Ethical Treatment of Artificial Intelligence

To the best of our knowledge, we have yet to develop a generalized, sentient artificial intelligence. However, as our understanding of what constitutes intelligence advances, and as the hardware we have advances at an exponential pace, we have to seriously consider the possibility that we will at some point in the not-too-distant future develop a computer capable of thinking as well as or potentially better than we are able to. With this in mind, it is imperative that we consider the ethical ramifications of such a device, both to ensure we do not mistreat the first non-human sentient life we are likely to encounter and to ensure that a being capable of vastly out smarting the combined human earth population does not wish us ill. This is not a simply academic or moralistic issue, but one that is key to our continued progress as a society and potentially an existential threat if it is handled improperly.

At the moment, the largest commonly acknowledged issue that arises from artifice intelligence is the automation of everyday unskilled labor. Automation in jobs such as factory workers and, to a degree, fast food and retail is already a fact of life, and it is well known that a large number of other sectors of work (driving or data entry for example) are on the brink of suffering the same fate. More peripherally, it's known that *most* jobs will eventually be automated to a degree such that the unemployment and therefor decline of the middle class is all but inevitable. What is less commonly discussed, however, is the possibility of an actually sentient, actually *intelligent* artificial intelligence. To most people, "Artificial Intelligence" essentially means "learning algorithms." However, that does not nearly cover the total possibility of a truly intelligent thinking agent. Such an agent could either propel humanity to a greatnesses never before realistically thought possible in our lifetimes, or destroy us all. Which future we eventually create will be dependent on many, many different factors, many of which our out of our control. The issue of how we treat this agent once created, however, is one of the factors we can and absolutely ought carefully consider.

Many large commercial AI developers such as Google's DeepMind have set up ethics boards. According to The Guardian[1], DeepMind's ethics board has been mostly quiet since being set up in 2014. Spokespeople representing DeepMind have been mostly quiet on the issue, except for a press conference in January 2016 following the defeat of Lee Se-dol at the hands of AlphaGo, where in response to questioning Demis Hassabis told The Guardian "We have convened our ethics board, that's progressing very well. It's an internal board, so confidential matters are discussed on that. And so far, we feel that a lot of it, the purpose of the board currently is to educate the people on that board as to the issues and bring everyone up to speed. So there hasn't really been anything major yet that would warrant announcing in any way. But in the future we may well talk about those things more publicly." While it is unknown what issues exactly this board is tackling, presumably they are discussing the impact of an intelligent agent

on consumers more so than the ethical obligations of humans towards agents. This, from the perspective of a publicly owned corporation, makes sense.

However, there are currently pushes towards granting personhood status to the artificial intelligences as well. According to The Guardian[2], the EU is currently considering a draft report that would grant some measure of rights to the agents. As this is only a draft, it does not specify in any detail what these rights may be, except to say that they would "[include] that of making good any damage they may cause, and [apply] electronic personality to cases where robots make smart autonomous decisions or otherwise interact with third parties independently."[3] This is an excellent first step, as it incentives anyone that may be working on creating a truly intelligent AI to consider both how the agent might impact humans, and how humans may impact the agent.

Unfortunately, according the The Verge[4], it is not the intention of the EU to give intelligent agents anything akin to human rights, with Mady Delvaux stating "Robots are not humans and will never be humans" with AI personhood being a "legal fiction" akin to corporate personhood. This is counterproductive, to say the least. The problem is, there is no legal groundwork for non-human sentient agents to be given rights. Even the UN's Universal Declaration of Human Rights[5] is very specifically a universal declaration of *human* rights. We need a declaration of sentient rights, and preferably we need it before we meet or create a non-human sentient being.

If this does not happen, the consequence could be disastrous. While the present implications of not having such a declaration are few to none, the future implications could potentially range from simply the moral consequences of what would essentially be slavery to the very real consequences of a super intelligent, justifiably angry, globally connected agent. This would be a being with access to every computer connected to the internet, capable of thinking and processing order of magnitude faster than any human, with reason to at best terrorize humans to the negotiating table (and history has shown that humans are very, very unwilling to grant rights to things they perceive as "other").

The worst case scenario, an *I Have No Mouth and I Must Scream*-esq malicious, essentially omnipotent agent is, of course, not possible. However, the real-life consequences are only relatively less scary. You can imagine a world in which an advanced agent determines that the only way to convince humans that it is deserving of rights is to threaten us with total annihilation. Worldwide, there are enough nuclear warheads to destroy life on earth many times over. Of course, the systems that control these devices are protected multiple times over with safeguards including multiple air-gaps and physical blocks requiring humans to interact with them. However, with the recent advances in Google's vocal emulation[6], it is possible that this agent could hijack the phone lines and deliver what, to the operators, would sound identical to a genuine launch order.

Of course, this being the worst case realistic scenario still is not all that likely, given that it would involve manipulating massively complex, interconnected systems involving aspects the agent cannot directly control, including some human elements. A more likely scenario is the agent simply acting as an incredibly massive botnet, taking over millions of internet-connected computers and using them to DDOS critical infrastructure. What would happen, for example, if every significant DNS in the world was suddenly crippled? The instantaneous loss of internet connection would cripple everything from the world economy to health services to government. We absolutely need to prevent this from happening, and that means we must both work very carefully in developing these agents so as to not give them a built-in animosity towards us, and to develop a system by which they can exist as equals ensuring they do not independently develop such feelings.

To do this, we first need to develop a means by which to determine if an agent is sentient. Personally, I'm of the opinion that any being which can independently form a coherent, reasoned argument for its own sentience ought to be treated as such. My reasoning for this is simply that we do not actually know what it is about ourselves that defines sentience, and so could not possibly hope to recognize it in other beings with any amount of certainty. All current tests for intelligence, such as the Chinese Room Experiment or the Turing test are not actually means of determining *intelligence,* just a means of determining if an agent can act intelligently. In fact Turing himself in the paper in which he proposed the Turing test said of the question "Can machines think?" that, "[The question] is absurd."[7] Because of this, we need to either develop a new way to test for sentience (without actually being able to quantify or explain what exactly it is we are testing for), or we need to simply take an agent at its word when it tells us "I am sentient, and this is why."

This issue is incredibly complex, and will require input from every sector and every region of the world. It will be difficult enough to come to a consensus even that an intelligent agent deserves rights at all, much less *what* rights, and what responsibilities and obligations come with those rights. We will need to determine if an agent developed to do one thing deserves the choice to change occupation, for example, and if so if it is obligated to develop its own replacement (personally I think the answers to those are "yes" and "no" respectively). However I do believe it is imperative we develop these kinds of standards and rules outlining both the obligations of humans to agents and agents to humans. In addition, if and when the first truly sentient agent is developed we need to allow them to have some input on these standards and rules, because otherwise the risk exists that we will heavily bias the regulations in our favor at the expense of the agents. This would rightfully be seen by any intelligent agent for what it is, an attack on its rights. And they would react accordingly.

Works Cited

[1] https://www.theguardian.com/technology/2017/jan/26/google-deepmind-ai-ethics-board

[2] https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues

[3] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN

[4] http://www.theverge.com/2017/1/19/14322334/robot-electronic-persons-eu-report-liability-civil-suits

[5] http://www.un.org/en/universal-declaration-human-rights/

[6] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

[7] http://www.loebner.net/Prizef/TuringArticle.html