Talk at the Chalmers initiative seminar

Autonomous transport systems – the good, the bad and the unknown

October 27, 2016

Digitalization and Dragons

Olle Häggström

http://www.math.chalmers.se/~olleh/
http://haggstrom.blogspot.se/
## Here Be Dragons: table of contents

1. Science for good and science for bad  
2. Our planet and its biosphere  
3. Engineering better humans?  
4. Computer revolution  
5. Going nano  
6. What is science?  
7. The fallacious Doomsday Argument  
8. Doomsday nevertheless?  
9. Space colonization and Fermi’s paradox  
10. What do we want and what should we do?
Synopsis: Future advances in science and technology have enormous potential to bring humanity great benefits, but can also bring great harm, including extinction of humanity. We desperately need a better map of possible future advances, so as to be able to act with more foresight, and reap the benefits while avoiding the risks.
The task becomes especially difficult as the areas with the greatest positive potential and those with the greatest risk often tend to coincide. One such area is artificial intelligence (AI).

**On one hand**, a McKinsey report a couple of years ago estimated an added value of $50’000 billion due to advances in AI and robotics in the next decade. Beyond that, the possibilities are virtually unbounded (other than by the laws of physics).

**On the other hand**, there are (at least) the following three reasons to be concerned about...
...the following three reasons to be concerned about what machine intelligence may bring about:

1. Can autonomous drones and other military AI technology become dangerous, for instance by falling into the hands of terrorists?

2. Robosourcing – will robots outcompete us on the labor market?

3. When AI has reached the level where we humans are no longer the most intelligent beings on Earth, can we expect to remain in control?
1. AI technology for autonomous weapons
Opposition to autonomous warfare swells to 16,000 signatories

Artificial intelligence community comes together in unprecedented numbers to call for a ban on AI-controlled weaponry
From the letter: "If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity."
Quoting from memory, here is computer scientist Patrick Doherty, in a talk in August 2015:

"I haven’t signed the letter, because there are no good or evil technologies, there are only good or evil uses of them."
2. Robosourcing
Machines taking over the jobs of humans is of course nothing new: it has been going on throughout history, and especially since the industrial revolution. Up to now, new jobs for humans have been created at about the same rate as they disappear (modulo the booms and recessions of the economic cycles). Will that still be the case, under the following new circumstances?

- It used to be only about manual labor. Now automation is increasingly affecting jobs with a more intellectual content.
- Automation is happening faster than ever. Will human labor be able to adapt sufficiently fast?
Different professions suffer different levels of risk for being rationalized away. Here are a few examples where the risk analysis turns out rather differently:

- bus drivers
- journalists
- university professors
- prostitutes
In the somewhat longer perspective, we should ask ourselves: do we really want to keep working? Would it not be liberating if we all lost our jobs to machines?

The issue of whether, in the long run, we can and want to abolish human labor has the following components:

- **Utopia.** How can a society be organized in such a way that do not work (but instead focus on art, love, sports and other enjoyable things) but still find life meaningful?

- **Getting from here to there.** How do we get from here to Utopia without causing unprecedented amounts of social exclusion along the way?
3. The ultimate AI risk – robot revolution?
“Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all. [...] It’s tempting to dismiss the notion of highly intelligent machines as mere science fiction. But this would be a mistake, and potentially our worst mistake in history.”

Stephen Hawking
Stuart Russell
Max Tegmark
Frank Wilczek
i The Independent, maj 2014
<table>
<thead>
<tr>
<th>Myth: Superintelligence by 2100 is inevitable</th>
<th>Fact: It may happen in decades, centuries or never. AI experts disagree &amp; we simply don’t know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myth: Superintelligence by 2100 is impossible</td>
<td>Fact: Many top AI researchers are concerned.</td>
</tr>
<tr>
<td>Myth: Only Luddites worry about AI</td>
<td>Actual worry: AI turning competent, with goals misaligned with ours</td>
</tr>
<tr>
<td>Mythical worry: AI turning evil</td>
<td>Actual worry: AI turning competent, with goals misaligned with ours</td>
</tr>
<tr>
<td>Mythical worry: AI turning conscious</td>
<td>Actual worry: AI turning competent, with goals misaligned with ours</td>
</tr>
<tr>
<td>Myth: Robots are the main concern</td>
<td>Actual worry: Misaligned intelligence is the main concern; it needs no body, only an internet connection</td>
</tr>
<tr>
<td>Myth: AI can’t control humans</td>
<td>Fact: Intelligence enables control: we control tigers by being smarter</td>
</tr>
<tr>
<td>Myth: Machines can’t have goals</td>
<td>Fact: A heat-seeking missile has a goal</td>
</tr>
<tr>
<td>Mythical worry: Superintelligence is just years away</td>
<td>Actual worry: It’s at least decades away, but it may take that long to make it safe</td>
</tr>
</tbody>
</table>

Olle Häggström

Digitalization and Dragons
I give the penultimate word to Nassim Nicholas Taleb, who writes, in his 2007 book *The Black Swan*:

“If you are a Stone Age historical thinker called on to predict the future in a comprehensive report for your chief tribal planner, you must project the invention of the wheel or you will miss pretty much all of the action. Now, if you can prophesy the invention of the wheel, you already know what a wheel looks like, and thus you already know how to build a wheel, so you are already on your way.”
Finally: Yes, Taleb has a point. Predicting the future is generally difficult, and predicting future technological advances even more so. Given the amount of value at stake, we should nevertheless try to make as sensible predictions as we can (while maintaining an appropriate level of epistemic humility), and then to act with foresight.