Machine Morality: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties

Wendell Wallach (wendell.wallach@yale.edu)
ISPS, Interdisciplinary Bioethics Center; 87 Trumbull Street
New Haven, CT 06520-8209 USA

Colin Allen (colallen@indiana.edu)
Indiana University, Department of History and Philosophy of Science
1011 E. Third Street, Bloomington, IN 47405 USA

Iva Smit(iva.smit@planet.nl)
E&E Consultants
Slanijkseweg 11, 7077 AM Netterden, The Netherlands

Abstract

The implementation of moral decision-making abilities in AI is a natural and necessary extension to the social mechanisms of autonomous software agents and robots. Engineers exploring design strategies for systems sensitive to moral considerations in their choices and actions will need to determine what role ethical theory should play in defining control architectures for such systems. The architectures for morally intelligent agents fall within two broad approaches: the top-down imposition of ethical theories, and the bottom-up building of systems that aim at specified goals or standards which may or may not be specified in explicitly theoretical terms. In this paper we wish to provide some direction for continued research by outlining the value and limitations inherent in each of these approaches.

Introduction

Moral judgment in humans is a complex activity, and it is a skill that many either fail to learn adequately or perform with limited mastery. Although there are shared values that transcend cultural differences, cultures and individuals differ in the details of their ethical systems and mores. Hence it can be extremely difficult to agree upon criteria for judging the adequacy of moral decisions. Difficult value questions also often arise where information is inadequate, and when dealing with situations where the results of actions can not be fully known in advance. Thus ethics can seem to be a fuzzy discipline, whose methods and application apply to the most confusing challenges we encounter. In some respects ethics is as far away from science as one can get. While any claim that ethics can be reduced to a science would at best be naive, we believe that determining how to enhance the moral acumen of autonomous software agents is a challenge that will significantly advance the understanding of human decision making and ethics. The engineering task of building autonomous systems that safeguard basic human values will force scientists to break down moral decision making into its component parts, recognize what kinds of decisions can and cannot be codified and managed by essentially mechanical systems, and learn how to design cognitive and affective systems capable of managing ambiguity and conflicting perspectives. This project will demand that human decision making is analyzed to a degree of specificity as yet unknown and, we believe, it has the potential to revolutionize the philosophical study of ethics. A computer system, robot, or android capable of making moral judgments would be an artificial moral agent (AMA) and the proper design of such systems is perhaps the most important and challenging task facing developers of fully autonomous systems (Allen et al., 2000). Rosalind Picard, director of the Affective Computing Group at MIT, put it well when she wrote, “The greater the freedom of a machine, the more it will need moral standards.” (Picard, 1997) The capacity to make moral judgments is far from simple, and computer scientists are a long way from substantiating the collection of affective and cognitive skills necessary for moral reasoning in artificial intelligence (AI). While there are aspects of moral judgment that can be isolated and codified for tightly defined contexts, where all available options and variables are known, moral intelligence for autonomous entities is a complex activity dependent on the integration of a broad array of discrete faculties. Over the past few decades it has become apparent that moral judgment in humans is much more than a capacity for abstract moral reasoning. Emotional intelligence (Damasio, 1995; Goleman, 1996), sociability, the ability to learn from experience and social interactions, consciousness, the capacity to understand the semantic content of symbols, a theory of mind, and the ability to “read the minds” of others all contribute to moral intelligence. Social mechanisms contribute to various refinements of behavior that people expect from each other, and will also be necessary for robots that function to a high degree of competence in social contexts. The importance of these “supra-rational” faculties raises the question of the extent to which an artificial agent must emulate human faculties to function as an adequate moral agent. Computer scientists and philosophers have begun to consider the challenge of developing computer systems capable of acting within moral guidelines, under various rubrics such as “computational ethics,” “machine ethics,” and “artificial morality.” One obvious question is whose morality or what morality should computer scientists try to implement in AI? Should the systems’ decisions and actions conform to religiously or philosophically inspired value systems, such as Christian, Buddhist, utilitarian, or other sources of social norms? In principle AMAs could be Islamic androids, Kantian software agents, or robots designed to exemplify Classical virtues, but the kind of morality people wish to implement will suggest radical differences in the underlying structure of the systems in which computer scientists implement that morality. The more central question from the perspective of a computer scientist is how to implement moral decision-making capabilities in AI. What kinds of decisions can be substantiated within computational systems given the present state of computer technology and the progress we anticipate over the next five to ten years? Attention to what can be implemented may also help us discern what kind of moral system is feasible for artificial systems. For example, some ethical theories seem to provide a kind of moral calculus which allows, in principle, the course of action that will maximize welfare to be quantitatively determined, while other ethical theories appear to require higher order mental faculties that we are a long way from knowing how to reproduce in AI. Reproducing human faculties is not the only route for implementing moral decision making in computational systems. In fact, computers may be better than humans in making moral decisions in so far as they may not be as limited by the bounded rationality that characterizes human decisions, and they need not be vulnerable to emotional hijacking (Allen, 2002; Wallach, 2004). It is worth expanding on the role of emotions in moral decision making because it illustrates many of the complexities of designing AMAs and may help us understand why AMAs or robots could function differently from humans. Despite appearances, the observation that AMAs need not be subject to emotional hijacking does not introduce any contradiction with our previous claim that emotional intelligence may be necessary for moral decision making. First, a large component of affective intelligence concerns the ability to recognize and respond appropriately to emotional states in others (Picard, 1997) and it is an open question whether this kind of intelligence itself requires a computer system to have the capacity for emotions of their own. Second, in humans, while emotions are beneficial in many circumstances, this is compatible with certain emotions being disadvantageous or even dysfunctional in other circumstances. Recognizing this fact presents an opportunity for engineers to design AMAs whose moral faculties operate in a way that makes them less susceptible to emotional interference or dysfunctionality. The literature on affective computing is burgeoning and provides a good starting point for thinking about these issues, but it is also noteworthy that this literature has thus far failed to deal with the role of emotions in *moral* decisions. This may partially reflect the lasting influence of Stoicism on scientific thinking about morality. The Stoics believed that moral reasoning should be dispassionate and free of emotional prejudice, which has been presumed to mean that emotions should be banned entirely from moral reflection. (The Stoic ideal is exemplified in such science fiction icons as Mr. Spock in *Star Trek*.) The emotional intelligence literature demonstrates that the capacity for moral reasoning is complex and that emotions must be considered an integral aspect of that capacity. Some of this complexity has been explored for rather different ends in the philosophical moral psychology literature. The development of AMAs will require us to combine the insights from affective computing, emotional intelligence, and moral psychology in order to give engineers a chart outlining the kinds of emotions and their roles in moral decision making. Whether an AMA is a computer system or robot, implementing moral faculties in AI will require that scientists and philosophers study and break down moral decision making in humans into computationally manageable modules or components. An interplay exists between, on the one hand, analyzing and theorizing about moral behavior in humans and, on the other hand, the testing of these theories in computer models. For example, in artificial life (ALife) experiments that attempt to simulate evolution, there is evidence that certain moral principles, such as reciprocal altruism and fairness, may have emerged naturally through the essentially mechanical unfolding of evolution. If specific values are indeed naturally selected during evolution and our genes bias development towards these values, this will have profound ramifications for the manner in which we build AMAs. In any case, it is essential that these systems function so as to be fully sensitive to the range of ethical concerns for the health and well-being of humans and other entities (animals, corporations, etc.) worthy of moral consideration. Given the need for this range of sensitivities, AMAs have the potential to cause considerable discomfort to human beings who are not, for example, used to having machines detect their emotional states. The ability to detect emotions has broad ethical ramifications that pose a particularly complex challenge to sociologists, computer scientists, roboticists, and engineers concerned with human reactions to machines. Successful human-machine interactions may well require that we incorporate the entire value systems underlying such interactions. This is essential if humans are going to trust sophisticated machines, for trust depends on the felt belief that those you are interacting with share your essential values and concerns, or at least will function within the constraints suggested by those values. Having set the stage, our task now is to explore some bottom-up and top-down approaches to the development of artificial systems capable of making moral decisions. By “top-down” we mean to combine the two slightly different senses of this term, as it occurs in engineering and as it occurs in ethics. In the engineering sense, a top-down approach analyzes or decomposes a task into simpler subtasks that can be directly implemented and hierarchically arranged to obtain a desired outcome. In the ethical sense, a top-down approach to ethics is one which takes an antecedently specified general ethical theory (whether philosophically derived, such as utilitarianism, or religiously motivated, such as the “Golden Rule”) and derives its consequences for particular cases. In our merged sense, a top-down approach to the design of AMAs is any approach that takes the antecedently specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory. By contrast, bottom-up approaches, if they use a prior theory at all, do so only as a way of specifying the task for the system, but not as a way of specifying an implementation method or control structure.

In bottom-up engineering, tasks can also be specified *a*theoretically using some sort of performance measure (such as winning chess games, passing the Turing Test, walking across a room without stumbling, etc.). Various trial and error techniques are available to engineers for progressively tuning the performance of systems so that they approach or surpass the performance criteria. High levels of performance on many tasks can be achieved, even though the engineer lacks a theory of the best way to decompose the task into subtasks. Post-hoc analysis of the system can sometimes yield a theory or specification of the relevant subtasks, but the results of such analyses can also be quite surprising and typically do not correspond to the kind of decomposition suggested by *a priori* theorizing. In its ethical sense, a bottom-up approach to ethics is one that treats normative values as being implicit in the activity of agents rather than explicitly articulated (or even articulatable) in terms of a general theory. In our use of the term, we wish to recognize that this may provide an accurate account of the agents' understanding of their own morality and the morality of others, while we remain neutral on the ontological question of whether morality is the kind of concept for which an adequate general theory can be produced. In practice, engineers and roboticists typically build their most complex systems using both top-down and bottom-up approaches, assembling components that fulfill specific functions guided by a theoretical top-down analysis that is typically incomplete. Problem solving for engineers often involves breaking down a complex project into discrete tasks and then assembling components that perform those discrete functions in a manner that adequately fulfills the goals specified by the original project. Commonly there is more than one route to meet the project goals, and there is a dynamic interplay between analysis of the project's structure and the testing of the system designed to meet the goals. Failures of the system may, for example, reveal that secondary considerations have been overlooked in one's original analysis of the challenge, requiring additional control systems or software that accommodates complex relationships. Because the top-down/bottom-up dichotomy is too simplistic for many complex engineering tasks, we should not expect the design of AMAs to be any different. Nevertheless, it serves a useful purpose for highlighting two potential roles of ethical theory for the design of AMAs, in suggesting actual algorithms and control structures and in providing goals and standards of evaluation.

Engineering Challenges: Bottom-Up Discrete Systems

There is a range of approaches for building any artificially intelligent system. In this paper we focus our attention on the subset of approaches which corresponds best to current research in AI and robotics: the assembly of subsystems implementing relatively discrete human capacities with the goal of creating a system that is complex enough to provide a substrate for artificial morality. Such approaches are "bottom-up" in our sense because the development and deployment of these discrete subsystems is not itself explicitly guided by any ethical theory. Rather, it is hoped that by experimenting with the way in which these subsystems interact, something that has suitable moral capacities can be created. Computer scientists and roboticists are already working on a variety of discrete AI-related skills that are presumably relevant to moral capacities. The techniques provided by artificial life, genetic algorithms, connectionism, learning algorithms, embodied or subsumptive architecture, evolutionary and epigenetic robotics, associative learning platforms, and even good old-fashioned AI, all have strengths in modeling specific cognitive skills or capacities. Discrete subsystems might be built around the most effective techniques for implementing discrete cognitive capacities and social mechanisms. But computer scientists following such an approach are then confronted with the difficult (and perhaps insurmountable) challenge of assembling these discrete systems into a functional whole. This discrete-systems approach can be contrasted with more holistic attempts to reproduce moral agency in artificial systems, such as the attempt to evolve AMAs in an ALife environment, or training a single large connectionist network to reproduce the input-output functions of real moral agents. While the former approach seems promising, it is limited by our present inability to simulate environments of sufficient social and physical complexity to favor the complex capacities required for real-world performance. The latter approach has no real practitioners (to our knowledge) for reasons that illustrate our point that practical AMA development is going to require a more structured approach; for it is surely naive to think any single current technique for training homogeneous artificial neural networks could satisfactorily replicate the entire suite of moral behavior by actual human beings. The development of artificial systems that act with sensitivity to moral considerations is presently still largely confined to designing systems with "operational morality," that is, "ensuring that the AI system functions as designed" (Wallach, 2003). This is primarily the extension of the traditional engineering concern with safety into the design of smart machines that can reliably perform a specified task, whether that entails navigating a hallway without damaging itself or running into people, visually distinguishing the presence of a human from an inanimate object, or deciphering the emotional state implicit in a facial expression. Thus engineers and computer scientists are still in the process of designing methods for adequately fulfilling the discrete tasks that might lead cumulatively to more complex activities and greater autonomy. The word "discrete" here should not be taken to mean "isolated." In fact, among the most promising approaches to robotics are those which exploit dynamic interaction between the various subtasks of visual perception, moving, manipulating, and understanding speech. For example, Roy (submitted) exploits such interactions in the development of the seeing, hearing, and talking robotic arm he calls "Ripley." Ripley's speech understanding and comprehension systems develop in the context of carrying out human requests for actions related to identifying and manipulating objects within its fields of vision and reach, while balancing internal requirements such as not allowing its servos to overheat – a very real mechanical concern, as it turns out. Especially interesting for our purposes is that Roy (2005) explicitly places his project in the context of Asimov's Three Laws of Robotics (Asimov, 1950): doing no harm to humans, obeying humans, and preserving self. Roy depicts the various subsystems involved in speech processing, object recognition, movement, and recuperation, as nodes in a graph whose edges indicate interactions between the subsystems and their relationships to the laws. It is obvious, for example, how speech comprehension is essential to the second law, and motor cooling is essential to the third law. But the most telling feature of Roy's graph for our purposes is that the line representing an edge leading from Asimov's first law to a presumed implementation of moral capacities trails off into dots. Without necessarily endorsing Asimov's laws, one may understand the challenge to computer scientists as how to connect the dots to a substantial account of ethical behavior. How do we get from discrete skills to systems that are capable of autonomously displaying complex behavior, including moral behavior, and that are capable of meeting challenges in new environmental contexts and in interaction with many agents? Some scientists hope or presume that the aggregation of discrete skill sets will lead to the emergence of higher order cognitive faculties including emotional intelligence, moral judgment, and consciousness. While "emergence" is a popular word among some scientists and philosophers of science, it is still a rather vague concept, which implies that more complex activities will somehow arise synergistically from the integration of discrete skills. Perhaps the self-organizing capacity of evolutionary algorithms, learning systems or evolutionary robots provides a technique that will facilitate emergence. Systems that learn and evolve might discover ways to integrate discrete skills in pursuit of specific goals. It is important to note that the goals here are not predefined as they would be by a top-down analysis. For example, in evolutionary systems, a major problem is how to design a fitness function that will lead to the emergence of AMAs, without explicitly applying moral criteria. The slogan "survival of the most moral" highlights the problem of saying what "most moral" amounts to in a non-circular (and computationally tractable) fashion, such that the competition for survival in a complex physical and social environment will lead to the emergence of moral agents. The individual components from which bottom-up systems are constructed tend to be mechanical and limited in the flexibility of the responses they can make. But when integrated successfully these components can give rise to complex dynamic systems with a range of choices or optional responses to external conditions and pressures. Bottom-up engineering thus offers a kind of dynamic morality, where the ongoing feedback from different social mechanisms facilitates varied responses as conditions change. For example, humans may be born trusting their parents and other immediate caregivers, but the manner in which humans, both children and adults, test new relationships and feel their way over time to deepening levels of trust is helpful in understanding what we mean by dynamic morality. We humans invest each relationship with varying degrees of trust, but there is no simple formula for trust or no one method for establishing the degree to which a given person will trust a new acquaintance. A variety of social mechanisms including low risk experiments with cooperation, the reading of another's emotions in specific situations, estimations of the other's character, and calculations regarding what we are willing to risk in a given relationship all feed into the dynamic determination of trust. Each new social interaction holds the prospect of altering the degree of trust invested in a relationship. The lesson for AI research and robotics is that while AMAs should not enter the world suspicious of all relationships, they will need the capacity to dynamically negotiate or feel their way through to elevated levels of trust with the other humans or computer systems with which they interact. A potential strength of complex bottom-up systems lies in the manner they dynamically integrate input from differing social mechanisms. Their weakness traditionally lies in not knowing which goals to use for evaluating choices and actions as contexts and circumstances change. Bottom-up systems work best when they are directed at achieving one clear goal. When the goals are several or the available information is confusing or incomplete, it is a much more difficult task for bottom-up engineering to provide a clear course of action. Nevertheless, progress in this area is being made, allowing adaptive systems to deal more effectively with transitions between different phases of behavior within a given task (e.g., Smith et al., 2002) and with transitions between different tasks (Roy, 2005). A strength of bottom-up engineering lies in the assembling of components to achieve a goal. Presuming, however, that a sophisticated capacity for moral judgment will just emerge from bottom-up engineering is unlikely to get us there, and this suggests that the analysis provided by top-down approaches will be necessary.

Top-Down Theories of Ethics

The philosophical study of ethics has focused on the question of whether there are top-down criteria that illuminate moral decision making. Are there duties, principles, rules, or goals to which the countless moral judgments we make daily are all subservient? In a changing world it is impossible to have rules, laws, or goals that adequately cover every possible situation. One of the functions of legislatures, courts, and even differing cultures is to adjudicate and prioritize values as new challenges or unanticipated events arise. Top-down ethical systems come from a variety of sources including religion, philosophy, and literature. Examples include the Golden Rule, the Ten Commandments, consequentialist or utilitarian ethics, Kant's moral imperative and other duty based theories, legal codes, Aristotle's virtues, and Asimov's three laws for robots. To date very little research has been done on the computerization of top-down ethical theories. Those few systems designed to analyze moral challenges are largely relegated to medical advisors that help doctors and other health practitioners weigh alternative courses of treatments or to evaluate whether to withhold treatment for the terminally ill. Computer-aided decision support is also popular for many business applications, but generally the actual decision making is left to human managers. There are, nevertheless, many questions as to whether such systems might undermine the moral responsibility of the human decision makers (Friedman & Kahn, 1992). Will, for example, doctors and hospitals, feel free to override the advice of an expert system with a good track record, when there will always be a computerized audit trail available to enterprising lawyers, if the human's decision goes awry? Throughout industry and in financial applications values are built into systems that must make choices among available courses of actions. Generally these are the explicit values of the institutions who use the systems and the engineers who design the systems, although many systems also substantiate often unrecognized implicit values (Friedman & Nissenbaum, 1996). One would be hard pressed to say that these systems are engaged in any form of explicit moral reasoning, although they are capable of making choices that their designers might not have been able to anticipate. The usefulness of such systems lies largely in their ability to manage large quantities of information and make decisions based on their institution's goals and values, at a speed that humans can not approach. Top-down ethical theories are each meant to capture the essence of moral judgment. Nevertheless, the history of moral philosophy can be viewed as a long inquiry into the limitations inherent in each new top-down theory proposed. These known limitations also suggest specific challenges in implementing a top-down theory in AI. For example, if you have many rules or laws, what should the system do when it encounters a challenge where two or more rules conflict? Can laws or values be prioritized or does their relative importance change with a change in context? To handle such problems philosophers have turned toward more general or abstract principles from which all the more specific or particular principles might be derived. Among ethical theories, there are two "big picture" rivals for what the general principle should be. Utilitarians claim that ultimately morality is about maximizing the total amount of "utility" (a measure of happiness or well being) in the world. The best actions (or the best specific rules to follow) are those that maximize aggregate utility. Because utilitarians care about the consequences of actions for utility, their views are called "consequentialist." Consequentialists have much computing to do, because they need to work out many, if not all, of the consequences of the available alternatives in order to evaluate them morally. The competing "big picture" view of moral principles is that ethics is about duties to others, and, on the flip side of duties, the rights of individuals. Being concerned with duties and rights, such theories fall under the heading of "deontology," a 19th century word which is pseudo-Greek for the study of obligations. In general, any list of specific duties or rights might suffer the same problem as a list of commandments (in fact, some of the traditional commandments are duties to others). For example, a duty to tell the truth might, in some circumstances, come into conflict with a duty to respect another person's privacy. One way to resolve these problems is to submit all *prima facie* duties to a higher principle. Thus, for instance, it was Kant's belief that all legitimate moral duties could be grounded in a single principle, the categorical imperative, which could be stated in such a way as to guarantee logical consistency. It turns out that a computational Kantian would also have to do much computing to achieve a full moral evaluation of any action. This is because Kant's approach to the moral evaluation of actions requires a full understanding of the motives behind the action, and an assessment of whether there would be any inconsistency if everyone acted on the same motive. This requires much understanding of psychology and of the effects of actions in the world. Other general deontological principles that seek to resolve conflicts among *prima facie* duties would face similar issues. Both consequentialist (e.g., utilitarian) and deontological (e.g., Kantian) approaches provide general principles, whose relevance to the design of ethical robots was first noted by Gips (1995). These different approaches raise their own specific computational problems, but they also raise a common problem of whether any computer (or human, for that matter) could ever gather and compare all the information that would be necessary for the theories to be applied in real time. Of course humans apply consequentialist and deontological reasoning to practical problems without calculating endlessly the utility or moral ramifications of an act in all possible situations. Our morality, just as our reasoning, is bounded by time, capacity, and inclination. Parameters might also be set to limit the extent to which a computational system analyzes the beneficial or imperatival consequences of a specific action. How might we set those limits on the options considered by a computer system, and will the course of action taken by such a system in addressing a specific challenge be satisfactory? In humans the limits on reflection are set by heuristics and affective controls. Both heuristics and affect can at times be irrational but also tend to embody the wisdom gained through experience. We may well be able to implement heuristics in computational systems, but affective controls on moral judgments represent a much more difficult challenge.

Our brief survey by no means exhausts the variety of top-down ethical theories that might be implemented in AI. Rather, it has been our intention to use these two representative theories to illustrate the kinds of issues that arise when one considers the challenges entailed in designing an AMA. Consequentialism comes in many variations and each religious tradition or culture has its own set of rules. Consequences and rules are combined in some theories. Arguably one of these variations might have built in heuristics that cut through some of the complexity we have discussed. Rule consequentialists, for example, determine whether an act is morally wrong based on a code of rules that have been selected solely in terms of their average or expected consequences. The consequences of adopting the rule must be more favourable than unfavourable to all parties involved. Presumably these rules, once selected, will cut down on much of the calculation that we described earlier for an act consequentialist who will have to evaluate all the consequences of his action. But there is a trade-off. The rule consequentialist must deal with conflicting rules, the prioritization of rules, and any exceptions to the rules. When thinking of rules for robots, Asimov's laws come immediately to mind. On the surface these three laws, plus a "zeroth" law that he added in 1985 to place humanity's interest above that of any individual, appear to be intuitive, straightforward, and general enough in scope to capture a broad array of ethical concerns. But in story after story Asimov demonstrates problems of prioritization and potential deadlock inherent in implementing even this small set of rules (Clark, 1994). Apparently Asimov concluded that his laws would not work, and other theorists have extended this conclusion to encompass any rule based ethical system implemented in AI (Lang, 2002). In addition, we will require some criteria for evaluating whether a moral judgment made by a computational system is satisfactory. When dealing with differences in values and ethical theories this is by no means an easy matter, and even criteria such as a Moral Turing Test, designed to manage differences in opinion as to what is 'right' and 'good,' have inherent weaknesses (Allen et al., 2000). The strength of top-down theories lies in their defining ethical goals with a breadth that subsumes countless specific challenges. But this strength can come at a price: either the goals are defined so vaguely or abstractly that their meaning and application is subject for debate, or they get defined in a manner that is static and fails to accommodate or may even be hostile to new conditions.

Merging Top-Down and Bottom-Up

As we indicated at the end of the introduction, the top-down/bottom-up dichotomy is somewhat simplistic. Top-down analysis and bottom-up techniques for developing or evolving skills and mental faculties will undoubtedly both be required to engineer AMAs. To illustrate the way in which top-down and bottom-up aspects interact, we consider two cases. First, we'll look at the possibility of utilizing a connectionist network to develop a computer system with good character traits or virtues and then we'll look at attempts to develop complex mental faculties and social mechanism, such as a theory of mind for an embodied robotic system. We'll also address the challenge of bringing it all together in an agent capable of interacting morally in a dynamic social context.

Virtue Ethics

Public discussions of morality are not just about rights (deontology) and welfare (utility); they are often also about issues of character. This third element of moral theory can be traced back to Aristotle, and what is now known as "virtue ethics." Virtue ethicists are not concerned with evaluating the morality of actions on the basis solely of outcomes (consequentialism), or in terms of rights and duties (deontology). Virtue theorists maintain that morally good actions flow from the cultivation of good character, which consists in the realization of specific virtues. Virtues are more complex than mere skills, as they involve characteristic patterns of motivation and desire. Socrates claimed that virtues couldn't be misused, because if people have a certain virtue, it is impossible for them to act as if they did not have it. This led him to the conclusion that there is only one virtue, the power of right judgment, while Aristotle favored a longer list. Just as utilitarians do not agree on how to measure utility, and deontologists do not agree on which list of duties apply, virtue ethicists do not agree on a standard list of virtues that any moral agent should exemplify. Rather than focus on these difference in this paper, our attention will be directed at the computational tractability of virtue ethics: could one make use of virtues as a programming tool? Virtues affect how we deliberate and how we motivate our actions, but an explicit description of the relevant virtue rarely occurs in the content of the deliberation. For instance, a kind person does kind things, but typically will not explain this behavior in terms of her own kindness. Rather, a kind person will state motives focused on the beneficiary of the kindness, such as "she needs it," "it will cheer him up," or "it will stop the pain" (Williams, 1985). Besides revealing some of the complexities of virtue theory, this example also demonstrates that the boundaries between the various ethical theories, in this case utilitarianism and virtue based ethics, can be quite fuzzy. Indeed, the very process of developing one's virtues is hard to imagine independently of training oneself to act for the right motives so as to produce good outcomes. Aristotle contended that the moral virtues are distinct from practical wisdom and intellectual virtues, which can be taught. He believed that the moral virtues must be learned through habit and through practice. This emphasis on habit, learning, and character places virtue theory in between the top-down explicit values advocated by a culture, and the bottom-up traits discovered or learned by an individual through practice. Building computers with character can be approached as either a top-down implementation of virtues or the development of character by a learning computer. The former approach views virtues as characteristics that can be programmed into the system. The latter approach stems from the recognition of a convergence between modern connectionist approaches to AI and virtue-based ethical systems, particularly that of Aristotle. Top-down implementations of the virtues are especially challenged by the fact that virtues comprise complex patterns of motivation and desire, and manifest themselves indirectly. For example, the virtue of being kind can be projected to hundreds of different activities. If applying virtue-theory in a top-down fashion, an artificial agent would have to have considerable knowledge of psychology to figure out which virtue, or which action representing the virtue, to call upon in a given situation. A virtue-based

AMA, like its deontological cohorts, could get stuck in endless looping when checking if its actions are congruent with the prescribed virtues, and then reflecting upon the checking, and so on. The problems here bear some relationship to the well-known Frame Problem, but may also be related to other issues that frequently arise in AI and database contexts concerning predictions of indirect effects over extended time periods. By linking the virtues to function (as in the Greek tradition), and tailoring them sharply to the specific tasks of an AMA, perhaps some of these computational problems can be mitigated. In principle, the stability of virtues – that is, if one has a virtue, one cannot behave as if one does not have it – is a very attractive feature considering the need for an AMA to maintain "loyalty" under pressure while dealing with various, not always legitimate, information sources. In humans, the stability of virtue largely stems from the emotional grounding of virtues, which motivates us to uphold our image as honorable beings. The challenge for a designer of AMAs is to find a way to implement the same stability in a 'cold' unemotional machine. A virtuous robot may require emotions of its own as well as emotionally rooted goals such as happiness. Perhaps the artificial simulation of an admirable goal or desire to meet the criterion of being virtuous will suffice, but in all likelihood we will only find out by going through the actual exercise of building a virtue-based computational system. Several writers have mentioned that connectionism or parallel distributed processing has similarities to Aristotle's discussion of virtue ethics (DeMoss, 1998; Howell, 1999). Connectionism provides a bottom-up strategy for building complex capacities, recognizing patterns or building categories naturally by mapping statistical regularities in complex inputs. Through the gradual accumulation of data the network develops generalized responses that go beyond the particulars on which it is trained. Rather than relying on abstract, linguistically-represented theoretical knowledge, connectionist systems "seem to emphasize the immediate, the perceptual, and the non-symbolic" (Gips, 1991; see also Churchland, 1995; and DeMoss, 1998). After stating his virtue-based theory, Aristotle spends much of the *Nicomachean Ethics* discussing the problem of how one is to know which habits will lead to the "good," or happiness. He is clear at the outset that there is no explicit rule for pursuing this generalized end, which is only grasped intuitively. The end is deduced from the particulars, from making connections between means and ends, between those specific things we need to do and the goals we wish to pursue. Through intuition, induction and experience – asking good people about the good – our generalized sense of the goal comes into focus, and we acquire practical wisdom and moral excellence. Howell (1999) argues that connectionism is capable of explaining how human minds develop intuitions through the unconscious assimilation of large amounts of experience; he writes: "Random experiences, connectionist learning through exposure to instances, reinforcement learning, and advice-seeking and taking, all play a part in childhood development, and make us what we are." It is interesting and suggestive to note the similarity between Aristotelian ethics and connectionism, but the challenge of implementing virtues within a neural network remains a formidable one. Existing connectionist systems are a long way from tackling the kind of complex learning tasks we associate with moral development. Nevertheless, the prospect that artificial neural networks might be employed for at least some dimensions of moral reasoning is an intriguing possibility that deserves extensive consideration.

Consciousness, Theory of Mind, and Other Supra-Rational Faculties and Social Mechanisms

A socially viable robot will require a rich set of skills to function properly as a moral agent in multi-agent contexts that are in a dynamic state of flux. The relationships between these agents will be constantly evolving as will the social context within which they operate. Customs change. Particularly in regards to the AMAs themselves, one might imagine increasing acceptance and latitude for their actions as humans come to feel that their behavior is trustworthy. Conversely, if AMAs fail to act appropriately, the public will demand laws and practices that add new restrictions upon the AMAs behavior. Morality evolves and the AMAs will be active participants in working through new challenges in many realms of activity. Computers actually function quite well in multi-agent environments where auctions, bargaining, or other forms of negotiation are the primary mode of interactions. But each transaction can also change the relationships between agents. Within specific contexts, roles, status, wealth, and other forms of social privilege give form to relationships and the perception of what kinds of behavior are acceptable. The AMA will need to understand customs and when to honor and when to ignore these prerogatives, which can easily change – for example, when an individual person moves into a new role over time or even many times during a single day. In recent years the focus on high-level decision making has been directed less at theories of ethics and more toward particular faculties and social mechanisms that enhance the ability to reason in concrete situations, especially emotions (Damasio 1995; Clark 1998). Navigating a constantly changing environment with both human and non-human agents will require AMAs to have both sophisticated social mechanisms as well as rich informational input about the changes in the context, and among other agents with which it interacts. Whether we think of these faculties – which include things such as consciousness, emotions, embodied intelligence, and theory of mind – as supra-rational, or as illuminating once hidden dimensions of reason, is not of critical concern. What is important is that agents without these tools will either fail in their moral decision-making abilities or be constrained to act in limited domains. Good habits, good character, supra-rational faculties and social mechanisms, are all relevant to our current understanding of moral acumen. If each of these faculties and mechanisms is, in turn, a composite of lower level skills, designers of

AMAs can expect to borrow ideas from other attempts to model other high-level mental capacities. For example, in Igor Aleksander et al.'s (2005) top-down analysis, consciousness can be broken down into five broad skills including imagining, attending to input, thinking ahead, emotions, and being in an out-there world. Each of these skills is in turn a composite or set of more limited tools. Each is likely to be independently relevant to the design of AMAs, but there is also the possibility that AMAs will be limited unless they are fully conscious, and it also seems likely that consciousness cannot be reduced to just five broad skills. Scassellati's (2001) work on a theory of mind, the ability of an entity to appreciate the existence of other minds or other complex entities affected by its actions, illustrates the challenge of implementing a complex social mechanism within the design of a robot. Utilizing the theories of cognitive scientists who have broken down a theory of mind into discrete skills, Scassellati and other computer scientists have tried to substantiate each of these skills in hardware. While collectively these skills might lead to a system that actually acts as if it had a theory of mind, to date they have had only limited success in substantiating a few of these attributes in AI. The hard work of coordinating or integrating these skill sets lies ahead. Perhaps advances in evolutionary robotics will facilitate this integration. To date we not only lack systems with a theory of mind, but we do not yet know whether we have an adequate hypothesis about the attributes that are necessary for a system to have a theory of mind. The development of other faculties and social mechanisms for computer systems and robots, such as being embodied in the world, emotional intelligence, the capacity to learn from experience, and the ability to 'understand' the semantic content of symbols are also each in similar primitive states of development. If the components of a system are well designed and integrated properly the breadth of choices open to an AMA in responding to challenges arising from its environment and social context will expand. Presumably the top-down capacity to evaluate those options will lead to selecting those actions which both meet its goals and fall within acceptable social norms. But ultimately the test of a successful AMA will not rely on whether its bottom-up components or top-down evaluative modules are individually satisfactory. The system must function as a moral agent, responding to both internal and externally arising challenges. The immediate and ongoing challenge for computer science and robotics is the ability of individual components and modules to work together harmoniously. Given that the complexity of this integration will grow exponentially as we bring in more and more systems, scientists and engineers should focus on evolutionary and other self-organizing techniques.

Conclusion

Autonomous systems must make choices in the course of flexibly fulfilling their missions, and some of those choices will have potentially harmful consequences for humans and other subjects of moral concern. Systems that approximate moral acumen to some degree, even if crude, are more desirable than ethically "blind" systems that select actions without any sensitivity to moral considerations. Ultimately the degree of comfort the general public will feel with autonomous systems depends on the belief that these systems will not harm humans and other entities worthy of moral consideration, and will honor basic human values or norms. Short of this comfort, political pressures to slow or stop the development of autonomous systems will mount. Deep philosophical objections to both the possibility and desirability of creating artificial systems that make complex decisions are unlikely to slow the inexorable progression toward autonomous software agents and robots. Whether strong artificial intelligence is possible remains an open question, yet regardless of how intelligent AI systems may become, they will require some degree of moral sensitivity in the choices and actions they take. If there are limitations in the extent to which scientists can implement moral decision making capabilities in AI, it is incumbent to recognize those limitations, so that we humans do not rely inappropriately on artificial decision makers. Moral judgment for a fully functioning AMA would require the integration of many discrete skills. In the meantime, scientists will build systems that test more limited goals for specific applications. These more specialized applications will not require the full range of social mechanisms and reasoning powers necessary for complicated ethical judgments in social contexts. Emotional intelligence would not be required for a computer system managing the shutdown procedures of an electrical grid during a power surge. While such systems are not usually thought to be engaged in making moral decisions, presumably they will be designed to calculate minimal harm and maximize utility. On the other hand, a service robot tending the elderly would need to be sensitive to the possibility of upsetting or frightening its clients, and should take appropriate action if it senses that its behavior caused fear or any other form of emotional disturbance. We have suggested that thinking in terms of top-down and bottom-up approaches to substantiating moral decision-making faculties in AI provides a useful framework for grasping the multi-faceted dimensions of this challenge. There are limitations inherent in viewing moral judgments as subsumed exclusively under either bottom-up or top-down approaches. The capacity for moral judgment in humans is a hybrid of both bottom-up mechanisms shaped by evolution and learning, and top-down mechanisms capable of theory-driven reasoning. Morally intelligent robots require a similar fusion of bottom-up propensities and discrete skills and the top-down evaluation of possible courses of action. Eventually, perhaps sooner rather than later, we will need AMAs which maintain the dynamic and flexible morality of bottom-up systems that accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles that represent ideals we strive to meet. A full appreciation of the way in which these elements might be integrated leads into important metaethical issues concerning the nature of moral agency itself that are beyond the scope of the present paper. Nevertheless, we hope to have indicated the rich and varied sources of insights that can immediately be deployed by scientists who are ready to face the challenge of creating computer systems that function as AMAs.

Acknowledgments

An earlier version of the paper was prepared for and presented at the Android Science Cog Sci 2005 Workshop in Stresa, Italy. We wish to thank Karl MacDorman, co-organizer of workshop, for his encouragement, detailed comments, and editing suggestions. We are also grateful for the helpful comments provided by three anonymous reviewers arranged by the Workshop organizers.

References

Aleksander, I., Lahstein, M., & Lee, R. (2005). Will and emotions: A machine model that shuns illusion. In *AISB '05: Proceedings of the symposium on next generation approaches to machine consciousness.*

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence 12*, 251–261.

Allen, C. (2002). Calculated morality: Ethical computing in the limit. In Smit, I. & Lasker, G.E. (Eds.), *Cognitive, emotive and ethical aspects of decision making and human action Vol. I.* Windsor, Canada: IIAS.

Asimov, I. (1950). *I, Robot.* Gnome Press.

Churchland, P.M. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain.* Cambridge, Massachusetts: MIT Press.

Clark, A. (1998). Being there: Putting brain, body, and world together again. Cambridge, MA: MIT Press.

Clark, R. (1993, 1994). Asimov's laws of robotics: Implications for information technology. Published in two parts, in IEEE Computer *26, 12* (December 1993) pp. 53-61 and 27,1 (January 1994), pp. 57-66.

Damasio, A. (1995). *Descartes error.* New York: Pan Macmillan.

Danielson, P. (2003). Modeling complex ethical agents. Presented at the conference on *Computational modeling in the social sciences*, Univ. of Washington. *http://www.ethics.ubc.ca/pad/papers/mcea.pdf.*

DeMoss, D. (1998). Aristotle, connectionism, and the morally excellent brain. The Paideia Archive. Proceedings of the 20th World Congress of Philosophy.

Friedman, B. & Kahn, P. (1992). Human agency and responsible computing. *Journal of Systems and Software 17*, 7–14.

Friedman, B. & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems 14*, 330–347.

Gips, J. (1995) Towards the ethical robot. In Ford, K., Glymour, C., & Hayes, P. (eds.) *Android epistemology.* MIT Press, Cambridge, MA, pp. 243-252.

Goleman, D. (1995). *Emotional intelligence.* New York: Bantam Books.

Howell, S.R. (1999). Neural networks and philosophy: Why Aristotle was a connectionist. http://www.psychology.mcmaster.ca/beckerlab/showell/aristotle.pdf

Lang, C. (2002). Ethics for artificial intelligence. Available at http://philosophy.wisc.edu/lang/AIEthics/index.htm

Picard, R. (1997). *Affective computing.* Cambridge, MA: MIT Press.

Roy, D. (2005). Meaning machines. Talk given at Indiana University, March 7, 2005.

Roy, D. (submitted). Grounding language in the word: Schema theory meets semiotics. *http://web.media.mit.edu/~dkroy/papers/pdf/aij_v5.pdf.*

Scassellati, B. (2001). Foundations for a theory of mind for a humanoid robot. *http://www.ai.mit.edu/projects/lbr/hrg/2001/scassellati-phd.pdf.*

Smith, T., Husbands, P., & Philippides, A. (2002). Neuronal plasticity and temporal adaptivity: GasNet robot control networks. *Adaptive Behavior* 10, 161-183.

Skyrms, B. (2000). Game theory, rationality and evolution of the social contract in evolutionary origins of morality. In Katz, L. (Ed.), *Evolutionary origins of morality* (pp. 269–285). Exeter, UK: Imprint Academic.

Wallach, W. (2003). Robot morals and human ethics in Smit, I., Lasker, L. & Wallach, W., *Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence Vol. II.* Windsor, Canada: IIAS.

Wallach, W. (2004). Artificial morality: bounded rationality, bounded morality and emotions. In Smit, I., Lasker, L. & Wallach, W. (Eds.), *Cognitive, emotive and ethical aspects of decision making and human action Vol. III.* Windsor, Canada: IIAS.

Williams, B. (1985). *Ethics and the limits of philosophy.* Cambridge, MA: Harvard University Press.