10

Why AIs Fail at Moral Dilemmas

In the control center of the US Robotics company, Spooner, the psychologist Dr. Calvin, and the (good) robot Sonny fight against an army of (evil) robots, all of which are controlled by the radically utilitarian software system VIKI. With an eerie red glow inside their metal bodies, they take decisive action against Spooner, Calvin, and Sonny. But for all their numerical superiority, the good guys have something valuable with which they can first destroy VIKI and, by extension, all evil robots: a kind of syringe that, when properly placed, can instantly turn off VIKI. Just as the robot Sonny is about to insert the syringe into VIKI's central computer, Calvin slips. With the last of her strength, she manages to hold on to a metal beam. Below her it goes down 100 Meters—if she let's go she is dead. Sonny the robot must decide: Should he kill VIKI—and thus save humanity—or save the life of Dr. Calvin, a single human? Sonny is visibly overwhelmed. He doesn't want to let Dr. Calvin die, but, on the other hand, he wants to protect mankind from VIKI. For Spooner, however, it's clear what should be done: "Save Calvin!" he shouts to Sonny.

As we have seen, the practice of deliberation cannot be algorithmized. This is especially evident in situations involving moral dilemmas. A moral dilemma exists when there is no satisfactory resolution to a moral conflict. When a person has two or more obligations that she cannot meet together and she feels guilty whatever she does, then there is a moral dilemma. She regrets not fulfilling the obligation even though there was another obligation that made it impossible for her to fulfil it. In moral dilemma situations, the obligations persist; they are not removed by the conflict.

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_10

Not every moral conflict is a genuine moral dilemma. In many cases, it is possible to arrive at a clear recommendation by weighing different moral reasons. Weighing conflicting moral reasons need not necessarily lead to a genuine moral dilemma: I promised to take my daughter to the movies this afternoon. On the way there, I get a call that my other daughter has a high fever and needs taking to the doctor. After a brief deliberation, I decide to prioritize the duty to help the sick daughter over the duty to keep my promise. There is no moral dilemma here, but merely the conflict of two grounds of obligation, which, however, is clearly to be resolved in favor of one of the two. One could say that the obligation to keep my promise to take one daughter to the movies is nullified by the priority obligation to help the sick daughter.

In some cases, however, there seems to be no resolution of such a moral conflict. A genuine moral dilemma arises when conflicting grounds of obligation persist and I am, in a sense, guilty regardless of what I do. Ancient tragedy literature developed particular excellence in fictionalizing such dilemma situations, which inevitably lead to moral guilt. A striking, if gruesome, example is William Styron's novel *Sophie's Choice*. This book is about a Jewish woman (Sophie) who is taken to a concentration camp by the Germans during World War II. The sadistic concentration camp warden gives Sophie a choice: she must choose which one of her two children to keep and which one would be gassed. If she chooses neither, both must die. Sophie chooses to save her son. No matter what Sophie decides to do, she will burden herself with immense guilt: either because she sacrifices one of the children for the sake of the other or because she fails to prevent the murder of one of the children who would otherwise live. Sophie survives. But even years later, she has not been able to forgive herself and eventually kills herself.

The British ethicist Bernhard Williams has presented a variant of this dilemma.[1] On a trip to South America, the tourist Jim passes through a small town. He sees 20 tied up Indians standing against a wall. In front of them are several men in uniforms. Their leader, Pedro, explains to Jim that the men must be shot to make an example after protesting against the government. Pedro now offers Jim, as a guest in this country, the honor of shooting one of the Indians. If he does so, the others will be set free. If he shoots none, all 20 will die, as planned. Jim can neither escape nor bargain with Pedro. He must choose. The Indians ask him to accept the offer. What ought Jim to do? No matter what he does, he is guilty, either because he makes himself the murderer of a human being or because he becomes responsible for the death of 20 Indians.

[1] Smart and Williams (1973).

Williams makes a point of noting that the mere fact that the tourist refuses to participate in this gruesome game does not mean that he can be accused of causing the deaths of 20 people. The guerrilla leader will always remain the one who brought about this situation in the first place. Still, one will not be reassured by the fact that doing nothing spares moral guilt.

Utilitarian (consequentialist) ethics rejects the existence of genuine moral dilemmas. The reason is obvious. If action is judged according to the optimization criterion (maximize the expected value of utility) there can be no conflict, at best indifference: It may be that two courses of action have the same maximum expected utility value. In order for the person to be able to act and not starve to death like Buridan's ass,[2] the utilitarian motivated person will choose or roll the dice on one of the two options between which he is indifferent.

Genuine moral dilemmas are characterized by the fact that one cannot roll the dice between conflicting obligations; the situation is too serious for that. One could also say that the decision is *existential* insofar as it provides information about the fundamental attitude of this person. There is much to be said for interpreting the existence of moral dilemmas as an expression of the general non-computability of our moral deliberations. Digital computers are defined as Turing machines and deliver unambiguous results. For this reason alone, they cannot be a model of practical reason.

The helplessness of robots in the face of real moral dilemmas is also a recurring motif in films. Not only Sonny is at a loss at the end of the film as to whom he should save (one single human being or possibly the freedom of an entire city), but other artificial beings also fail in such situations. But unlike Sophie from the novel *Sophie's Choice,* robots are not expected to feel guilty for the rest of their lives and end up committing suicide—like Sophie—because they cannot live with the feeling of having acted wrongly.

[2] "Buridan's ass" is a Persian parable that tells of a donkey that cannot decide between two haystacks of equal size and distance and eventually starves to death.