Colin Allen:
Morality and Artificial Intelligence

How and why did you get interested in the field of machine morality?

The question of how to make a machine behave ethically was on a list compiled by one of artificial intelligence's luminaries of topics where philosophers could help. It seemed like an interesting challenge and I had just been invited to write an article for an artificial intelligence journal, so I decided to see where I could take it.

Artificial Intelligence, Machine learning and genetic programming, just to name a few branches, are highly complex fields of research. Coming as you did from a meta-science, how did you approach this challenge from an ethical perspective?

Well, let me start by saying I am not an ethicist! I'm a philosopher of science and philosopher of mind who to that point had mostly worked on issues in animal cognition, but I had also taken quite a few post-graduate courses in computer science, specializing in artificial intelligence. So, the first thing I did was to talk to an ethicist colleague of mine, Gary Varner, about my ideas for the article and he agreed to be a co-author. My approach was initially to ask the same technical questions about whether ethical theories such as Kant's or Bentham's could in fact be computed. Later, in the book with Wendell Wallach, this became what we called the "top down" approach.

Your book "Moral Machines" discusses the field of machines as moral agents. Should we define morality as purely human quality or should we use a concept of different qualities of morality? Also from a practical perspective: what concept of morality should we use while discussing the issues right at hand?

Wendell and I wrote the book with a very practical question in mind: How, as a matter of fact, would one improve the kinds of machines that are already being built so that they could be better, morally speaking? As such, we didn't want to prejudge any questions about the nature of morality, who or what has it, etc. We recognized that philosophers tend to gravitate towards the hard cases where there is much disagreement, because this is where theories get tested against intuitions. But despite this, there's a surprising amount of agreement about practical ethics. Whether you're a utilitarian or Kantian, Christian or Buddhist, you can agree that stabbing the stranger sitting next to you on the train is morally bad, or, more subtly, that anyone to whom we cause a harm has a prima facie moral claim against us. Of course, there's lots of room for disagreement about what constitutes a harm, and when it is acceptable to cause a harm, but our basic premise was that most machines, robots and software bots, that are currently making harmful decisions don't even have the means to take those harms into account when making these decisions.

You have used the term “artificial moral agents”, why and how would you differentiate natural from artificial moral agents?

Like artificial anything, we want to acknowledge that deliberately engineered products will not be the same as those that have grown organically. Artificial sweeteners aren't the same as sugars, and artificial intelligence only resembles biological intelligence. Whether artificial moral agents ever become as capable as biological moral agents is a question for science fiction and futurism. I should also acknowledge that for some ethical theorists, the central problem of moral agency is the conflict between selfish inclination and moral duty, but this assumes a form of psychology that may not apply to artificial agents. Nevertheless, for the time being we know that any artificial system we place in an ethically charged decision making situation will have strengths and limitations. Many of those limitations stem from our not really understanding, either at a scientific or humanistic level, what goes into making us moral agents. (Lots of theories, no consensus.) So in part the project of building artificial moral agents is partly a project of self- evaluation. If we don't flag what we're doing with the term "artificial" there's a risk of losing sight of our own role in shaping these systems.

Are there beneficial aspects of looking at morality from the perspective of the artificial intelligence theory?

One of the interesting things, I think, that comes out of the attempt to think in computational terms about morality or ethics is a richer conception of the space in which ethical behavior operates. Rather than seeing these as opposite poles, I'm more inclined to see them as separate axes or dimensions of the decision space. The time- and information-bounded nature of most decision making makes embodied dispositions an essential part of moral agency. There simply isn't enough time in the world to compute all of the consequences, actual or logical, of an action, even if one had perfect information. So, moral agents must be disposed to react in ways that are morally acceptable.

These bottom up reactivities are also, however, subject to top-down evaluation, and, here emotions like pride, regret, or shame can serve to strengthen or weaken dispositions, but so can a reasoned determination to live up to an abstract principle. Given the abstract nature of most top-down principles, however, it is hardly surprising that they sometimes conflict with each other and with our dispositionally-formed intuitions. The result is that any moral principle could be overridden in a specific situation. As socially-enculturated human beings, it is natural for us to want to come up with some higher principle to adjudicate these conflicts, but in the absence of such a principle, what one has is a decision space in which duties, consequences, and dispositions are all relevant dimensions, but none is paramount. Moral agency involves a hybrid of bottom up and top down processes, often operating over different time scales. "Shoot first, ask questions later" is the wrong slogan because we can ask some questions first, but our ability to do so is often limited and we must return to the questions in retrospect, hoping to calibrate the shooting response better next time we are in a similar situation.

We are a long way from being able to build hybrid architectures for artificial moral agents to have such sophistication. But a chief goal of the book is to start a discussion about whether providing machines with just part of the bottom up or top down capacities for moral decision making would be better than having machines that are ethically insensitive to such considerations. What information does a battlefield robot or your bank's computer have to have in order to make decisions that a human moral agent would endorse? What reasoning capabilities would it need to be able to weigh collective outcomes against individual rights, either prospectively or retroactively?

Most people see robots and computers as predetermined machines without any ability to transcend into the sphere of decision making. How was your approach to this topic and how did people respond to your concept of artificial moral agents?

Whether predetermined or not, the fact is that machines are involved in all sorts of decisions, from approving credit card transactions to allocating resources in hospitals. They are even being deployed as automatic sentries on national borders. I'm not sure whether this means that they have "transcended into the sphere of decision making" but it does mean that without direct human oversight machines are selecting among options that have moral consequences. The metaphysical questions about whether this is "really" decision making don't concern me as much as the practical questions about whether these machines can be made sophisticated enough to weigh the factors that are important for ethical decision making.

People react to the idea of artificial moral agents in several ways. Some assume that we are talking about human-level artificial intelligence and dismiss the topic as pure science fiction, and others assume we must be concerned with whether robots themselves deserve rights. For me, however, it is important to avoid science fiction and stay focused on what is likely to happen in the next decade or so. A different kind of worry comes from those who say that by using the word "agents" for machines we are contributing to the abdication of human responsibility for the consequences of our own technologies. I recognize the seriousness of the concern, but I think it's also likely that by referring to artificial moral agents we set up a kind of dissonance that might help users recognize that they should be wary of overestimating the capacities of these machines.

So what you are saying is, that right now we should focus more on the practical ethical challenges at hand which arise from the use of these systems (e.g. the Future Attribute Screening Technology (FAST) – Hostile Intent Detection of the Department of Homeland Security[1] than to engage in speculation on full moral agency of machines. Do you think that your book could be something like a whistleblower by starting this discussion?

It was certainly our intention to help start such a discussion. And it's interesting that we seem to be in the middle of a small explosion of interest in the topic. Just after our book came out, Peter Singer's more journalistic Wired for War came out to significant press coverage, and now Ron Arkin's Governing Lethal Behavior in Autonomous Robots has just been released, the first book to provide an actual design specification for robots capable of exercising lethal force. While these other books focus on military applications, we think it's important to recognize that the issues go far beyond the battlefield.

In your book you have put forward two dimensions for artificial moral agents: ethical sensitivity and autonomy. On this framework you differentiate between operational and functional morality as well as finally full moral agency. How can we understand these moralities and where on this framework are robots now (and where can they probably be finally)?

There are not intended to be hard and fast distinctions, but operational morality is intended to include cases where the decisions about what is a morally acceptable behavior are largely in the hands of the designers and programmers, whereas functional morality implies some built-in capacities for moral reasoning or decision making. Operational morality generally applies to machines that operate in relatively closed environments with relatively few options for action.

Under these circumstances, designers may be able to anticipate the situations the machine will encounter and pre-specify the morally preferred actions in those circumstances. In more open environments where machines have greater autonomy, they must be designed to detect ethically relevant features of the situation, and select among options accordingly. We use the term "functional morality" primarily to acknowledge that these capacities may fall short of the full moral agency of human beings, although I would like to maintain that it's an open question whether there are any limits to what machines can do. At the current time, machine autonomy is increasing, meaning that machines are operating in more open environments without human oversight and with more options available to them. But aside from a few A.I. projects that are described in chapters 9 and 10 of the book, there is relatively little work on giving machines the kind of ethical sensitivity that, in combination with autonomy, would be necessary for functional morality.

Why do you think it is like that? It seems obvious that there is a need for research on this matter.

I don’t think it is a deliberate omission, but a sign of how new the field is. Engineers tend to prefer well-defined problems, and as I’ve already mentioned, philosophers like controversial topics. For this and other reasons it’s actually quite a challenge to bring the two cultures together. But it is coming. In addition to our book and the others that have recently appeared, a scholarly collection of essays edited by the computer scientist-philosopher husband-wife team of Michael and Susan Anderson is in the works. And a couple of graduate student projects that I’m aware of show that they are starting to pay attention are thinking creatively about how ethical capabilities might be important in a variety of online and real-world contexts.

What can robots with representations of emotions – like the projects KISMET and later on Nexi MDS – do for the development of artificial moral agents?

I think emotion-representing robots do two things for artificial moral agents. One is perhaps quite dangerous, in that it may cause people to attribute more understanding of their own emotions to the machines than there really is. If Kismet or Nexi reacts to a person's sad face by looking sad, there is a risk that the person will assume more empathy than exists.

This is dangerous if the machine lacks the capacity to help the person properly deal with the situation that is causing the sadness. The other thing may be essential, however, since part of the ethical sensitivity required for functional morality involves being able to detect and react to the emotional responses of people interacting with the robot. All other things being equal, if a robot through its actions causes anger or sadness, then it needs to reevaluate that action. This is not to say that robots should always change their behavior whenever they detect a negative emotional response, or do whatever it takes to get a positive emotional response from the people it is interacting with. But such responses are crucial pieces of information in assessing the moral appropriateness of actions.

The KISMET Project has been very well documented and the emotional responses you refer to can be seen on videos on the webpage of the MIT Computer Science and Artificial Intelligence Laboratory[2]. What do you think about the use of robots in the entertainment industry? In some countries in Asia robots are being developed explicitly as "personal companions". What impact will that have on interpersonal relations of humans, especially children growing up with robotic pets?

The sex industry has driven a lot of technology development, from the earliest faxes through postcards to videotape recording and online video on demand. The more "respectable" face of robotic companions for the elderly and toys for children are just the tip of a very large iceberg. I think it's hard to say what kind of impact these technologies will have for human interpersonal relationships. It will probably bring benefits and costs, just as with the Internet itself. It's easy to find lots of people who lament the replacement of face-to-face interactions with Facebook, Twitter, and the like. But at the same time probably all of us can think of old friendships renewed, or remote relationships strengthened by the use of email and online social networking. I don't think robotic pets are inherently bad for children, although I am sure there are those who will complain that one doesn't have to be as imaginative with a robot as with a stuffed toy. I'm not so sure this is correct. With a robotic toy, a child may be able to imagine different possibilities, and a robotic pet will likely serve as a nexus of interactions in play with other children. And just as we are finding that highly interactive video games can bring cognitive benefits to young[3] and old[4] alike, we may find that robotic companions do likewise. Of course there will be problems too, so we must remain vigilant without being fearful that change is always a bad thing.

Free will, understanding and consciousness are seen as crucial for moral decisions though they are often attributed exclusively to humans. You have argued that functional equivalence of behaviour is what really matters in the practical issues of designing artificial moral agents. What is your perspective on these three fields concerning artificial moral agents?

All of these are again looking towards more futuristic end of this discussion. People in A.I. have for over 50 years been saying that we'll have full human equivalency in 50 years. I don't know whether it will be 50 years or 100 years or never, because I don't think we know enough about human understanding, consciousness, and free will to know what's technologically feasible. My stance, though, is that it doesn't really matter. Military planners are already sponsoring the development of battlefield robots that will have greater autonomous capacities than the hundreds of remote-operated vehicles that are already in use. The military are sufficiently concerned about the ethical issues that they are funding research into the question of whether autonomous robots can be programmed to follow the Geneva conventions and other rules of war. These questions are pressing regardless of whether these machines are conscious or have free will. But if you want my futuristic speculation, then I'm a bit more pessimistic than those who are predicting a rapid take-off for machine intelligence in the next 25-30 years, but I would be very surprised if my grand- children or great-grandchildren aren't surrounded by robots that can do anything a person can do, physically or cognitively.

As you said military robots are a reality on the battlefields today and it seems clear that their number and roles will expand, probably faster than most of us think or would like them to. Do you think that the military is actually ready for the changes these semiautonomous systems bring to the army?

I'm encouraged by the fact that at least some people in the military understand the problem and they are willing to support research into solutions. Both the U.S. Navy and Army have funded projects looking at ethical behavior in robots. Of course, it's possible to be cynical and assume that they are simply trying to provide cover for more and more impersonal ways of killing people in war. But I think this underestimates the variety and sophistication of military officers, many of whom do have deep moral concerns about modern warfare. Whether the military as a whole is ready for the changes is a different matter perhaps, because for someone on the front lines, sending a robot into a cave with authorization to kill anything that moves may seem like a pretty attractive idea. There will be missteps – there always have been – and I'm fairly sure that the military is not actually ready for all the changes that these systems will bring because some of those changes are unpredictable.

One of your other fields of study has been animal cognition. Have you found this helpful while developing your perspectives on artificial moral agents?

It's a good question because I started off really treating these as separate projects. However, thinking about the capacities of nonhuman animals, and the fact that it isn't really a dog-eat-dog world, leads to some ideas about the behavioral, cognitive, and evolutionary foundations of human morality. Various forms of pro-social (and proto-ethical) behavior are increasingly being reported by experimentalists and observers of natural behavior of animals. Of course, nonhuman animals aren't, as far as we know, reflective deliberators, but neither is all of basic human decency and kindness driven by explicit ethical reasoning. Animals give us some ideas about the possibilities for machines that aren't full moral agents.

So you are referring to studies like Benjamin Libet's through which the absolute predominance of reason in human decision making is questioned in favour of subconscious processes. It is easily comprehensible that these concepts will be seminal, though it seems to be harder to create a model of ethical behaviour by the means of animals, considering the complexity of the mind, than developing a simpler rule-based behaviour system. What do you think are the main areas where the development of artificial morality could benefit from the research in animal cognition? Or maybe one could even say, that concepts which stem from this field are crucial for a realistic approach to artificial morality?

One of the things we are learning from animals is that they can be quite sensitive to reciprocity of exchange in long term relationships. If one animal shares food with or grooms another, there doesn't have to be an immediate quid pro quo. Speaking only slightly anthropomorphically one could say that they build relationships of trust, and there is even evidence that early play bouts may provide a foundation for such trust. These foundations support generally "pro-social" behavior.

Humans are no different, in that we establish trust incrementally. However, what's remarkable about human society is that we frequently trust total strangers and it usually turns out all right. This is not a consciously reasoned decision and, as recent research in behavioral economics shows, may even involve acting against our immediate self interests. Artificial moral agents will also have to operate in the context of human society with its mixture of personal relationships based on medium to long term reciprocity and transactions with strangers that depend for their success on local social norms. Ethical robots have to be pro-social, but not foolishly so. Animal studies can do a lot to help us understand the evolution and development of pro-social behavior, and some of this will be transferable to our robot designs.

The purpose of the already mentioned NEXI MDS project at the MIT Personal Robots Group[5] is to support research and education goals in human-robot interaction, teaming, and social learning. Do you think projects like this which focus on the improvement of robots for interpersonal relations could benefit from the research in animal behaviour?

I recently attended a conference in Budapest on comparative social cognition that had both animal and robot researchers, so these are two communities that are already in dialogue. Particularly interesting, I think, is that we are finding a variety of social learning capabilities not just in the species most closely related to humans, the anthropoid apes, but in species that are much more distant from us. Especially interesting in this regard are dogs, who in many respects are even more human-like than chimpanzees in their capacity for social interaction and cooperation with us. By studying dogs, and which signals from us they attend to, we can learn a lot about how to design robots to use the same cues.

You have identified two main approaches to artificial moral agents, the top-down approach (one could say a rule-based approach) and the bottom-up approach (which is often seen in connection with genetic programming). How can these to approaches help in building artificial moral agents and where lie their strengths and weaknesses?

A strength of top-down approaches is that the ethical commitments are explicit in the rules. The rules can also be used to explain the decision that was taken. However, it is hard to write rules that are specific enough to be applied unambiguously in all circumstances.

Also, the rules may lead to what we have called a "computational black hole" meaning that it is really impossible to gather and process all the information that would really be necessary to make a decision according to the rules. Bottom-up approaches, and here I'd include not just genetic algorithms but various kinds of learning techniques, have the strength of being able to adaptively respond and generalize to new situations based on limited information, but when systems become sufficiently complex they have the drawback that it is often unclear why a particular outcome occurred.

To overcome the restraints of both approaches you have suggested merging these two to a hybrid moral system. How can we imagine this?

I believe that we will need systems that continuously engage in a self-evaluative process. We describe it as a virtue-based approach because it has some things in common with Aristotle's ethics. Bottom-up processes form a kind of reactive layer that can be trained to have fundamentally sound responses to moral circumstances. A robot following an instruction by a human must not be completely opportunistic in the means it takes to carry out that instruction, running roughshod over the people for whom it is not directly working.

Rules alone can't capture what's needed. One can't say, for instance, "never borrow a tool without asking" or "never violate a direct order from a human being" for we want agents that are flexible enough to recognize that sometimes it is acceptable, and perhaps even obligatory, to do so. Such decisions are likely to require a lot of context-sensitivity, and for this, a bottom-up approach is best.

However, we will want these same systems to monitor and re-evaluate the outcomes in light of top-down principles. Sometimes one cannot know whether another's welfare is affected or rights violated until well after the fact, but a reflective moral agent, on learning of such an outcome, should endeavor to retrain its reactive processes, or reform its principles. But this is a very hard problem, and is perhaps where the project of artificial moral agents really does slide down the slope into science fiction. But by pointing out that there are reasons to think that neither a top-down or a bottom-up approach will alone be sufficient, we hope to have initiated a debate about how to develop machines that we can trust.

Would this monitor and evaluation system be something like the “ethical governor” which Ronald Arkin proposed in his project on “Governing Lethal Behaviour”?

Overall, there’s considerable similarity between our hybrid approach and Arkin’s “deliberative/reactive” architecture. However, because his “ethical governor” operates immediately prior to any action being taken, actually what I have been describing is something closer to his “ethical adaptor” which is another component of his ethical control architecture, and which is responsible for updating the ethical constraints in the system if an after-the-fact evaluation shows that a rule violation occurred. A significant difference between our approach and Arkin’s is that the rules themselves (e.g. the Geneva Conventions) are considered to be known and fixed, and not themselves subject to interpretation or revision. This approach is possible because he considers only the case of robots operating in a well-defined battlefield and engaging only with identifiable hostile forces. Arkin believes that in such circumstances, intelligent robots can actually behave more ethically than humans can. Humans get angry or scared and commit war crimes, and Arkin’s view is that robots won’t have these emotional reactions, although he recognizes that some sort of affective guidance is important.

Besides research and teaching you are also maintaining a blog on the theory and development of artificial moral agents and computational ethics[6], so I guess you will be working on these fields in the future? And which projects are you currently working on?

Right, I’ll continue to keep an eye on machine morality issues, but I’m currently being reactive rather than pursuing any new lines of research in this area. My biggest current ongoing project is something completely different – with funding from the U.S. National Endowment for the Humanities we are developing software to help us build and maintain a complete representation of the discipline of philosophy, that we call the Indiana Philosophy Ontology, or InPhO for short[7]. I’m also continuing to work actively on topics in the philosophy of cognitive science, and I’m currently working on papers about the perceptual basis of symbolic reasoning and about the use of structural mathematical models in cognitive science, among other topics.

[1] http://www.dhs.gov/xres/programs/gc_1218 480185 439.shtm.

[2] e.g. http://www.ai.mit.edu/projects/sociable /movies/affective-intent-narrative.mov.

[3] e.g. http://discovermagazine.com/2005/jul/ brain-on-video-games.

[4] e.g. http://www.sciencedaily.com/releases/ 2008/12/081211081442.htm.

[5] http://robotic.media.mit.edu/projects/robots/ mds/overview/overview.html.

[6] http://moralmachines.blogspot.com.

[7] http://inpho.cogs.indiana.edu.