Some have proposed https://ieeexplore.ieee.org/document/1667949 49 https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-001.pdf 50 https://dl.acm.org/doi/10.5555/1811259.1811320 51 https://www.researchgate.net/publication/259084235_What_Statistics_Could_Do_for_Ethics_The_Idea_of_Common_Sense_Processing_Based_Safety_Valve 52 that we teach machines a moral code with case-based machine learning. The basic idea is this: Human judges would rate thousands of actions, character traits, desires, laws, or institutions as having varying degrees of moral acceptability. The machine would then find the connections between these cases and learn the principles behind morality, such that it could apply those principles to determine the morality of new cases not encountered during its training. This kind of machine learning has already been used to design machines that can, for example, detect underwater mineshttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.6963&rep=rep1&type=pdf 53 after feeding the machine hundreds of cases of mines and not-mines.

There are several reasons machine learning does not present an easy solution for Friendly AI. The first is that, of course, humans themselves hold deep disagreements about what is moral and immoral. But even if humans could be made to agree on all the training cases, at least two problems remain.

The first problem is that training on cases from our present reality may not result in a machine that will make correct ethical decisions in a world radically reshaped by superintelligence.

The second problem is that a superintelligence may generalize the wrong principles due to coincidental patterns in the training data.https://intelligence.org/files/AIPosNegFactor.pdf 54 Consider the parable of the machine trained to recognize camouflaged tanks in a forest. Researchers take 100 photos of camouflaged tanks and 100 photos of trees. They then train the machine on 50 photos of each, so that it learns to distinguish camouflaged tanks from trees. As a test, they show the machine the remaining 50 photos of each, and it classifies each one correctly. Success! However, later tests show that the machine classifies additional photos of camouflaged tanks and trees poorly. The problem turns out to be that the researchers' photos of camouflaged tanks had been taken on cloudy days, while their photos of trees had been taken on sunny days. The machine had learned to distinguish cloudy days from sunny days, not camouflaged tanks from trees.

Thus, it seems that trustworthy Friendly AI design must involve detailed models of the underlying processes generating human moral judgments, not only surface similarities of cases.

Further reading:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk