AI

Yesterday at lunch a friend asked me what tech trend he should pay attention to but was probably ignoring.

Without thinking much I said “artificial intelligence”, but having thought about that a bit more, I think it’s probably right.

To be clear, AI (under the common scientific definition) likely won’t work. You can say that about any new technology, and it’s a generally correct statement. But I think most people are far too pessimistic about its chances - AI has not worked for so long that it’s acquired a bad reputation. CS professors mention it with a smirk. Neural networks failed the first time around, the logic goes, and so they won’t work this time either.

But artificial general intelligence might work, and if it does, it will be the biggest development in technology ever.

I’d argue we’ve gotten closer in lots of specific domains - for example, computers are now better than humans at lots of impressive things like playing chess and flying airplanes. But rather than call these examples of AIs, we just say that they weren’t really that hard in the first place. And to be fair, none of these really feel anything like a computer that can think like a human.

There are a number of private (or recently acquired) companies, plus some large public ones, that are making impressive progress towards artificial general intelligence, but the good ones are very secretive about it.

There are certainly some reasons to be optimistic. Andrew Ng, who worked or works on Google’s AI, has said that he believes learning comes from a

single algorithm - the part of your brain that processes input from your ears is also capable of learning to process input from your eyes. If we can just figure out this one general-purpose algorithm, programs may be able to learn general-purpose things.

There have been promising early results published from this sort of work, but because the brain is such a complex system so dependent on emergent behavior it's difficult to say how close to the goal we really are. We understand how individual neurons work pretty well, and it's possible that's all we need to know to model how intelligence works. But the emergent behavior of 100 billion of them working together on the same principles gets extraordinarily complex, and difficult to model in software. Or, as Nick Sivo says, "it's like reverse engineering the latest Intel processor with only the basic knowledge of how a transistor works." It's also possible that there's some other phenomenon responsible for intelligence, and the people working on this are on the wrong track.

The biggest question for me is not about artificial intelligence, but instead about artificial consciousness, or creativity, or desire, or whatever you want to call it. I am quite confident that we'll be able to make computer programs that perform specific complex tasks very well. But how do we make a computer program that decides what it wants to do? How do we make a computer decide to care on its own about learning to drive a car? Or write a novel?

It's possible--probable, even--that this sort of creativity will be an emergent property of learning in some non-intuitive way. Something happened in the course of evolution to make the human brain different from the reptile brain, which is closer to a computer that plays pong. (I originally was going to say a computer that plays chess, but computers play chess with no intuition or instinct--they just search a gigantic solution space very quickly.)

And maybe we don't want to build machines that are concious in this sense. The most positive outcome I can think of is one where computers get really good at doing, and humans get really good at thinking. If we never figure out how to make computers creative, then there will be a very natural division of labor between man and machine.