Machine intelligence, part 1

This is going to be a two-part post—one on why machine intelligence is something we should be afraid of, and one on what we should do about it. If you're already afraid of machine intelligence, you can skip this one and read the second post tomorrow—I was planning to only write part 2, but when I asked a few people to read drafts it became clear I needed part 1.

WHY YOU SHOULD FEAR MACHINE INTELLIGENCE

Development of superhuman machine intelligence (SMI) [1] is probably the greatest threat to the continued existence of humanity. There are other threats that I think are more certain to happen (for example, an engineered virus with a long incubation period and a high mortality rate) but are unlikely to destroy every human in the universe in the way that SMI could. Also, most of these other big threats are already widely feared.

It is extremely hard to put a timeframe on when this will happen (more on this later), and it certainly feels to most people working in the field that it's still many, many years away. But it's also extremely hard to believe that it isn't very likely that it will happen at some point.

SMI does not have to be the inherently evil sci-fi version to kill us all. A more probable scenario is that it simply doesn't care about us much either way, but in an effort to accomplish some other goal (most goals, if you think about them long enough, could make use of resources currently being used by humans) wipes us out. Certain goals, like self-preservation, could clearly benefit from no humans. We wash our hands not because we actively wish ill towards the bacteria and viruses on them, but because we don't want them to get in the way of our plans.

(Incidentally, Nick Bostrom's excellent book "Superintelligence" is the best thing I've seen on this topic. It is well worth a read.)

Most machine intelligence development involves a "fitness function"—something the program tries to optimize. At some point, someone will probably try to give a program the fitness function of "survive and reproduce". Even if not, it will likely be a useful subgoal of many other fitness functions. It worked well for biological life. Unfortunately for us, one thing I learned when I was a student in the Stanford AI lab is that programs often achieve their fitness function in unpredicted ways.

Evolution will continue forward, and if humans are no longer the most-fit species, we may go away. In some sense, this is the system working as designed. But as a human programmed to survive and reproduce, I feel we should fight it.

How can we survive the development of SMI? It may not be possible. One of my top 4 favorite explanations for the Fermi paradox is that biological intelligence always eventually creates machine intelligence, which wipes out biological life and then for some reason decides to makes itself undetectable.

It's very hard to know how close we are to machine intelligence surpassing human intelligence. Progression of machine intelligence is a double exponential function; human-written programs and computing power are getting better at an exponential rate, and self-learning/self-improving software will improve itself at an exponential rate. Development progress may look relatively slow and then all of a sudden go vertical—things could get out of control very quickly (it also may be more gradual and we may barely perceive it happening).

As mentioned earlier, it is probably still somewhat far away, especially in its

ability to build killer robots with no help at all from humans. But recursive self-improvement is a powerful force, and so it's difficult to have strong opinions about machine intelligence being ten or one hundred years away.

We also have a bad habit of changing the definition of machine intelligence when a program gets really good to claim that the problem wasn't really that hard in the first place (chess, Jeopardy, self-driving cars, etc.). This makes it seems like we aren't making any progress towards it. Admittedly, narrow machine intelligence is very different than general-purpose machine intelligence, but I still think this is a potential blindspot.

It's hard to look at the rate or improvement in the last 40 years and think that 40 years for now we're not going to be somewhere crazy. 40 years ago we had Pong. Today we have virtual reality so advanced that it's difficult to be sure if it's virtual or real, and computers that can beat humans in most games.

Though, to be fair, in the last 40 years we have made little progress on the parts of machine intelligence that seem really hard—learning, creativity, etc. Basic search with a lot of compute power has just worked better than expected.

One additional reason that progress towards SMI is difficult to quantify is that emergent behavior is always a challenge for intuition. The above common criticism of current machine intelligence—that no one has produced anything close to human creativity, and that this is somehow inextricably linked with any sort of real intelligence—causes a lot of smart people to think that SMI must be very far away.

But it's very possible that creativity and what we think of us as human intelligence are just an emergent property of a small number of algorithms operating with a lot of compute power (In fact, many respected neocortex

researchers believe there is effectively one algorithm for all intelligence. I distinctly remember my undergrad advisor saying the reason he was excited about machine intelligence again was that brain research made it seem possible there was only one algorithm computer scientists had to figure out.)

Because we don't understand how human intelligence works in any meaningful way, it's difficult to make strong statements about how close or far away from emulating it we really are. We could be completely off track, or we could be one algorithm away.

Human brains don't look all that different from chimp brains, and yet somehow produce wildly different capabilities. We decry current machine intelligence as cheap tricks, but perhaps our own intelligence is just the emergent combination of a bunch of cheap tricks.

Many people seem to believe that SMI would be very dangerous if it were developed, but think that it's either never going to happen or definitely very far off. This is sloppy, dangerous thinking.

[1] I prefer calling it "machine intelligence" and not "artificial intelligence" because artificial seems to imply it's not real or not very good. When it gets developed, there will be nothing artificial about it.