ARTIFICAL INTELLIGENCE: THREATS AND PROMISES

O. Snytnikova, *IN-33*

According to some authors, who research the field of Artificial Intelligence (AI), one of the greatest dangers of AI is that people conclude too early that they understand it. The critical inference is not that AI is hard, but that, for some reason, it is very easy for people to think they know far more about Artificial Intelligence than they actually do.

Artificial Intelligence is not settled science. And on the topic of global catastrophic risks of Artificial Intelligence, there is virtually no discussion in the existing technical literature. When something is universal enough in our everyday lives, we take it for granted to the point of forgetting it exists. But when something new appears, it may cause some difficulties.

We can observe many threats and promises in different fields of its influence. But it is a risky intellectual endeavor to predict specifically how a benevolent AI would help humanity, or an unfriendly AI harm it. Another important point is the possibility of technical and philosophical failures.

It would be a very good thing if humanity knew how to choose into existence a powerful optimization process with a particular target. Or in more colloquial terms, it would be nice if we knew how to build a nice AI.

To describe the field of knowledge needed to address that challenge, scientists have proposed the term "Friendly AI". But common reaction is that people immediately declare that Friendly AI is an impossibility, because any sufficiently powerful AI will be able to modify its own source code to break any constraints placed upon it.

There are any number of vaguely plausible reasons why Friendly AI might be humanly impossible, and it is still more likely that the problem is. But one should not so quickly write off the challenge, especially considering the stakes.

Lytvynenko G.I. EL *advisor*