolutionary theory and "simpler, safer, and wiser" than Friendly AI/Coherent Extrapolated Volition. Rather than seeking to create a purely selfless utility-maximizer, Waser argues that we should develop systems that find cooperation to be in their own self-interest (since, in fact, it is) and that see the flourishing of a cooperative society as a motivating goal. Such a system would conclude that actions "defending humans, the future of humankind, and the destiny of humankind" are rationally required and would carry them out. It should be fairly clear from the earlier discussion that there are also computational issues with Waser's approach, and there is a fine line between rational universal good and rational universal evil. Indeed, given the potential for ethical "black swans" arising from complex social system dynamics, it will not always be clear how a RUB system would act in the real world, and what the actual consequences would be. Furthermore, as Shulman argues (Shulman, 2010), the propensity of an agent to cooperate depends, among other factors, on the probability it perceives of being able to attain resources through cooperation vs. conflict. Thus, Waser's proposal may make sense in particular situations where cooperation is, in fact, desirable for the agent, but may fail if the intelligence or power differential becomes so great that an agent decides to take over the world and maximize its own utility and that of its descendants, which it perceives as having more potential value than that of existing humans (see, ex., Robert Nozick's famous Utility Monster as an example of such logic).

The Insufficiency of Machine Ethics in Guaranteeing Positive AI Outcomes

Even if the machine/AGI ethics problem were "solved," the deployment of AGI systems on a large scale may not actually lead to positive social outcomes. First, one can imagine a malevolent human training a machine on experiences that would lead to the development of a warped sense of morality, akin to the ways in which cults limit the information available to their members. Absent a totalitarian state in which academic freedom and the ability to tinker with one's purchased technologies are eliminated, people will be able to reprogram machines to fit their preferences. As such, a "technical" solution to machine ethics may mean little in a world in which unethical humans exist and have access to advanced technology.

A second related concern is that even if the software problem is solved, cybersecurity vulnerabilities on the hardware side may lead to unethical outcomes from advanced AGI systems. In

humans, the concepts of benevolence, competence, and integrity are often distinguished in the context of trust (Mayer et al., 1995). These terms may not mean the same thing in machines as they do in humans, but some interesting lessons can be learned. Humans have a background set of psychological traits which are typically assumed by default in trusting relationships, including empathy, and it is hoped that they will not be steered away from benevolent behaviour by malevolent individuals, so the strength of one's moral convictions/integrity is crucial for establishing trust. One does not expect a human to instantaneously become malevolent after a lifelong pattern of benevolent behaviour (although it should be noted that the environment matters greatly in determining whether people will violate ethical constraints, even when all involved people do, in fact, have empathy). Insofar as there is a machine analogue of moral integrity, it is cybersecurity or information assurance – that is to say, a virtual or physical autonomous system could be manipulated into actions with harmful consequences. Insofar as the consequences of a machine's actions and existence reflects back upon it as a moral agent, it can be seen as having more or less integrity to the extent to which its software and hardware are immune to corruption by outside sources. In other words, a system that is "unhackable" will be and should be considered as more likely to be sustainably benevolent than one which is vulnerable to hacking, but given the frequency of news reports about incursions into supposedly secure information systems, machine ethics needs to be considered in the context of the security of the broader human-information technology ecosystem.

Third, equity issues may arise in the deployment of AGI systems. Such systems, which could assist humans in attaining arbitrary goals, would be of great personal and professional value. Yet the history of technological diffusion suggests that there is no reason to assume that all will benefit equally, or even to any extent, from a given technology (see, ex., the continuing lack of access to purified water and electricity by billions of people today). Thus, even if machines act in ethical fashions in their prescribed domains, the overall impact may not be broadly beneficial, particularly if control of AGI technologies is concentrated in already powerful groups such as the militaries or large corporations that are not necessarily incentivized to diffuse the benefits of these technologies.

Fourth, risks may arise at the system level that are not apparent or relevant to the decision-making of individual AGI's. For example, if Hanson (2001) and Bringsjord and Johnson's (2012) analyses are correct, then ubiquitous AGIs could lead to a decline in human wages that is not necessarily intended by any particular AGI or human. Also, unintended systemic effects may arise due to ubiquitous ethical actions carried out by previously amoral artefacts being used by self-interested humans, akin to the unintended effects that can arise from ubiquitous defecting behaviour by humans (i.e. there could be some sort of "reverse tragedy of the commons" arising from inefficient social coordination). For example, it is not necessarily clear that explicitly altruistic behaviour by all agents will necessarily be in the best interest of society since many social institutions (including how corporations do business) are based on action according to a profit motive. Unanticipated herding or other interactive effects between AGIs could also arise with unintended negative consequences, such as the systemic risks that have arisen due to the adoption of high-frequency machine trading in the stock market.

Before proceeding to the conclusion, one final comment should be made about the place of machine/AGI ethics in a broader portfolio of approaches for ensuring positive outcomes from AGI (Muehlhauser et al., unpublished). The Machine Intelligence Research Institute has argued that because of the possibility of an intelligence explosion, developing Friendly AI (their term for what, in the parlance of this paper, is a reliable variant of machine ethics) is an urgent engineering challenge. While this paper makes no judgment about the plausibility of an intelligence explosion, our response to such a possibility should be informed not just by what would seem, in principle, to help solve the problem

(such as Friendly AI) but what is foreseeable given current values and technology. All of the considerations discussed so far in this paper seem to point in the direction of Friendly AI (in the sense of a machine that is guaranteed to act in an ethical fashion at a large scale for a long period of time) being unattainable. Thus, if this is true and the argument for the plausibility of an intelligence explosion is also true and that this carries substantial risks, then more attention should be given to alternative AGI risk approaches such as hardware constraints, AGI "boxing," attentiveness to and enhancement of human values, etc. rather than hoping that our ethical intuitions will become systematized in the near future.

Conclusion

Several broad classes of possible machine ethics failure modes have been identified in this paper:

1. Insufficient knowledge and/or computational resources for the situation at hand
    a. Making an exception to a rule when an exception shouldn't have been made based on the morally relevant factors
    b. Not making an exception when an exception should have been made based on the morally relevant factors
2. Moral dilemmas facing an agent result in the sacrificing of something important
3. The morals being modeled by the system are wrong
    a. Due to insufficient training data
    b. Due to folk morality being flawed
    c. Due to extrapolated human values being flawed or because of the extrapolation process itself
4. Loss of understanding or control of ethical AGI systems
    a. Due to complexity
    b. Due to extrapolation of our values far beyond our current preferences

This list is not exhaustive, and does not include some of the specific concerns raised about particular machine ethics proposals, but it illustrates the variety of issues which may prevent the creation of a reliable computational instantiation of ethical decision-making. Furthermore, there are a variety of ways (such as failures by humans in training the agents, intelligent agents being hacked, and undesired systemic effects) in which even reliable machine ethics would not ensure positive social outcomes from the diffusion of advanced A(G)I.

I will conclude by assessing machine ethics from the perspective of Nelson and Sarewitz's three criteria for "technological fixes" for social problems discussed in (Nelson and Sarewitz, 2008), where here I am referring to negative social outcomes from AI as a possible social problem. Nelson and Sarewitz argue that an effective technological fix must embody the cause-effect relationship connecting problem to solution; have effects that are assessable using relatively unambiguous and uncontroversial criteria; and build on a pre-existing, standardized technological "core." It should be fairly obvious based on the preceding analysis that machine ethics is not an adequate technological fix for the potential risks from AI according to these criteria. Machine ethics does not embody the cause-effect relationship associated with AI risks because humans are involved and responsible in various ways for the social outcomes of the technology, and there is more to successful ethical behaviour than having a good algorithm. Additionally, ethics is hardly unambiguous and uncontroversial – not only are there disagreements about the appropriate ethical framework to implement, but there are specific topics in ethical theory (such as, ex., population ethics and the other topics identified by Crouch) that appear to elude any definitive resolution regardless of the framework chosen. Finally, given the diversity of AI systems today and for the foreseeable future and the deep dependence of ethical behaviour on context, there appears to be no hope of machine ethics building on an existing technical core. All of this

suggests that machine ethics research may have some social value, but it should be analysed in a broader lens of the inherent difficulty of intelligent action in general and the complex social context in which humans and computational agents will find themselves in the future.

Acknowledgments: The following individuals helped shape my thinking concerning the topics in this paper: John Fraxedas, Jake Nebel, Amul Tevar, Stuart Armstrong, John Fraxedas, Micah Clark, David Atkinson, Luke Muelhauser, Michael Vassar, and all of my colleagues at the Consortium for Science, Policy, and Outcomes, especially Clark Miller, Brad Allenby, Dave Guston, Dan Sarewitz, Erik Fisher, Michael Burnam-Fink, and Denise Baker.

References:

Allen, C. et al. (2000). Prolegomena to any future artificial moral agent, *Journal of Experimental and Theoretical Artificial Intelligence*, **12**, 251-261.

Allenby, B. (2009). The ethics of emerging technologies: Real time macroethical assessment, *IEEE International Symposium on Sustainable Systems and Technology, ISSST '09.*

Allenby, B., and Sarewitz, D. (2011). Out of control: How to live in an unfathomable world, *New Scientist*, **2812**, 28-29.

Anderson, M., and Anderson, S. L. (2011). *Machine Ethics.* New York: Cambridge University Press.

Arkin, R., (2009). *Governing Lethal Behavior in Autonomous Robots.* London: Chapman and Hall.

Berker, S. (2009). The Normative Insignificance of Neuroscience, *Philosophy and Public Affairs*, **37(4)**, 293-329.

Berlin, I. (1990). *The Crooked Timber of Humanity: Chapters in the History of Ideas.* New York: Alfred A. Knopf.

Boehm, C. (2012). *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* New York: Basic Books.

Bringsjord, S. (2009). Unethical but Rule-Bound Robots Would Kill Us All, accessed February 28, 2013, http://kryten.mm.rpi.edu/PRES/AGI09/SB_agi09_ethicalrobots.pdf

Bringsjord, S. et al. (2011). Piagetian Roboethics via Category Theory: Moving beyond Mere Formal Operations to Engineer Robots Whose Decisions Are Guaranteed to be Ethically Correct. In Anderson, M., and Anderson, S. L. (2011). *Machine Ethics.* New York: Cambridge University Press.

Bringsjord, S., and Bello, P. (2012). On How to Build a Moral Machine, *Topoi*, 0167-7411.

Bringsjord, S., Johnson, J. (2012). Rage against the machine, *The Philosophers' Magazine*, **57**, 90-95.

Cloos, C. (2005). The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism, *American Association for Artificial Intelligence*.

Crouch, W. (2012). The most important unsolved problems in ethics, *Practical Ethics* (blog), accessed February 28, 2013, http://blog.practicalethics.ox.ac.uk/2012/10/the-most-

important-unsolved-problems-in-ethics-or-how-to-be-a-high-impact-philosopher-part-iii/.

Cushman, F. A., et al. (2010). Our multi-system moral psychology: Towards a consensus view. In J. Doris et al. (Eds.) *The Oxford Handbook of Moral Psychology.* New York: Oxford University Press.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex.* London: John Murray.

Deghani, M. et al. (2011). An Integrated Reasoning Approach to Moral Decision Making. In Anderson, M., and Anderson, S. L. (2011). *Machine Ethics.* New York: Cambridge University Press.

Freeman, T. (2009). Using Compassion and Respect to Motivate an Artificial Intelligence, accessed February 28, 2013, http://www.fungible.com/respect/paper.html

Gert, B. (2007). *Common Morality: Deciding What to Do.* New York: Oxford University Press.

Gigerenzer, G. (2010). Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality, *Topics in Cognitive Science*, **2**, 528-554.

Goertzel, B. (2006). Apparent Limitations on the "AI Friendliness" and Related Concepts Imposed by the Complexity of the World, accessed February 28, 2013, http://www.goertzel.org/papers/LimitationsOnFriendliness.pdf.

Gomila, A., Amengual, A. (2009). Moral emotions for autonomous agents. In Vallverdu, J., Casacuberta, D. (eds.), *Handbook of research on synthetic emotions and sociable robotics: new applications in affective computing and artificial intelligence.* Hershey: IGI Global.

Gowans, C. (1987). *Moral Dilemmas.* New York: Oxford University Press.

Grau, C. (2006). There Is No "I" In "Robot": Robots and Utilitarianism, *IEEE Intelligent Systems*, **21(4)**, 52-55.

Guarini, M. (2011). Computational Neural Modeling and the Philosophy of Ethics: Reflections on the Particularism-Generalism Debate. In Anderson, M., and Anderson, S. L. (2011). *Machine Ethics.* New York: Cambridge University Press.

Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion.* New York: Pantheon.

Hanson, R. (2001). Economic growth given machine intelligence, *Journal of Artificial Intelligence Research*.

Helbing, D. (2010). Systemic Risks in Society and Economics. Lausanne: International Risk Governance Council.

Horgan, T., and Timmons, M. (2009). What Does the Frame Problem Tell us About Moral Normativity?, *Ethical Theory and Moral Practice*, **12**, 25-51.

Human Rights Watch and International Human Rights Clinic. (2012). *Losing Humanity: The Case Against Killer Robots.* New York: Human Rights Watch.

Klein, C. (2011). The Dual Track Theory of Moral Decision-Making: a Critique of the Neuroimaging Evidence*, Neuroethics*, **4(2)**, 143-162.

Lenman, J. (2000). Consequentialism and Cluelessness, *Philosophy and Public Affairs*, **29(4)**, 342-370.

Mackworth, A. (2011). Architectures and Ethics for Robots: Constraint Satisfaction as a Unitary Design Framework. In Anderson, M., and Anderson, S. L. (2011). *Machine Ethics.* New York: Cambridge University Press.

Matthias, A. (2011). Is the Concept of an Ethical Governor Philosophically Sound?, In *TILTing Perspectives 2011: "Technologies on the stand: legal and ethical questions in neuroscience and robotics"* Tilburg: Tilburg University.

Mayer, R. C. et al. (1995). An Integrative Model of Organizational Trust, *Academy of Management Review*, **20(3)**, 709-734.

McLaren, B. (2011). Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions. In Anderson, M., and Anderson, S. L. (2011). *Machine Ethics.* New York: Cambridge University Press.

Muehlhauser, L., Yampolskiy, R., and Sotala, K., unpublished, Responses to AGI Catastrophic Risk: A Survey.

Müller, V. (2012). Autonomous cognitive systems in real-world environments: Less control, more flexibility and better interaction, *Cognitive Computation*, **4(3),** 212-215.

Nelson, R., and Sarewitz, D. (2008). Three rules for technological fixes, *Nature*, **456,** 871-872.

Omohundro, S. (2008). The Basic AI Drives, *Proceedings of the First AGI Conference: Frontiers in Artificial Intelligence and Applications*, 171. Amsterdam: IOS Press.

Parfit, D. (2011). *On What Matters.* New York: Oxford University Press.

Pojman, L. (2005). *Ethics: Discovering Right & Wrong.* Belmont: Wadsworth Publishing Company.

Pontier, M., Widdershoven, G., Hoorn, J. (2012). Moral Coppelia – Combining Ratio with Affect in Ethical Reasoning. In *IBERAMIA 2012*, 442-451.

Powers, T. (2011). Prospects for a Kantian Machine. In Anderson, M., and Anderson, S. L. (2011). *Machine Ethics.* New York: Cambridge University Press.

Reynolds, C. (2005). On the Computational Complexity of Action Evaluations. In *Sixth International Conference of Computer Ethics: Philosophical Enquiry.* Enschede: University of Twente.

Ross, W. D. (1988). *The Right and the Good.* Cambridge: Hackett Publishing Company.

Savulescu, J. & Persson, I. (2012). *Unfit for the Future: The Need for Moral Enhancement.* New York: Oxford University Press.

Shaw, W. (1999). *Contemporary Ethics: Taking Account of Utilitarianism.* Hoboken: Wiley-Blackwell.

Shulman, C. (2010). Omohundro's "Basic AI Drives" and Catastrophic Risks, Machine Intelligence Research Institute.

Shulman, C., et al. (2009). Which Consequentialism? Machine Ethics and Moral Divergence. In *Proceedings of AP-CAP 2009*.

Sinnott-Armstrong, W. (1988). *Moral Dilemmas (Philosophical Theory).* Hoboken: Blackwell.

Taleb, N. (2007). *The Black Swan.* New York: Random House.

Tarleton, N. (2010). Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics, Machine Intelligence Research Institute.

Tosic, P., Agha, G. (2005). On the Computational Complexity of Predicting Dynamical Evolution of Large Agent Ensembles. In *Proceedings of the Third European Workshop on Multi-Agent Systems EUMAS '05.* Brussels: Flemish Academy of Sciences.

Wallach, W., and Allen, C. (2010). *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press.

Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence.* New York: Springer.

Waser, M. (2011). Rational Universal Benevolence: Simpler, Safer, and Wiser Than "Friendly AI," *Artificial General Intelligence: Lecture Notes in Computer Science*, **6830**, 153-162.

Williams, B. (1973). *Problems of the Self.* Cambridge: Cambridge University Press.

Williams, B., and Smart, J. J. C. (1973). *Utilitarianism: For and Against.* New York: Cambridge University Press.

Wolf, S. (1982). Moral Saints, *The Journal of Philosophy*, **79(8)**, 419-439.

Yudkowsky, E. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures, Machine Intelligence Research Institute.

Yudkowsky, E. (2007). Levels of Organization in General Intelligence. In *Artificial General Intelligence.* New York: Springer.

Yudkowsky, E. (2011). Complex Value Systems are Required to Realize Valuable Futures, *Proceedings of AGI 2011.* New York: Springer.