Artificial Intelligence with an “I don’t know” answer

Damjan Krstajić

The saying “I know that I know nothing” is attributed to the Greek philosopher Socrates, and one might wonder what the actual point of it is.

As a young man Socrates visited and conversed with various scholars in ancient Athens, and he realised that they were all talking about how much they knew, but really they did not. Therefore, by admitting how much he does not know he becomes wiser than them. In simple terms, the main lesson from Socrates is that by accepting the limitations of our knowledge we would avoid ignorance and wrongdoing. I think the same applies to artificial intelligence (AI).

Substantial investments and many hopes are placed in AI research today. We are overwhelmed with news about how AI will transform future science and our everyday life. However, is AI capable of responding with “I don’t know”? How are the limitations of AI being defined? Who is responsible for doing that? What are the benefits of knowing such limitations? As far as I am aware, there is not a great deal of interest in tackling such issues.

I think that there are two important questions related to the use of AI:

a) When to trust AI?

b) How to develop and extend AI continuously as a dynamic system?

In my opinion, knowing the limitations of AI is crucial to answering both questions, and it is the responsibility of the creators of AI systems to inform users about them. However, defining the limitation of AI is not an easy task at all. The creators of AI systems are probably aware that AI does have limitations, but are they able to define them? And how?

I think that by asking AI creators to design AI with an “I don’t know” answer will force them to come up with a solution regarding the limitations of the AI systems they develop. For us humans, the answer “I don’t know” comes naturally as the result of our introspection, and we rarely hear anyone

asking themselves: “How do I know that I don’t know?”. However, the question of how AI would know that it does not know is probably ignored because there is no obvious answer to it, as AI does not have the luxury of introspection.

References:

1. West, Thomas G., and Grace Starry West, eds. *Four Texts on Socrates: Plato's Euthyphro, Apology, and Crito, and Aristophanes' Clouds*. Cornell University Press, 1998.