There is also this solution: the creators of AI do not want to build conscious machines, but only intelligent machines, ones whose job is to complement human computing skills and thus create a human-friendly world. Here, however, the problem arises whether AI will be able to understand the human world of values and the specific nature of the moral obligations that result from those values. Will artifacts equipped with artificial intelligence be able to recognize the world of human values in their complex nature and will they be able to read the principles that govern the process of making difficult decisions by man? We still do not have satisfactory answers to these questions, even though this issue has been taken up by many researchers, including Eliezer Yudkowsky and Bostrom. When it comes to the knowledge of values, an important role is played by axiological intuition, a distinctly human ability and one which machines lack. At this point, however, another question comes into view, whether it is possible to write down the human structure of values and express it in the form of an algorithm, which can then be inscribed in the operational structure of an intelligent machine.

An advanced-level, super-intelligent AI can perform complex computational operations involving the collecting and segregating of data, while not being aware of its distinctiveness.[36] It seems reasonable to argue, however, that an entity cannot be the bearer of responsibility without having moral awareness. To solve complex axiological and moral dilemmas, a machine, besides intelligence, must also have consciousness. A conscious machine, capable of recognizing values and solving moral dilemmas, would need to have a will capable of choosing and acting independently of humans.[37] Autonomous machines, fully independent of the man controlling their operation, seem to be—from the perspective of the designers—an undesirable coincidence, unless of course this is also an unintended result of AI techno-evolution, one that we cannot control. This kind of operation, independent of the constructor and the user, is treated as a design error. The argument from the "designer's unintentional error" has had its reflections in popular culture, in narratives where autonomous robots want to take control over people, who are "less" intelligent.

The creation of an artificial intelligence that imitates the human world of values has also a negative side to it. Apart from positive values, axiology distinguishes negative values, which result in the desire for destruction, death, falsehood, and the creation of distorted (demonic) images of the sacred. Consequently, this leads to behaviors that we consider morally wrong, among them the propensities to be aggressive, to cheat, and to treat other people instrumentally. We cannot assume that man represents the highest level of consciousness and moral competence. Inscribing the human world of values into an intelligent machine can prove problematic. This is the rationale behind the building of "ethical robots," namely, that the super-intelligent machines thus created will be devoid of human flaws.[38] However, this means that this type of ethics becomes a utopian AI construct equipped with an "angelic" set of qualities (such as kindness, forbearance, the ability to cooperate) focused on the fostering of community values while devoid of human "demonic" tendencies. In other words, such projects dehumanize machines and make them into entities which are "artificial" in another sense of the word.

Such a project was created by Yudkowsky, who presented the development of intelligent machines the operation of which is based on positive va-

[36] Yudkowsky, *Complex Value Systems are Required to Realize Valuable Futures*, 38n.

[37] See Susan Schneider, *Artificial You: AI and the Future of Your Mind* (Princeton and Oxford: Princeton University Press, 2019), 16n.

[38] See Yudkowsky, *Complex Value Systems are Required to Realize Valuable Futures*, 40n.

lues. To realize this purpose, he used the “semantics of external reference,” which demonstrates that an increase in a machine’s knowledge about values such as kindness will in due course result in an increase in actual kindness in that machine’s operation. In the case of AI, this is the knowledge entered by programmers into the IT system. This solution does not work in the case of human beings in that it does not take into account the factor of free will and the situational dynamics in which the subject who is making the moral decision finds themselves. In fact, it is a machine variant of ethical intellectualism. Intelligent machines, like humans, may know the rules and yet act without conforming to them. Following Aristotle, it is necessary to distinguish in this case between machine techne knowledge and machine praxis knowledge. The former is responsible for collecting and segregating data, while the latter for choosing and action.

What would the phronetic knowledge of machines consist in? Yudkowsky introduces the formula of “semantics of causal significance” in this case. It assumes that AI should not do exactly what programmers have written into it, but something similar, something that results from a certain skill in solving difficult situations. Developers are not able to take into account all circumstances, so it should be assumed that the AI will be equipped with the ability to modify decisions. This means that the solution proposed by AI does not have to suit us. As the source principle of AI normativity, Yudkowsky adopts: “do the right thing,” which is based on the principle of reflective equilibrium. This is one variant of the Greek rule of the ethics of moderation, introduced by Aristotle into ethical thought. Here, however, the problem arises whether the reflective balance of the man being and that of AI are based on the same principles. The premise of the ethics of moderation is to have human experiences resulting from corporeality and communion with other people (ethics of friendship). For AI, experience will only be information inscribed in the system, not an individual or personal experience. Thus, the thesis that a machine can “think and act like a human” is merely an approximation.