8. GOOD MACHINE, NICE MACHINE...

I.J. Good, a brilliant early researcher into artificial intelligence (AI), wrote in a 1965 article, "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make provided that the machine is docile enough to tell us how to keep it under control."

Scenarios in which highly advanced machines threaten the survival of humanity, once the exclusive domain of science fiction, have increasingly generated interest and concern. If machines become "smarter" than humans, and capable of redesigning themselves, their evolution could quickly spiral out of control, producing what researchers now refer to as a "technological singularity." For instance, the home page of the Machine Intelligence Research Institute (MIRI) announces that the institute's researchers are trying to make sure that intelligent machines will not develop in unintended, harmful ways. MIRI researchers are focusing on the notion of a machine using principles rather than genetic algorithms (GA) to reason about its own behavior by. One might argue, however, that a superintelligent machine would by definition be able to outsmart its human nannies.

Simply building a known theory of ethics into a superintelligent machine might not ensure that it will treat us well, and may even kill us. Muehlhauser and Helm from MIRI point out in a 2012 document that, were a machine "superoptimizer" to optimize one of the familiar moral theories, the results might be far from desirable. Optimized hedonistic utilitarianism might lead the machine to hook us all up to machines that continuously administer chemical or neurological experiences; while optimized negative utilitarianism (to minimize suffering, rather than maximize pleasure), might lead it to euthanize all humans painlessly, "no humans, no suffering."

Case from the 2015 Intercollegiate Ethics Bowl National Competition