CS461 Homework 6 **25/12/2019**
Adahan Yalçınkaya 21502369
Bengi Dönmez 21602237
Emre Sülün 21502214
Eray Şahin 21502758
Kazım Ayberk Tecimer 21502531

Yoshua Bengio

In the book Architects of Intelligence in Martin Ford's interview with Yoshua Bengio, Bengio says that "My understanding of the current science as it is now, and as I can foresee it, is that those kinds of scenarios [Existential threat from super intelligent AI and getting into a recursive improvement loop] are not realistic" [1]. In this essay, we will argue that existential threat from super intelligent AI and getting into a recursive improvement loop is a realistic scenario and current developments in AI world is making us closer to these catastrophes.

As the technology advances, computers had a monumental augmentation in their computing and processing power. This leads to advancement in AI technology. Today, we use AI extensively in our lives. But, there are negative examples and scenarios about the risks [2]. We are in the opinion that the most important risk is the potential of becoming super intelligence. Although it is manmade, it may improve and renew itself without human control and beyond the understanding capability of humans. "People now control the planet, not because we're the strongest, fastest or biggest, but because we're the smartest. If we're no longer the smartest, are we assured to remain in control?" says Tegmark, the President of the Future of Life Institute [3]. While we are trying to survive, to make easy our life and pursuing power, we changed our environment by ignoring and even harming other living and inanimate beings. There's no guarantee AI won't do the same to us.

There are mainly two types of AI namely specialized and general (AGI). The difference between them is that "specialized AI is created to do one thing, AGI is created to learn to do anything"[4]. Today specialized AI is widely used for specific tasks such as Siri, facial recognition, Snapchat filters etc. On the other hand, "AGI would be able to learn, plan, reason, communicate in natural language, and integrate all of these skills to apply to any task" [4]. Recursive self-improvement peculiarity of AGI grants a system to make improvements and renewing. This system leads AI to improve itself in every cycle via feedback, in the end resulting AI to become Super Intelligence. it is predicted to happen within a few decades.

Knight from MIT Technology Review says that "We've never before built machines that operate in ways their creators don't understand [5]. As an example, in the early 2000s AI

was used to make smart toys but now Nvidia Corporation created a self-driving car. Unlike its predecessors Nvidia's car learned driving while watching others, instead of following instructions given by the engineer [5]. This shows that how AI evolves faster than we think. In the last decade we applied AI to anything healthcare, finance, smart homes, security, chatbots and even social media. AI is used everywhere in our life and it will affect every part of our lives when it gets out of control.

We cannot control how AI evolves even now. Most of the cases, AI uses machine learning algorithms and deep learning. For AI to learn it needs to be fed with data. When it is fed data on the internet without any control over it, it can result with unwanted results. For example, let's look at the case of TAY (Thinking About You) Microsoft's chatbot released on Twitter on March 23, 2016 [6]. After few hours that it had been released, TAY started tweeting many racist and inappropriate phrases. Microsoft had to shut it down 16 hours after its release [6]. This is a very important example that we have no control over AI's learning on open internet. Another example of this is an OpenAI's language modeling algorithm, named GPT-2. GPT-2 is trained with 8 million web pages and it can generate texts that cannot be differentiated from real humans writings. This brings along lots of problems with it like fake news. Therefore, OpenAi does not share whole algorithm with the public [7]. This is a temporary solution to prevent GPT-2 to become like TAY bot. They feed GPT-2 with controlled data. It looks like a successful achievement for AI world even though GPT-2 hasn't receiving data from the outside, it is in its safe haven. Someday in the future, GPT-2 should contact the open internet, when that day comes there is no reason GPT-2 not to become like TAY bot.

To conclude, AI is a rapidly growing branch of the computer science and it is intertwined with our lives. It also has a potential risk. Current developments in AI may lead to catastrophic outcomes for humanity. As a people of science, we must acknowledge these risks and take precautions.

References

[1] M. Ford, Architects of Intelligence, Birmingham UK, Packt Publishing Ltd., 2018.

[2] A.Turchin and D. Denkenberger, "Classification of global catastrophic risks connected with artificial intelligence" AI & Soc 2018, p.2. [Online]. Available: https://www.researchgate.net/publication/324935393_Classification_of_global_catastrophic_risks_connected_with_artificial_intelligence [Accessed Dec. 22, 2019]

[3] M. Tegmark "Benefits and Risks of Artificial Intelligence". [Online]. Available: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/ [Accessed Dec.23,2019]

[4] M. Bullock "Artificial General Intelligence in plain English", 10/10/2019. [Online]. Available: from https://towardsdatascience.com/artificial-general-intelligence-in-plain-english-

e8f6e9a56555 [Accessed Dec. 23, 2019]

[5]W. Knight “The Dark Secret at the Heart of AI” MIT Technology Review, 11, April 2017. [Online]. Available:https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai [Accessed Dec. 23, 2019]

[6] V. Mathur, Y. Stavrakas and S. Singh, “Intelligence analysis of Tay Twitter bot” 2nd International Conference on Contemporary Computing and Informatics(IC3I 2016), Noida,India(IC3I), December 14-17,2016, Institute of Electrical and Electronics Engineers Inc., 2016 pp.231-236. [Online]. Available: https://www.researchgate.net/publication/316727714_Intelligence_analysis_of_Tay_Twitter_bot [Accessed Dec. 23, 2019]

[7] A. Radford, J. Wu, D. Amodei, D. Amodei, J. ClarkMiles, B. Sutskever “Better Language Models and Their Implications” 14, February 2019 [Online]. Available: https://openai.com/blog/better-language-models/ [Accessed Dec 23 2019]

Word count: 777 (without references)