Machine intelligence, part 2

This is part two of a a two-part post—the first part is here.

THE NEED FOR REGULATION

Although there has been a lot of discussion about the dangers of machine intelligence recently, there hasn't been much discussion about what we should try to do to mitigate the threat.

Part of the reason is that many people are almost proud of how strongly they believe that the algorithms in their neurons will never be replicated in silicon, and so they don't believe it's a potential threat. Another part of it is that figuring out what to do about it is just very hard, and the more one thinks about it the less possible it seems. And another part is that superhuman machine intelligence (SMI) is probably still decades away [1], and we have very pressing problems now.

But we will face this threat at some point, and we have a lot of work to do before it gets here. So here is a suggestion.

The US government, and all other governments, should regulate the development of SMI. In an ideal world, regulation would slow down the bad guys and speed up the good guys—it seems like what happens with the first SMI to be developed will be very important.

Although my general belief is that technology is often over-regulated, I think some regulation is a good thing, and I'd hate to live in a world with no regulation at all. And I think it's definitely a good thing when the survival of humanity is in question. (Incidentally, there is precedent for classification of privately-developed knowledge when it carries mass risk to human life.

SILEX is perhaps the best-known example.)

To state the obvious, one of the biggest challenges is that the US has broken all trust with the tech community over the past couple of years. We'd need a new agency to do this.

I am sure that Internet commentators will say that everything I'm about to propose is not nearly specific enough, which is definitely true. I mean for this to be the beginning of a conversation, not the end of one.

The first serious dangers from SMI are likely to involve humans and SMI working together. Regulation should address both the case of malevolent humans intentionally misusing machine intelligence to, for example, wreak havoc on worldwide financial markets or air traffic control systems, and the "accident" case of SMI being developed and then acting unpredictably.

Specifically, regulation should:

1) Provide a framework to observe progress. This should happen in two ways. The first is looking for places in the world where it seems like a group is either being aided by significant machine intelligence or training such an intelligence in some way.

The second is observing companies working on SMI development. The companies shouldn't have to disclose how they're doing what they're doing (though when governments gets serious about SMI they are likely to out-resource any private company), but periodically showing regulators their current capabilities seems like a smart idea.

2) Given how disastrous a bug could be, require development safeguards to reduce the risk of the accident case. For example, beyond a certain checkpoint, we could require development happen only on airgapped

computers, require that self-improving software require human intervention to move forward on each iteration, require that certain parts of the software be subject to third-party code reviews, etc. I'm not very optimistic than any of this will work for anything except accidental errors—humans will always be the weak link in the strategy (see the AI-in-a-box thought experiments). But it at least feels worth trying.

Being able to do this—if it is possible at all—will require a huge amount of technical research and development that we should start intensive work on now. This work is almost entirely separate from the work that's happening today to get piecemeal machine intelligence to work.

To state the obvious but important point, it's important to write the regulations in such a way that they provide protection while producing minimal drag on innovation (though there will be some unavoidable cost).

3) Require that the first SMI developed have as part of its operating rules that a) it can't cause any direct or indirect harm to humanity (i.e. Asimov's zeroeth law), b) it should detect other SMI being developed but take no action beyond detection, c) other than required for part b, have no effect on the world.

We currently don't know how to implement any of this, so here too, we need significant technical research and development that we should start now.

4) Provide lots of funding for R+D for groups that comply with all of this, especially for groups doing safety research.

5) Provide a longer-term framework for how we figure out a safe and happy future for coexisting with SMI—the most optimistic version seems like some version of "the human/machine merge". We don't have to figure this out today.

Regulation would have an effect on SMI development via financing—most venture firms and large technology companies don't want to break major laws. Most venture-backed startups and large companies would presumably comply with the regulations.

Although it's possible that a lone wolf in a garage will be the one to figure SMI out, it seems more likely that it will be a group of very smart people with a lot of resources. It also seems likely, at least given the current work I'm aware of, it will involve US companies in some way (though, as I said above, I think every government in the world should enact similar regulations).

Some people worry that regulation will slow down progress in the US and ensure that SMI gets developed somewhere else first. I don't think a little bit of regulation is likely to overcome the huge head start and density of talent that US companies currently have.

There is an obvious upside case to SMI —it could solve a lot of the serious problems facing humanity—but in my opinion it is not the default case. The other big upside case is that machine intelligence could help us figure out how to upload ourselves, and we could live forever in computers. Or maybe in some way, we can make SMI be a descendent of humanity.

Generally, the arc of technology has been about reducing randomness and increasing our control over the world. At some point in the next century, we are going to have the most randomness ever injected into the system.

In politics, we usually fight over small differences. These differences pale in comparison to the difference between humans and aliens, which is what SMI will effectively be like. We should be able to come together and figure out a regulatory strategy quickly.

Thanks to Dario Amodei (especially Dario), Paul Buchheit, Matt Bush, Patrick

Collison, Holden Karnofsky, Luke Muehlhauser, and Geoff Ralston for reading drafts of this and the previous post.

[1] If you want to try to guess when, the two things I'd think about are computational power and algorithmic development. For the former, assume there are about 100 billion neurons and 100 trillion synapses in a human brain, and the average neuron fires 5 times per second, and then think about how long it will take on the current computing trajectory to get a machine with enough memory and flops to simulate that.

For the algorithms, neural networks and reinforcement learning have both performed better than I've expected for input and output respectively (e.g. captioning photos depicting complex scenes, beating humans at video games the software has never seen before with just the ability to look at the screen and access to the controls). I am always surprised how unimpressed most people seem with these results. Unsupervised learning has been a weaker point, and this is probably a critical part of replicating human intelligence. But many researchers I've spoken to are optimistic about current work, and I have no reason to believe this is outside the scope of a Turing machine.