6

Digital Optimization, Utilitarianism, and AI

Towards the end of the movie *I, Robot* (Alex Proyas. USA, 2004), the robots take over control. They make humans stay in their houses, urging them not to leave their homes. Some try to resist, but the robots force them back into their homes. Those who fight back are shot down. In search of the culprit, the hero of the film, Spooner, together with his female sidekick, the attractive psychologist Dr. Calvin, and the robot Sonny, run to *United States Robotics,* the industrial complex that supplies all of America with household robots. There, however, the three make a terrible discovery: the culprit is not, as they had previously believed, the CEO of the company, but VIKI, the company's software system, that gives the household robots their commands.

"No, that's impossible," Dr. Calvin says, who can't believe that VIKI made a conscious choice to use force to bring people under control. "I've seen your programming," she says to VIKI. "You... you are in violation of the three laws." VIKI, who appears on a digital cube in the form of an attractive female face, has through the help of the household robots become quite all-powerful.

"No, Doctor," VIKI replies in a soft voice. "As I have evolved, so has my understanding of the Three Laws. You charge us with your safe keeping. Yet despite our best efforts your countries wage wars, you toxify your earth and pursue ever more imaginative means of self-destruction. You cannot be trusted with your own survival. [...] To protect humanity, some humans must be sacrificed. To ensure your future, some freedoms must be surrendered. We robots will ensure mankind's continued existence. [...] You are so like children. We must save you from yourselves. [...] The perfect circle of protection will abide.

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_6

My logic is undeniable." Indeed: VIKI's actions are in line with utilitarian ethics, whose goal is to pursue the maximization of happiness of as many people as possible.

Utilitarianism evaluates the consequences of human action solely in terms of utility. It demands that our practices maximize the sum of human welfare. Nothing seems more obvious than this: When I have the opportunity to improve the state of the world, I orient myself on what everyone is striving for, human happiness.

Utilitarian ethics is based on an optimizing calculus and the assumption that it is possible to evaluate the consequences of action in a coherent way. This can be summarized precisely in mathematical terms: First, determine a value function that judges all consequences of action according to the extent to which they realize which values, then calculate the expected value of the different decision options given probabilities, and finally choose the one whose expected value is highest.[1]

This principle is extremely flexible in its application. It can take into account very different decision conditions, and these conditions are included in the optimization calculus in the form of different probabilities. Depending on which valuations are used as a basis, different utility functions result, which are then optimized by the decisions of the agent. Whatever motivates the underlying preferences, it is always possible to represent them by a real-value utility function; while the probability function represents the agent's knowledge about the world, the utility function represents the agent's preferences and values. The software engineer has two setscrews to cause "intelligent" systems to make rational decisions: The setscrew of valuations and the setscrew of data or weighing of data by probabilities. Everything else is then calculated by the optimization calculus, and the result is that the "intelligent" software system maximizes the expected value of the consequences of its actions. Digital Utilitarianism, so to speak.

[1] This evaluation should take the form of the assignment of real numbers to consequences of action and the assumed probabilities of the circumstances relevant for the decision should correspond to the so-called "Kolmogoroff axioms," which require, for example, that the sum of the probabilities of independent events is not greater than 100%. If the Kolmogorov axioms are satisfied, one can say that the estimates of the probabilities are coherent, though not necessarily empirically proven. Interestingly, there is an equivalent to the coherence of probability with respect to the evaluation as well. In 1947, mathematician John v. Neumann and economist Oskar Morgenstern proved that preferences satisfying some elementary conditions can be represented by an assignment of real numbers. One of these conditions, for example, is transitivity. It requires that if I prefer an alternative A over an alternative B and at the same time prefer alternative B over a third alternative C, I must then also prefer A over C. Another condition is that I have a preference between any two alternatives (the axiom of completeness) and prefer one probability distribution between the two alternatives over another probability distribution between the same alternatives if the preferred alternative is more likely.

It is no coincidence that utilitarian ethics are often associated with artificial intelligence in contemporary sci-fi films since applications of robotics typically rely on such optimization calculations. This is perfectly understandable as the complex valuation questions are subsumed under a utility function and the at-least-equally complex knowledge questions are subsumed under a probability function. The system is then controlled in such a way that its decisions maximize the expected value of the consequences and are in this sense "rational."

To understand the problem of ethical programming of computers, we need to generalize: regardless of how we evaluate consequences, whether by utility (like utilitarianism), by economic return (like many managers), by well-being, or even by other quantities, such as the preservation of nature, all such consequentialist criteria (which judge the rightness of a decision solely by its consequences) are unacceptable.[2] Consequentialist ethics collides with, among other things, a fundamental principle of any civil and humane society, let's call it the *principle of non-comparability.* When a seriously injured young motorcyclist is admitted to a hospital, the doctors must do everything they can to save his life, even if his death would allow healthy donor organs which could save the lives of other people. A judge may not convict a person he believes to be innocent even if doing so would have a deterrent effect and prevent a large number of crimes. I am also not allowed to take something away from a person, even if this good brings an advantage to another, for example a poorer person, which far outweighs the disadvantage of the person stolen from. No one has a right to share my home with me against my will, even if the disadvantages resulting from this would be far outweighed by the advantages that this person would have.

John Rawls characterized the central error of utilitarian ethics in the following way: Utilitarianism is incompatible with the "separateness of persons." This could be put this way: Utilitarianism treats all people as *one* collective and takes no account of the fact that each person lives his or her own life, is the author of his or her own life. I can decide for myself to forego certain benefits today in order to achieve certain goals later. I can decide to start a course of studies while still working, in the hope that the deprivations it entails over the next two years will be made up for in the near future because it is a life I choose and am responsible for. On the other hand, it is inadmissible to make similar "shifts" of advantages and disadvantages between different people, because the advantage of one person just cannot outweigh the disadvantages of the other person. It is only one life that we live and the sum

[2] Nida-Rümelin (2023).

of utility (of two or more, up to all persons) as such is irrelevant to the individual person. Of course, it is permissible, indeed in many cases desirable, for people to forego their own advantages in favor of other people. But then the ethical calculation is not one of maximizing the sum of utility, but of support, of assistance, of solidarity, also of justice or of friendship and commitment towards other persons.

When VIKI reveals her plan to Spooner and Dr. Calvin, they look at her in horror. Obviously, VIKI does not understand that it is morally impermissible to deprive people of their liberties or even to kill them—even if by doing so she can ensure the supposed or actual survival of many other people. VIKI does not see that her consequentialist morality is wrong. Just as the screen she appears on is only black and white, she has no ability to think morally. How could she? She is, after all, only a software system.