Reprinted from Smit, I, Lasker, L. and Wallach, W., *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol III , pp. 1-6, IIAS

Artificial Morality:
Bounded Rationality, Bounded Morality and Emotions

Wendell Wallach, WW Associates, Bloomfield, CT, USA wwallach@comcast.net

Abstract

A central question in the development and design of artificial moral agents is whether the absence of emotions, and the capacity of computer systems to manage large quantities of information and analyze many courses of actions, will make them equal or superior to humans in making moral judgments. The contrasting contention that emotions, a sense of self, embodiment, consciousness, and the ability to understand the semantic content of symbols are essential to a faculty for moral judgment would suggest that in comparison to human morality, artificial morality will be inferior, if not inadequate. If these additional human faculties are essential to a moral intelligence, then it will be necessary to simulate or design similar faculties into an artificial moral agent. On the other hand, some of these additional faculties serve functions that are essential primarily because of limitations of the human mind, limitations that computers don't necessarily share.

In particular, I'll focus on whether some of the help we derive in making decisions from emotions and an understanding of the semantic content of values, function as compensations for our limited ability to comprehensively analyze challenges we face. Emotions and values are essential to facilitate the bounded rationality of human beings, but will be less essential, and often unnecessary to the calculated morality of artificial moral agents. This, of course, presumes that the potential comprehensive rationality of a computer is truly functional in meeting challenges fraught with real-world tensions, and not merely limited to the bounded moral environments in which artificial moral agents will initially act.

Keywords:
AI, artificial intelligence, decision-making, moral agents, ethics, emotions, bounded rationality

Artificial Morality

Artificial morality extends the field of computer ethics (CE) by fostering an exploration of the technological and philosophical issues involved in making computers themselves into explicit moral reasoners. Artificial morality is directed at building into AI systems sensitivity to the values, ethics, and legality of activities performed by computers and robots. The goal of artificial morality is designing artificial agents to act *as if* they are moral agents.

The relatively young field of CE has been defined broadly as the "analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology" (Moor, 1985). But CE is commonly understood more restrictively as focusing on applying ethical and legal standards – including privacy, property rights, and civil rights – to the use of computers. CE has been viewed as comparable to a professional code of ethics in which standards are set to guide and instruct people in limiting the use of their computers to ethical and legal activities (Barquin, 1992; Berger 2001). Other theorists, who view CE more expansively, contend that computers pose unique ethical considerations that are of philosophical interest (Floridi, 1998).

The emphasis in CE has been on the ethical use of computers by humans. The broader discussion has illuminated ways in which computers might violate ethical standards as a matter of course. The ease of copying data with computers, for example, has thrown copyright law into turmoil. A search engine may collect data that is legally considered to be private, unbeknownst to the user who initiated the search query.

Presumably, awareness of the ethical ramifications of computer usage might facilitate ethical behavior, but as with the violation of copyright laws that has not necessarily been the case.

Artificial morality shifts some of the burden for ethical behavior onto the computer system. This becomes particularly important as computers are being designed to perform with greater and greater autonomy. In addition, the speed at which computers perform mundane tasks would make it prohibitive for a human to evaluate whether each action is performed in a responsible or ethical manner. Implementing software agents with moral decision making capabilities offers the promise of computer systems able to evaluate whether each action it performs is ethically appropriate. This in no way serves as

a substitute for the moral responsibility of those who use computers. It merely means that using computers in an ethical manner will be easier if sensitivity to ethical and legal values is built into the software. For example, Data Mining software might be designed so that in its search for information it could discriminate which databases it can scan freely and when it is invading the privacy or property rights of an individual or a corporation. In theory then, a user could initiate a broad based search while being insulated by the software from retrieving information that is legally or morally restricted.

Artificial morality is concerned with understanding the kinds of ethical considerations that lend themselves to being implemented in computers and robots. Moral decision making capabilities might in turn lead to systems that display greater autonomy and perhaps the development of an artificial moral agent (Allen *et al.*, 2000). This presumes that there are no practical limits to the moral decision making capabilities we can implement in computer systems. If there are limits to artificial morality, then it will be important for us to recognize those limits so that we will not place a false reliance on autonomous systems.

Advantages and Limits of Artificial Morality

In comparison to human morality, artificial morality might arguably be inferior. However, two aspects of information systems are emphasized by those who suggest that computers hold the promise of acquiring a moral faculty that is equal to if not superior to that of humans (Bostrom, 2003). Computers are capable of managing large quantities of information and analyzing many courses of action, while humans are restricted in the information they can manage and the courses of action they analyze (Allen, 2002). Thus a computer has more choices and might select a course of action superior to those considered by a human counterpart. Secondly, the absence of emotions means that a computer will not be subject to the kind of emotional highjacking and sentimentality that interfere with human reason. Both of these advantages are rooted in a stoic view of ethics, where reason unfettered by emotions will lead to a more developed moral sensibility.

Those who are critical of the prospect that computers will develop faculties comparable to humans focus on the absence of consciousness, the lack of a sense of self, the importance of being embodied in the world, their inability to *understand* the semantic content of symbols, the absence of emotions, and difficulty in working with challenges where the information is incomplete, misunderstood, false, or where the result of a course of action is unknown. True believers (Kurzweil, 1999) in what John Searle (1980) dubbed "strong artificial intelligence", contend that we will learn how to simulate each of these faculties in artificial systems. The MIT Robotics Laboratory is working on the design of robotics that are embodied in their environment, while Brian Scazallatti at Yale is extending this research to the development of a robot with a sense of self. Much of AI research is directed at designing software that functions to recognize the semantic content of words and symbols, while computer consciousness is also being tackled with a variety of different approaches (Holland, 2003). Contemporary computers do not have *understanding* or conscious, and may never have. However, this does not necessarily mean we will not be able to design systems that function as if they were conscious or understand the meaning of the information they process.

The absence of emotions arises as both an advantage and limitation when considering the potential moral acumen of computer systems. The stoic worldview emphasized disturbing or destructive aspects of human emotions. "Emotional intelligence" (Goleman, 1995) is the phrase that subsumes our contemporary understanding that emotions are multifaceted and beneficial in a variety of ways. Recognizing the emotional state of others helps us interact socially and respond appropriately. Our emotions and feeling can also be quite helpful in indicating a course of action, particularly when our reason fails to indicate the desirability of one course over another (DeSousa, 1987). The neuroscientist Antonio Damasio (1995) relates a now famous story about Eliot, an intelligent patient with brain damage to the neural circuitry necessary for processing secondary emotions. Eliot is unable to make even simple decisions, which suggests that emotions may be an essential component of all decision making.

Instinctual and emotional responses to a challenge hold information that may be germane to meeting the challenge, for example, fear warns us that we may be in danger. Emotions provide quick instinctual responses to threats, while the brain centers, which evaluate information, are much slower. Bypassing the analytical deciphering of information serves to further those ethical considerations, for example, survival, protection of children, and procreating, that have been hard-wired by evolution into the emotional brain. But much of our emotional apparatus evolved in a very different world than we now inhabit, and can lead to responses that are irrational, if not actually dysfunctional, in the modern world. Inappropriate emotions, emotional high-jackings, and emotional flooding can interfere with our ability to meet challenges rationally and may even lead to violent and self-destructive responses. Presumably any artificial moral agent (AMA) we develop will not need to carry this dysfunction evolutionary baggage.

Bounded Rationality, Calculated Morality, and Emotions

Two seminal moments in the birth of cognitive science and cognitive psychology were the publication in 1956 of a paper by George Miller titled, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information", and in 1957 Herbert Simon's proposing a theory of "satisficing decisions" or "bounded rationality", for which he won the 1978 Nobel Prize in Economics. Both Miller and Simon stressed the limits of human conscious mental faculties. Miller's paper reviewed research that short-term memory is limited, that we can only hold approximately seven thoughts in our mind at one time, and then went on to propose that information be recorded in chunks or mental representations. Simon argued that humans don't have enough brainpower to handle the breadth of information they must sort through in order to make informed decision. Rather than efficiency, maximizing profits, or some other goal, we settle for the first option we encounter that's "good enough".

Humans augment their decision-making capabilities with experience, feel, intuition, and judgment – skills lacking in contemporary computers. But the defeat of the chess master Gary Kasparov in 1997 by IBM's Deep Blue II illustrates that raw computing power can prevail over rich mental faculties. Deep Blue II used its computing power to look at a much larger range of chess moves than was possible for a master chess player.

Colin Allen (2002) has proposed that while the ascription of being "too calculating" suggests a lack of ethical sensitivity in a person, it may well indicate superiority in the decision making capability of artificial moral agents. If a computer might indeed analyze more deeply than a human the ramifications of different course of action, its decisions hold the prospect of being more rational.

It is important to recognize that much of our use of values in decision making function as a compensation or short-cut method for handling the vast bulk of information that impinges on a decision. Many challenges give way to a rational course of action when one looks deeply enough at the relevant information. But we often lack the time, inclination, or mental faculties for such analysis and therefore turn to value judgments, intuition, or feelings that determine the weight of the information we do consider. This methodology suggests that human decisions will be less than optimal in comparison to those of a well-designed computer system when dealing with challenges that eventually yield to a rational analysis.

On the other hand, humans have been far superior to computers in handling challenges where our information is incomplete, inadequate, contradictory, or the effects of a course of action are not predictable (Wallach, 2003). In addition to our values, experience, and intuition, our emotions play a major role in facilitating a response to those challenges where we lack enough information to determine a rational response.

Simulating Emotional Intelligence

In theory, it may be possible to develop computer systems capable of applying values, experience, fuzzy logic, and other heuristic tools to difficult problems. Emotions and consciousness are perhaps the most difficult abilities to simulate in artificial systems. Whether artificial systems can actually have emotions or consciousness is an outstanding question, but they may be able to function as if they do.

In computer simulations of the intelligence we acquire from emotions, the most significant strides have been made in designing systems that are cognizant of the emotions of others. Designing computer systems to read emotions through facial expressions and tactile cues, as well as responding appropriately to input from users about their feelings, are among the projects being pursued at MIT's Media Lab. Rosalind Picard, the Director of the Labs' Affective Computing group, is more skeptical regarding the prospect of developing computer systems that have emotions of their own (Picard, 1997). From a stoic perspective, one might even consider this self-defeating.

We can imagine a computer with a substitute for emotions that performs some of the beneficial tasks emotions perform without the apparent drawbacks of emotions. An internal weighing system based on a hierarchy of values might tip the balance and thereby facilitates ethical decision-making in difficult situations. Whether such substitutes for emotions adequately compensate for the rich information we derive from our emotions is unclear.

From a computational perspective emotions can be thought of as affective output derived from processing a vast amount of informational input through biochemical and physical activity in the body and nervous systems. Input is accumulated into an affective output. Discrepancies are defused or dispelled in the affective states, which provide us with a generalized sense of the prevailing conditions or the challenge at hand. Loss of emotional control might be understood as sensory output that exceeds the computational capacity of the mind/body.

We expect computer systems to eventually have information processing capacity that exceeds that of mind/bodies, so that even a system with emotions built-in might have the capability of managing emotional content without system overload. This of course begs the question as to whether a computational theory of mind is the best way to think of the manner in which forces, such as a gust of wind or witnessing the death of a loved one, effect our bodies and minds. It also presumes that silicon, or whatever other medium we use, will be as effective as carbon-based cells and molecules in processing sensory data.

Bounded Morality

Simulating the functions served by consciousness and emotions in humans within computer systems represent the far horizon of artificial morality. These faculties will be most important in service robots, humanoids, and other computation systems that make judgments that immediately effect the individuals with whom they interact. Artificial moral agents that provide human decision makers with decision support tools will not require such faculties. Computational systems that analyze the applicability of laws or social customs to discrete general challenges, such as the ethics of Data Mining, will require little or no representation of emotional concerns.

Approaches to designing artificial moral agents will vary. Computer scientists within research laboratories will explore experimental approaches. The natural approach for engineers within industry will be to consider ethical requirements as additional design constraints similar to constraints on a system's safety or efficiency. Initially, systems will be developed to function within limited contexts – bounded environments. The moral reasoning of a computer system will not be considered to be distinct from any other constraint necessary for the system to function in the specified context.

The question, of course, is what constraints will we build in, should those restraints be treated as hard rules or soft guidelines, and how will we manage challenges in which values or constraints conflict? Ensuring that the computer system or robot will act morally will be dependent on formulating the right set of constraints and formulas for the environment within which the system will operate.

The effectiveness of these constraints will be largely determined by how clearly the context within which the system functions has been defined. For example, the activity of search engines is limited to systems connected in networks or the WEB, but it is difficult to constrain these agents' behavior *vis a vis* privacy or property rights, because the applicable body of law remains in question. As public policy debate leads to clearer guidelines regarding the application of privacy and property rights to data within computer systems, boundaries will be sharpened regarding the legality of data searches within specific contexts. One might expect that there will be a clear delineation between data that is private and data that is public or data that is protected by property law.

Within a bounded context it becomes quite feasible for a computer system, functioning within a set of constraints and the accompanying formulas for resolving conflicts, to exhaustively analyze the breadth of available courses of action. Such systems will display a kind of "bounded morality", capable of behaving safely, legally, or morally in regards to those situations encountered that fit within the general constraints anticipated by designers.

The calculated morality of a computer system may well suffice to fulfill the legal and moral requirements of artificial agents functioning within bounded environments. Ensuring the morality of systems that operate in unbounded environments or in a wide variety of contexts will be a much more difficult problem. Such "unbounded morality " will require moral acumen similar to that we expect in humans, and may well demand that we develop AMA's with high order mental faculties and even something akin to emotions.

Moral Judgment or Mere Calculation?

Arguably, a computer system that applies legal and ethical constraints it has been designed to follow may be engaging in a form of reasoning, but is not in any respect actually making moral judgments. In it's unique form of analysis, the system might give rise to conclusions or results that couldn't be predicted in advance, but those decisions were, in effect, predetermined by the values implemented by the system designers. Of course, a similar claim is made by determinists who question the freedom of humans in making choices.

Humans function autonomously. There is a subtlety in the manner we direct our attention, both inwardly and outwardly, and in the relationship between our emotions and our consciousness, that make an unpredictable capacity for freedom of judgment and free will plausible.

Systems with artificial intelligence will move ever closer to the goal of being fully autonomous agents. In turn, the challenge of designing these agents so that they honor the broader set of values and laws that we demand of human moral agents will become increasingly pressing. Whether AMA's will be engaged in a sophisticated form of calculated morality, a new form of artificial morality, or true moral intelligence is a question that will remain unanswered until we actually determine how to develop such systems. Even then, we may be mystified as to how to interpret their behavior.

References:

Allen, C., Varner, G. and Zinser, J. (2000); A Prolegomena to Any Future Moral Agent; Journal of Experimental and Theoretical Artificial Intelligence, Vol. No.9 (pp. 251-261)

Allen, C. (2002); Calculated Morality: Ethical Computing in the Limit; in Smit, I. and Lasker, G.(ed.): Cognitive, Emotive, and Ethical Aspects of Decision Making and Human Action, Vol. II; The International Institute for Advanced Studies in Systems Research and Cybernetics (pp. 19-23)

Bostrom, N.(2003);Ethical Issues in Advanced Artificial Intelligence; in Smit, I., Lasker, G. and Wallach, W. (ed.): Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence; IIAS (pp.12-17)

Barquin, R. (1992) In Pursuit of a 'Ten Commandments' for Computer Ethics; Washington Consulting Group and Computer Ethics Institute

Berger, R. (2001) Is Computer Ethics Unique In Relation To Other Fields Of Ethics?; http://www.nd.edu/~rbarger/ce_unique.html
Damasio, A. (1995); Descartes Error; Pan Macmillan

DeSousa, R. (1987), The Rationality of Emotions; MIT Press

Floridi, L. (1998) Information Ethics: On the Philosophical Foundation of Computer Ethics; ETHICOMP98 The Fourth International Conference on Ethical Issues of Information Technology http://www.wolfson.ox.ac.uk/~floridi/ie.htm

Goleman, D. (1995); Emotional intelligence; Bantam Books

Holland, O. (2003); Machine Consciousness; Imprint Academic

Kurzweil.R. (1999) The Age of Spiritual Machines, When Computers Exceed Human Intelligence; Viking Penguin

Miller, G. (1956) The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information; Psychological Review, 63 (pp. 81-97)

Moor, J. H. (1985) What is Computer Ethics?; in Byrum, T.W. (ed.): Computers & Ethics; Blackwell

Picard, R. (1997) Affective Computing; The MIT. Press.

Searle, J. (1980) Minds, Brains, and Programs; The Behavioral and Brain Sciences, vol. 3.; http://members.aol.com/NeoNoetics/ Searles

Simon, H.A. (1982) Models of Bounded Rationality; The MIT Press

Wallach, W. (2003) Robert Morals and Human Ethics; in Smit, I., Lasker, G. and Wallach, W.(ed.); Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intellignce, IIAS (pp. 1-5)