WHY DO WE NEED TO REVISE THE GOALS OF AI?

AI has provided many very useful solutions, so it is therefore the technology that is currently being pinned as the greatest hope for improving human life. At the same time, however, the proliferation of AI systems is causing increasing risks. The main philosophical determinants of these threats can be boiled down to four fundamental philosophical problems.

First, people are increasingly losing their decision-making power (i.e., a loss of primary decision agency). Second, they are also gradually losing control over their cognitive processes (i.e., a loss of primary epistemic agency). Third, they are gradually losing control over technical systems and the development of new technology (i.e., a loss of control). Fourth, AI systems have an overwhelming advantage when it comes to processing large amounts of data, allowing them to solve problems more efficiently and gain a significant advantage (i.e., the supremacy of computing power).

The lack of clarity surrounding the risks posed by AI has partly resulted from the peculiar ideology driving the development of AI. The project to create a "synthetic human," which is an explicit or implicit premise of the AI discussion, appeals to myths and metaphysical longings, and it may even be a challenge to the Creator. However, it should be noted that what is actually being undertaken is the construction of an ideological envelope for the AI program, because it effectively obscures the actual goals and problems by diverting people's attention to intriguing but irrelevant aspects. After all, the very use of the term "artificial intelligence" is very ideologically loaded, and this is particularly evident when comparing it to the program's original label, namely cybernetics.

It should also be noted here that according to the rationality of technical activities, AI is designed with the aim of achieving specific, often hidden, business goals. This makes it possible to develop the program, although such a choice leads to a reduced axiology for the activities undertaken. More and more voices claim that it is possible to account for other hierarchies of values and goals in the design of technical systems and thus humanize technology, but this has not led to a wider discussion, probably because of the limited awareness of the risks. Instead, a significant problem is a lack of ideas for an AI model that will benefit societies in the long term rather than just profit a narrow group in the short term. In other words, the current business model for AI development has no competitors, and this is a serious problem that needs to be confronted when addressing the threats of AI.

TOWARD BENEFICIAL AI

PRELIMINARIES

To organize the philosophical considerations surrounding AI, we should first pose some basic questions. In particular, we should question what we expect from AI:

(1) Do we need superhumans?
(2) Do we desire perfect slaves?
(3) Do we want synthetic humans?
(4) Is a symbiotic coexistence appealing?

The answers to such questions reveal important differences between philosophers, tech visionaries, and AI researchers and engineers, so we need to clarify some fundamental questions. We should therefore pose some philosophical questions rather than the previous set of questions:

(1) What kind of AI do we need as humanity?
(2) What kind of relationships do we need?
(3) What values should be preferred?
(4) Is the anthropocentric viewpoint on AI justified?

The first question sets the perspective for the whole deliberation, while the subsequent questions are refinements of this perspective, which we will call *Beneficial AI*. As we understand it, this concept represents the AI that we need as humans for beneficial development in the long term.

BENEFICIAL AI

What is Beneficial AI? Stuart J. Russell’s definition states that a beneficent machine, one driven by Beneficial AI, realizes our objectives rather than its own.[17] Of course, it would be simpler if we knew what we really wanted,[18] but this is not the case. Thus, Russell proposes some tentative guidelines under which beneficially inclined AI systems should operate. He qualifies his proposal by admitting that these are just guidelines rather than rules of any sort, because he fears that these may be taken like Isaac Asimov's notorious laws of robotics, which were originally proposed in Asimov's work *I, Robot*[19] and amended several times. Such an approach risks pushing the whole idea of Beneficial AI down the rabbit's hole.[20]

Russell's rules for Beneficial AI, which are not indented as laws,[21] state, firstly, that the machine objective is to maximize the realization of human preferences. Secondly, they assert that the machine does not know initially what these preferences should be. Thirdly, they posit that the machine learns these preferences from human behavior. Russell is fully aware that we do not actually know how to do this, technically, conceptually, or otherwise, but he is sure that if we want to avoid the potential calamities of unbridled AI development, we must pursue this endeavor.

The concept of Beneficial AI has also been elaborated in the Asilomar AI Principles.[22] This list of recommendations from the Beneficial AI Conference is a lengthy one,[23] but a few of the more important ones include:

(1) Ethics: AI systems should be designed and operated such that they are compatible with the ideals of human dignity, rights, freedoms, and cultural diversity.

(2) Value alignment: Highly autonomous AI systems should be designed such that their goals and behaviors are guaranteed to align with human values throughout their operation.

(3) Shared benefits: AI technologies should benefit and empower as many people as possible.

[17] See Russell, *Human Compatible*.

[18] See ibidem.

[19] See Isaac Asimov, *I, Robot* (Garden City, New York: Doubleday & Company, Inc., 1950).

[20] This expression is used especially in the phrase "going down the rabbit hole" or "falling down the rabbit hole." It is a metaphor for something that transports someone into a wonderful (or troublingly) surreal state or situation (see "Rabbit Hole,"Dictionary.com, https://www.dictionary.com/e/slang/rabbit-hole/). The expression dates back to the famous 1865 classic *Alice's Adventures in Wonderland* by Lewis Carroll (see Lewis Carroll, *Alice's Adventures in Wonderland* Cambridge: Cambridge University Press, 1865), who was less famously a mathematician.

[21] See Russell, *Human Compatible*, 172.

[22] "Asilomar AI Principles," Future of Life Institute, August 11, 2017, https://futureoflife.org/2017/08/11/ai-principles/.

[23] See "Beneficial AI 2017."

As we said, the list is long, with it comprising twenty three areas grouped into research issues, ethics and values, and long-term issues. These ideas are certainly on the mark, and one could say they benefit the discussion about AI. Anyway, we more or less know what Beneficial AI should be, but the problem is that we are not sure how to realize it. This is where the concept of domestication or taming comes in.