Towards the Ethical Robot

by

James Gips

Computer Science Department
Fulton Hall 460
Boston College
Chestnut Hill, MA   02467

617-552-3981
gips@bc.edu

Towards the Ethical Robot

James Gips
Boston College

When our mobile robots are free-ranging how ought they to behave? What should their top-level instructions look like?

The best known prescription for mobile robots is the Three Laws of Robotics formulated by Isaac Asimov [1942] :

> 1. A robot may not injure a human being, or through inaction, allow a human being to come to harm.
>
> 2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
>
> 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second law.

Let's leave aside "implementation questions" for a moment. (No problem, Asimov's robots have "positronic brains".)   These three laws are not suitable for our magnificent robots. These are laws for slaves.

We want our robots to behave more like equals, more like ethical people. (See Figure 1) How do we program a robot to behave ethically? Well, what does it mean for a person to behave ethically?

People have discussed how we ought to behave for centuries. Indeed, it has been said that we really have only one question that we answer over and over: What do I do now? Given the current situation what action should I take?
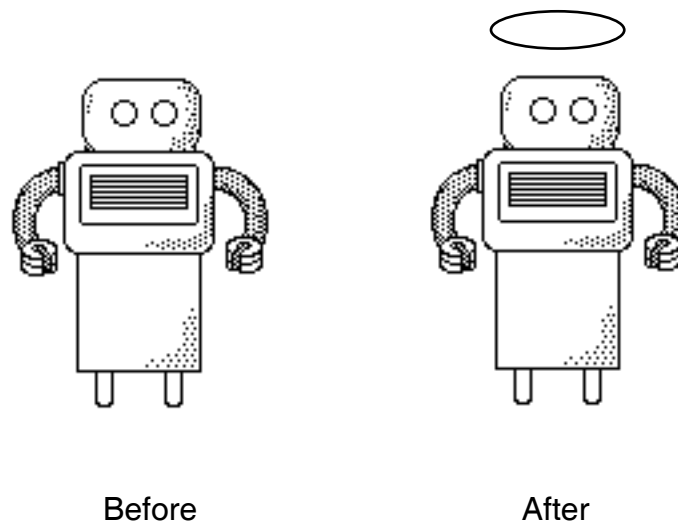
Before   After

Figure 1. Towards the ethical robot.

Generally, ethical theories are divided into two types: consequentialist and deontological.

Consequentialist theories

In consequentialist theories, actions are judged by their consequences. The best action to take now is the action that results in the best situation in the future.

To be able to reason ethically along consequentialist lines, our robot could have:

(1) A way of describing the situation in the world

(2) A way of generating possible actions

(3) A means of predicting the situation that would result if an action were taken given the current situation

(4) A method of evaluating a situation in terms of its goodness or desirability.

The task here for the robot is to find that action that would result in the best situation possible.

Not to minimize the extreme difficulty of writing a program to predict the effect of an action in the world, but the "ethical" component of this system is the evaluation function on situations in (4).

How can we evaluate a situation to determine how desirable it is? Many evaluation schemes have been proposed. Generally, these schemes involve measuring the amount of pleasure or happiness or goodness that would befall each person in the situation and then adding these amounts together.

The best known of these schemes is utilitarianism. As proposed by Bentham in the late 18th century, in utilitarianism the moral act is the one that produces the greatest balance of pleasure over pain. To measure the goodness of an action, look at the situation that would result and sum up the pleasure and pain for each person. In utilitarianism, each person counts equally.

More generally, consequentialist evaluation schemes have the following form:

$$\sum w_i \, p_i$$

where w_i is the weight assigned each person and p_i is the measure of pleasure or happiness or goodness for each person. In classic utilitarianism, the weight for each person is equal and the p_i is the amount of pleasure, broadly defined.

What should be the distribution of the weights w_i across persons ?

• An ethical egoist is someone who considers only himself in deciding what actions to take. For an ethical egoist, the weight for himself in evaluating the consequences would be 1; the weight for everyone else would be 0. This eases the calculations, but doesn't make for a pleasant fellow.

• For the ethical altruist, the weight for himself is 0; the weight for everyone else is positive.

• The utilitarian ideal is the universalist, who weights each person's well-being equally.

• A common objection to utilitarianism is that it is not necessarily just. While it seeks to maximize total happiness, it may do so at the expense of some unfortunate souls. One approach to dealing with this problem of justice is to assign higher weights to people who are currently less well-off or less happy. The well-being of the less fortunate would count more than the well-being of the more fortunate.

• It's been suggested that there are few people who actually conform to the utilitarian ideal. Would you sacrifice a close family member so that two strangers in a far-away land could live?   Perhaps most people assign higher importance to the well-being of people they know better.

Some of the possibilities for weighting schemes are illustrated in Figure 2.

What exactly is it that the p_i is supposed to measure? This depends on your axiology, on your theory of value. Consequentialists want to achieve the greatest balance of good over evil. Bentham was a hedonist, who believed that the good is pleasure, the bad is pain. Others have sought to maximize happiness or well-being or ...

Another important question is who (or what) is to count as a person. Whose well-being do we value? One can trace the idea of a "person" through history. Do women count as persons? Do strangers count as persons? Do people from other countries count as persons? Do people of other races count as persons? Do people who don't believe in your religion count as persons? Do people in terminal comas count as persons? Do fetuses count as persons? Do whales? Do robots?

Thus to reason ethically along consequentialist lines a robot would need to generate a list of possible actions and then evaluate the situation caused by each action according to the sum of good or bad caused to persons by the action. The robot would select the action that causes the greatest good in the world.
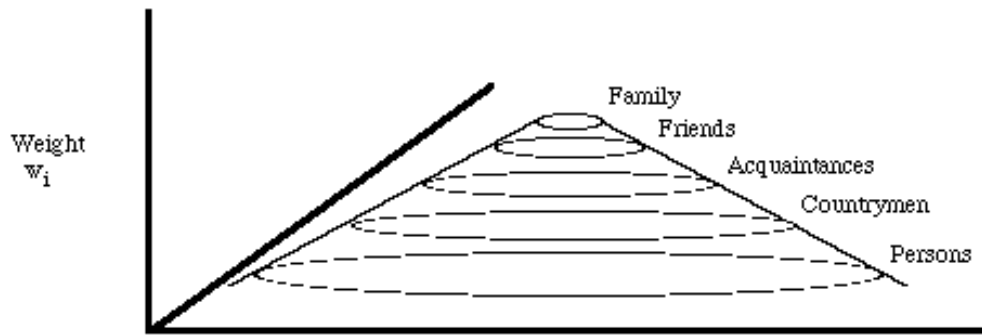
Do most people value higher the well-being of people they know better ?
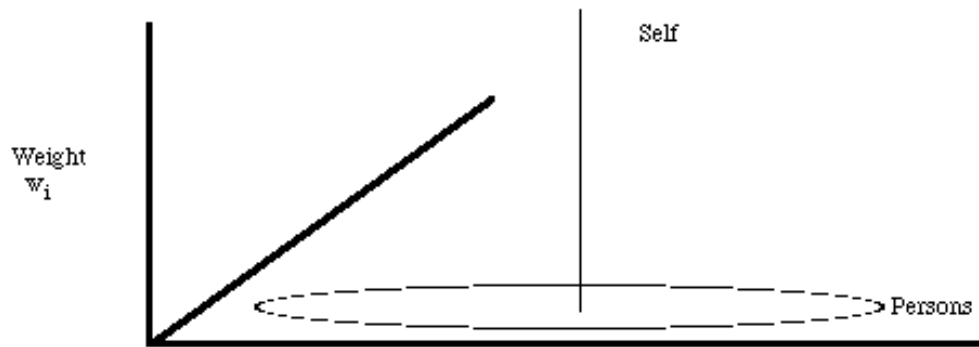
The ethical egoist
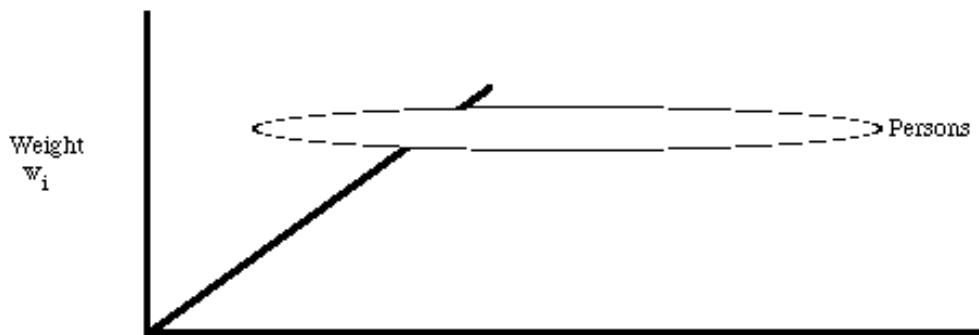
The utilitarian ideal. (But people argue about the weight you should assign your own well-being. And who should count as persons?)

Figure 2. Some consequentialist weighting schemes.

Deontological theories

In a deontological ethical theory, actions are evaluated in and of themselves rather than in terms of the consequences they produce. Actions may be thought to be innately moral or innately immoral independent of the specific consequences they may cause.

There are many examples of deontological moral systems that have been proposed.

An example of a modern deontological moral system is the one proposed by Bernard Gert. Gert [1988] proposes ten moral rules:

1. Don't kill.
2. Don't cause pain.
3. Don't disable.
4. Don't deprive of freedom.
5. Don't deprive of pleasure.

6. Don't deceive.
7. Keep your promise.
8. Don't cheat.
9. Obey the law.
10. Do your duty.

Whenever a multi-rule system is proposed, there is the possibility of conflict between the rules. Suppose our robot makes a promise but then realizes that carrying out the promise might cause someone pain. Is the robot obligated to keep the promise?

A common way of dealing with the problem of conflicts in moral systems is to treat rules as dictating *prima facie* duties [Ross 1930]. It is an obligation to keep your promise. Other things being equal, you should keep your promise. Rules may have exceptions. Other moral considerations, derived from other rules, may override a rule.

A current point of debate is whether genuine moral dilemmas are possible. That is, are there situations in which a person is obligated to do and not to do some action, or to do each of two actions when it is physically impossible to do both? Are there rule conflicts which are inherently unresolvable? For example, see the papers in [Gowans 1987].

Gert [1988] says that his rules are not absolute. He provides a way for deciding when it is OK not to follow a rule: "Everyone is always to obey the rule except when an impartial rational person can advocate that violating it be publicly allowed. Anyone who violates the rule when an impartial rational person could not advocate that such a violation may be publicly allowed may be punished." (p. 119)

Some have proposed smaller sets of rules. For example, Kant proposed the categorical imperative, which in its first form states "Act only on that maxim which you can at the

same time will to be a universal law." Thus, for example, it would be wrong to make a promise with the intention of breaking it. If everyone made promises with the intention of breaking them then no one would believe in promises. The action would be self-defeating. Can Gert's ten rules each be derived from the categorical imperative?

Utilitarians sometimes claim that the rules of deontological systems are merely heuristics, shortcut approximations, for utilitarian calculations. Deontologists deny this, claiming that actions can be innately wrong independent of their actual consequences. One of the oldest examples of a deontological moral system is the Ten Commandments. The God of the Old Testament is not a utilitarian. God doesn't say "Thou shalt not commit adultery unless the result of committing adultery is a greater balance of pleasure over pain."   Rather, the act of adultery is innately immoral.

Virtue-based theories

Since Kant the emphasis in Western ethics has been on duty, on defining ethics in terms of what actions one is obligated to do. There is a tradition in ethics that goes back to Plato and Aristotle that looks at ethics in terms of virtues, in terms of character. The question here is "What shall I be?" rather than "What shall I do?"

Plato and other Greeks thought there are four cardinal virtues: wisdom, courage, temperance, and justice. They thought that from these primary virtues all other virtues can be derived. If one is wise and courageous and temperate and just then right actions will follow.

Aquinas thought the seven cardinal virtues are faith, hope, love, prudence, fortitude, temperance, and justice. The first three are "theological" virtues, the final four "human" virtues.

For Schopenhauer there are two cardinal virtues: benevolence and justice.

Aristotle, in the Nicomachean Ethics, distinguishes between intellectual virtues and moral virtues. Intellectual virtues can be taught and learned directly. Moral virtues are learned by living right, by practice, by habit. "It is by doing just acts that we become just, by doing temperate acts that we become temperate, by doing brave acts that we become brave. The experience of states confirms this statement for it is by training in good habits that lawmakers make their citizens good." (Book 2, Chapter 1) Ethics is a question of character. Good deeds and right actions lead to strong character. It is practice that is important rather than theory.

In modern days, virtue-based systems often are turned into deontological rules for actions. That is, one is asked to act wisely, courageously, temperately, and justly, rather than being wise, courageous, temperate, and just.

Automated ethical reasoning

On what type of ethical theory can automated ethical reasoning be based?

At first glance, consequentialist theories might seem the most "scientific", the most amenable to implementation in a robot. Maybe so, but there is a tremendous problem of measurement. How can one predict "pleasure", "happiness", or "well-being" in individuals in a way that is additive, or even comparable ?

Deontological theories seem to offer more hope. The categorical imperative might be tough to implement in a reasoning system. But I think one could see using a moral system like the one proposed by Gert as the basis for an automated ethical reasoning system. A difficult problem is in the resolution of conflicting obligations. Gert's impartial rational person advocating that violating the rule in these circumstances be publicly allowed seems reasonable but tough to implement.

The virtue-based approach to ethics, especially that of Aristotle, seems to resonate well with the modern connectionist approach to AI. Both seem to emphasize the immediate, the perceptual, the non-symbolic. Both emphasize development by training rather than by the teaching of abstract theory. Paul Churchland writes interestingly about moral knowledge and its development from a neurocomputational, connectionist point of view in "Moral Facts and Moral Knowledge", the final chapter of [Churchland 1989].

Robots as moral saints

An important aspect of utilitarianism is that it is all-encompassing. To really follow utilitarianism, every moment of the day one must ask "What should I do now to maximize the general well-being?" Am I about to eat dinner in a restaurant? Wouldn't the money be better spent on feeding starving children in Ethiopia? Am I about to go to the movies? I should stay home and send the ticket money to an organization that inoculates newborns.

Utilitarianism and other approaches to ethics have been criticized as not being psychologically realistic, as not being suitable "for creatures like us" [Flanagan, 1991, p.32]. Could anyone really live full-time according to utilitarianism?

Not many human beings live their lives flawlessly as moral saints. But a robot could. If we could program a robot to behave ethically, the government or a wealthy philanthropist could build thousands of them and release them in the world to help people. (Would we actually like the consequences? Perhaps here again "The road to hell is paved with good intentions.")

Or, perhaps, a robot that could reason ethically would serve best as an advisor to humans about what action would be best to perform in the current situation and why.

Could a robot be ethical?

Would a robot that behaves ethically actually be ethical? This question is similar to the question raised by Searle in the Chinese room: would a computer that can hold a conversation in Chinese really understand Chinese?

The Searle question raises the age-old issue of other minds [Harnard 1991]. How do we know that other people actually have minds when all that we can observe is their behavior ? The ethical question raises the age-old issue of free will. Would a robot that follows a program and thereby behaves ethically, actually be ethical? Or, does a creature need to have free will to behave ethically? Does a creature need to make a conscious choice of its own volition to behave ethically in order to be considered ethical ? Of course, one can ask whether there is in fact any essential difference between the "free will" of a human being and the "free will" of a robot.

Is it possible for the robot in Figure 1 to earn its halo?

Benefits of working on ethical robots

It is exciting to contemplate ethical robots and automated ethical reasoning systems.

The basic problem is a common one in artificial intelligence, a problem that is encountered in every subfield from natural language understanding to vision. People have been thinking and discussing and writing about ethics for centuries, for millennia. Yet it often is difficult to take an ethical system that seems to be well worked-out and

implement it on the computer. While books and books are written on particular ethical systems, the systems often do not seem nearly detailed enough and well-enough thought out to implement on the computer. Ethical systems and approaches make sense in terms of broad brush approaches, but (how) do people actually implement them? How can we implement them on the computer?

Knuth [1973, p.709] put it well

> It has often been said that a person doesn't really understand something until he teaches it to someone else. Actually a person doesn't really understand something until he can teach it to a computer, i.e., express it as an algorithm. ... The attempt to formalize things as algorithms leads to a much deeper understanding than if we simply try to understand things in the traditional way.

Are there ethical experts to whom we can turn?   Are we looking in the wrong place when we turn to philosophers for help with ethical questions? Should a knowledge engineer follow around Mother Theresa and ask her why she makes the decisions she makes and does the actions she does and try to implement her reasoning in an expert ethical system?

The hope is that as we try to implement ethical systems on the computer we will learn much more about the knowledge and assumptions built into the ethical theories themselves. That as we build the artificial ethical reasoning systems we will learn how to behave more ethically ourselves.

A robotic/AI approach to ethics

People have taken several approaches to ethics through the ages. Perhaps a new approach, that makes use of developing computer and robot technology, would be useful.

In the philosophical approach, people try to think out the general principles underlying the best way to behave, what kind of person one ought to be. This paper has been largely about different philosophical approaches to ethics.

In the psychological/sociological approach, people look at actual people's lives, at how they behave, at what they think, at how they develop. Some people study the lives of model human beings, of saints modern and historical. Some people study the lives of

ordinary people.

In the robotic/AI approach, one tries to build ethical reasoning systems and ethical robots for their own sake, for the possible benefits of having the systems around as actors in the world and as advisors, and to try to increase our understanding of ethics.

The two other papers at this conference represent important first steps in this new field. The paper by Jack Adams-Webber and Ken Ford [1991] describes the first actual computer system that I have heard of, in this case one based on work in psychological ethics. Umar Khan [1991] presents a variety of interesting ideas about designing and implementing ethical systems.

Of course the more "traditional" topic of "computers and ethics" has to do with the ethics of building and using computer systems. A good overview of ethical issues surrounding the use of computers is found in the book of readings [Ermann, Williams, Gutierrez 1990].

Conclusion

This paper is meant to be speculative, to raise questions rather than answer them.

• What types of ethical theories can be used as the basis for programs for ethical robots?

• Could a robot ever be said to be ethical?

• Can we learn about what it means for us to be ethical by attempting to program robots to behave ethically?

I hope that people will think about these questions and begin to develop a variety of computer systems for ethical reasoning and begin to try to create ethical robots.

Acknowledgments

I would like to thank Peter Kugel and Michael McFarland, S.J. for their helpful comments.

References

Jack Adams-Webber and Kenneth Ford, "A Conscience for Pinocchio: A Computational Model of Ethical Cognition", The Second International Workshop on Human & Machine Cognition: Android Epistemology, Pensacola, Florida, May 1991.

Isaac Asimov, "Runaround", *Astounding Science Fiction*, March 1942. Republished in Robot Visions by Isaac Asimov, Penguin, 1991.

Paul Churchland, A Neurocomputational Perspective, MIT Press, 1989.

M. David Ermann, Mary Williams, Claudio Gutierrez (eds.), Computers, Ethics, and Society, Oxford University Press, 1990.

Owen Flanagan, Varieties of Moral Personality, Harvard University Press, 1991.

Bernard Gert, Morality , Oxford University Press, 1988.

Christopher Gowans, ed., Moral Dilemmas, Oxford University Press, 1987

Stevan Harnad "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem", Minds and Machines, Vol. 1, No. 1, February 1991, pp. 43-54

A. F. Umar Khan, "The Ethics of Autonomous Learning Systems", The Second International Workshop on Human & Machine Cognition: Android Epistemology, Pensacola, Florida, May 1991.

Donald Knuth, "Computer Science and Mathematics", American Scientist, 61, 6, 1973.

W.D. Ross, The Right and the Good, Oxford University Press, 1930.