[The original, German version of this article appeared on November 21, 2021, as part of the AI column “Aus dem Maschinenraum der KI” in the economics section of the German Sunday newspaper Welt am Sonntag, p. 24.]

[Translated with www.DeepL.com/Translator (free version) - AI technology made in Europe, see https://en.wikipedia.org/wiki/DeepL_Translator, and subsequently polished and modified by the author.]

Messages from the AI engine room

AI Alignment

Artificial intelligence needs a foundation of human values, including human dignity and moral agency

Can machines think? This question was raised by British mathematician Alan Turing (1912-1954) at the beginning of his essay "Computing Machinery and Intelligence," published in 1950. The article appeared seven years after he had cracked the German Wehrmacht's Enigma, which probably shortened the Second World War by months and already merits Turing a place in the history books.

Among many other ground-breaking ideas, Alan Turing has also developed the "Imitation Game." The idea is on first sight a bit airy-fairy but grew into fame far beyond the world of science as the so-called “Turing Test” and was the harbinger of a completely new scientific field: Artificial Intelligence (AI).

The imitation game is the following thought experiment: If a person communicating via a computer with an invisible counterpart fails to fiigure out whether this counterpart is a human being or a machine, then, according to Turing, it must be a truly intelligent machine whose thinking ability is equal to that of a human being. Turing was convinced that by the year 2000 computers would be able to fool humans in such a conversation. It did not happen quite that quickly, as we know. But the Turing test and its many variants are regarded as the definitive yardstick for artificial intelligence.

Back in the 1950s, Turing's thought experiment may have been provocative. Today, there is no doubt among many experts that it is only a matter of time before it is passed. For example, Duplex, the assistance software introduced by Google in 2019, can not only place online orders, but also call hair salons and restaurants, book appointments and reserve tables, including deceptively real pauses in speech and interspersed "ahem." The people on the other end of the line had no idea they were talking to a computer.

Current AI systems can already answer quote sophisticated questions. They can help us write song lyrics, poems, newspaper articles and simple computer programs. Yes, OpenAI’s Text AI GPT-3 is already creating the first credible, albeit very short, dialogs with AI decals of famous personalities such as Tom Hanks. For example, Paras Chopra, founder of the Indian tech company Wingify, asks Hank-AI what its favorite role is. Hank-AI replied, "As we get older, we realize how short life really is and how much more there is to see and do. I think in a way it can be a little depressing, but in a way that makes us appreciate a little more every day and the people around us. So my answer is that the best role for me is the next one because I want to break new ground. And remember: hold on! It only gets better

from here. " Can you distinguish this answer from a possible answer from the real Tom Hanks?

Sam Altman, CEO of OpenAI, believes that one of the next GPT generations could already be able to successfully complete the Turing test — however, he also suspects that it may not be worth the effort.

The fact that no AI — contrary to Turing's own prediction — has really passed the test can be seen as proof that we underestimate natural intelligence. It can also be viewed as proof that we humans constantly improve and are not easily bluffed by a computer program. Or both!

Who knows, maybe an AI system that is smart enough to pass the Turing test would also just be smart enough to not pass it — thereby fool us humans. Would that be troubling? Not really. After all, being smart does not imply being reasonable. And the later is what what we should care about. AI should be aligned with our values. We can and must bring the values of the European Enlightenment, human dignity, freedom and democracy into AI research and science. This will be the real Turing test.

And for the weekend, I recommend laymen and experts alike: Make yourself comfortable and watch the Oscar-winning film "The Imitation Game" again with Benedict Cumberbatch as Alan Turing.

Kristian Kersting is Professor of AI and Machine Learning at TU Darmstadt, co-director of the Hessian Center for AI (hessian.ai), and winner of the "German AI Award 2019". His AI column "Aus dem Maschinenraum der KI" appears regularly in the German Sunday newspaper Welt am Sonntag.