The Nature and Importance of Machine Ethics

James H. Moor

Department of Philosophy
Dartmouth College
Hanover, NH 03755
james.moor@dartmouth.edu

Normative notions have played a role in AI since its inception. The concept of intelligence is a normatively loaded notion. If every action counted as intelligent, AI would be very easy. But there are right ways and wrong ways to prove theorems, stack blocks, and make diagnoses. If there is a debate about machine ethics, then it is a debate about ethics or the capabilities of machines and not about the essential role of normativity within computing.

I will describe three different approaches to machines ethics: the *implicit ethics approach* in which computing technology is equipped with capabilities such that we judge what it is doing as ethical, the *interactive ethics approach* (or computer assisted ethics) in which computing technology and people collaborate to make better ethical decisions, and the *explicit ethics approach* in which computing technology instantiates ethical decision making which explicitly takes into account ethical factors in determining its decisions and actions.

I suspect that many will accept implicit ethics as a possibility, be dubious about interactive ethics, and be disbelievers or at least strong skeptics about an explicit ethics approach. I think the whole spectrum remains an open possibility for AI. I argue there is no good philosophical reason to oppose any of these approaches to machine ethics and that it is not an unreasonable bet that in time all three will be instantiated.

That said we need to be very realistic and tough in our assessments of the possibilities for machine ethics. We need to be clear about the nature of ethics itself. Theories that lend themselves to computing such as traditional utilitarianism are not the most philosophically defensible views. We cannot compromise the ethical requirements of a situation just to get a machine to do it. Moreover, we need to be clear about the limitations of formal systems used by machines. Standard deontic logic, for example, has some rather unintuitive consequences. We need to be good empiricists in evaluating the ethical capabilities of machines. Models of ethical reasoning that appear to be satisfactory on a priori inspection may not perform well in actual practice.

Finally, it might be asked, "Why bother with machine ethics?" Why not leave all ethical decision making to humans? The short answer is that this is not possible if we want to have the advantages that computers offer us. Often, but certainly not always, computers are reliable yet malleable control mechanisms. Very likely they will run ever more complex systems in complex situations in our society. To insist that humans must step in and make all of the "ethical decisions" at every point is not feasible and frankly may increase the probability of harm caused by complex systems. Of course, I am not suggesting that we should give up all decision making to machines. I am claiming that there may be situations such that all things considered humans are better protected by allowing machines to do some of the ethical decision making. For this reason machine ethics should be a priority in AI and computing in general.