Thoughts on the Value-Loading Problem

9/19/23

Not long ago, I read *Superintelligence* by Nick Bostrom, an overall careful accounting of the possible futures that might occur if/once some form of superintelligence arises. It is an oracular work, written nearly ten years ago, and it discusses many issues that are starting to become unavoidable. Nevertheless, I think the book is written in a certain spirit—what is often called "Scientism"—which pinpoints the problems of superintelligence accurately but prescribes the poison as the remedy.

Scientism, in brief, refers to the idea that the methods of the natural sciences are the only way to knowledge of the truth in all matters (including those—like culture, spirituality, and morality—that have traditionally been taken to be outside the purview of natural science).

In *Technopoly*, Neil Postman describes three types of society: the tool-using culture, the technocracy, and the technopoly. It is worth understanding the distinction between the first and last of these, the middle one being a sort of superposition or dialectical intermediate of the two. Perhaps the tidiest way to put the distinction is that a tool-using culture uses tools, and a technopoly is used by its tools, though this is probably an oversimplification. Or to put it another way, a tool-using culture uses its values to circumscribe its tools, while a technopoly allows its tools to circumscribe its values. In a technopoly, Scientism is a dominant strain of thought, and so is the associated idea that the scientific method "generates specific principles which can be used to organize society on a rational and humane basis," as Postman puts it.

It is in the scientistic spirit that many of the solutions to the problems of superintellgience are discussed in Bostrom's book. One of the central issues that arises with AI is the difficulty of instilling human values into a potentially powerful AI system. This is known as the *value-loading problem*, and it's a very thorny question. There are many facets to the problem, but the main one is that there is no widespread agreement on the meaning of "human values". The first hurdle in creating AI was to figure out how to get a computer to do something that typical humans can do but can't describe at a computer level of precision, like identifying pictures of dog breeds. The value-loading problem is the next level of this issue: how to get a computer to do something humans don't even agree they know how to do, like act ethically.

This is the context in which Bostrom proposes two "half-baked", in his words, solutions to the problem. I will mostly discuss not his solution, but Paul Christiano's, though they both take as a starting point the idea that even if we can't teach the AI our values, we can teach it good principles of ethical methodology. To some extent, this is more or less the only way to address the issue, since in any case we don't want the AI to learn static human values, but instead to make its beliefs subject to revision in light of new evidence or new ideas (as human values, to the extent that there are such values, have been over time). To be honest, though, I don't believe this makes the original problem any easier to solve: philosophers have not even been able to agree on the methodology of ethics, since, for example, there is still no agreement on whether virtue ethics, deontological ethics, or consequentialist ethics should be the dominant principle. In other words, this is just a kicking of the can down the road.

Paul Christiano's proposal starts off by assuming the consequentialist perspective. Consequentialism is also my horse in the race, but I will note that the choice seems to be more of a contortion of ethics to technology than the other way around, since it seems like consequentialism is the easiest ethical framework to teach to a computer, especially a computer/program designed to optimize for its objectives. The essence of the proposal is this: we are looking for a utility function U to teach to our AIs as the basis of their ethics. To do this, we 1) make a mathematical model of a single human brain and 2) specify an environment in which that human brain will spend centuries reflecting on the nature of the good. Then, U is whatever utility function that brain comes up with. The appeal of the proposal is that it gives a more or less precise definition of U without necessarily specifying how to compute U in practice. It has the flavor of a mathematical statement like "let n be the number of primes less than 2^{2^100}".