Will AI Destroy Humankind?

Stone Fang (Student ID: 19049045)

1 Introduction

Nowadays AI is under fast development and changing our world. Despite people are enjoying the benefits that AI brings, a few scientists and professionals are concerned about the potential threat from AI in the future(Wikipedia contributors, 2019c)(Bostrom, 2014). In fact, the question of "Will AI destroy humankind?" has been widely discussed and researched for a long time from serious academic scholars to fiction writers(Wikipedia contributors, 2019a). While some optimists may hold the belief that AI will benefit under absolute control of humans, by contrast, I think it is possible for AI to damage or even extinguish human race if we are incautious. In this essay, I will propose and analyze three conditions of AI takeover and what actions can be taken to prevent them.

If AI can destroy humankind, three conditions must be satisfied:

1. AI has the ability
2. AI has the chance
3. AI has the motivation

The following sections will analyze these three conditions in depth.

2 Ability

Obviously, AI must have the ability to destroy humankind before it does so. If AI can't, such destruction will never happen regardless of other two conditions.

Then it is natural to ask the next question: does AI already have such ability? At first glance it is easily to give an yes because we can build robots that can wound or kill a human or make use of existing robots which was not created for such purpose to do so. For example, robots can be equipped with weapons to form an AI army force, or an autonomous driving vehicle might be misused to cause a serious crash. This kind of answer seems to be straightforward and believable, but there are still something different if we think deeper.

A robot or autonomous vehicle can kill humans not because of superior in intelligence but rather a more powerful engine. On the other hand, machines do surpass humans on intelligence in some tasks such as calculation, playing chess or go, or even face recognition under some circumstances. However, such superior does not lead to harm to a human up to the present.

From the analysis above, we can see that AI agents like robots do not necessarily have to be highly intelligent in order to kill humans. Instead, the physical or mechanical aspect is more important.

If it runs faster, has stronger arms, or is equipped with powerful weapons, it does have the ability to harm or kill humans.

3 Chance

The second condition is more important at present. AI must have the chance to destroy humankind. If AI only have the ability to destroy humankind but cannot find a chance to exert such ability, the destruction will not happen.

3.1 Lost Control

The primary necessary condition of such chances is that AI is lost control by its creators. If every action or consequence can be accurately predicted and controlled by human, AI agents can never do harm to human unless they are programmed to do so.

3.2 Passive vs Active

There are two possible alternative reasons for an AI agent out of humans' supervision and monitoring. One is that it is complex enough so that it is impossible to predict and control every aspect of its behavior. In this case, AI agents do not hide or escape from humans' control on purpose. It is not because of AI's super intelligence but humans' limit in intelligence of controlling high complex objects. If this could happen, it implies that humans can create something complex enough and ultimately out of their control. In other words, the ability to control is less than the ability to create complexity. But really can we do that?

Alternatively, another reason is that AI agents have the ability to cheat humans. In this case, since humans are not likely to deliberately create something to deceive themselves, AI must be able to self-improve and ultimately emerge the capability of disguising and cheating. This implies humans should create AI agents that can learn something they are not supposed to learn, which is a high intelligent status beyond our current capability. All existing AI algorithms, are **not really intelligent** – they can do only what they are designed to do. On the other hand, a human can learn on arbitrary tasks, and even more, an infant can spontaneously start to learn without being told or designed what to learn.

3.3 Growth and Reproducibility

Are growth and reproducibility necessary for AI destroying humankind? Suppose an AI agent is out of control but can neither expand its zone of influence nor reproduce. It is possible for such AI agent to destroy humankind by coincidence. For example, if there is an AI program which controls the launching of nuclear weapons all around the world, it only requires one millisecond of lost control to destroy humankind.

But this case is not likely to be true because we will probably not build an AI program controlling all nuclear weapons. Therefore, AI agents must be able to grow and reproduce otherwise the bugs of lost control can be fixed.

4 Motivation

4.1 Learn of Motivation

Finally, AI must have the motivation to destroy humankind. This is the most tricky one among the three conditions. Before any AI agent evolves the motivation to destroy humankind, it must be able to create motivation or purpose. Currently an AI agent may have its purpose but it is programmed by human and cannot be changed without human's interference. Humans have motivations and purposes but we don't know the reason. This is a paradox – we know we are a species of purposefulness but we don't know why or how to create something purposeful. We need fundamental breakthrough in computing theory and cognitive science if we want AI algorithms that can "learn" new motivations.

4.2 Motivation of Destroying Humankind

However, if an AI agent can learn for any purpose and even create new motivations, does it necessarily evolve to have the motivation to destroy humankind? The answer is no and Asimov's "Three Laws of Robotics"(Wikipedia contributors, 2019b) is an answer to this question. If we can figure out that we are living in a world following the Three Laws, we are assured to be safe from AI. If we can find a way to create robots following the Three Laws, we can make sure the destruction by AI will never happen. But can we really do that? Can we create some form of AI that can create any motivation on its own, with only the motivation of destroying humankind being fortunately, coincidentally and trickly excluded?

5 Conclusion

To conclude, this essay proposed three conditions of the destroying of humankind by AI, analyzed whether the conditions have already been satisfied or will be satisfied in the future, and suggested how to prevent the destruction.

1. AI has the ability to destroy humankind. Though such ability may comes from power and strength instead of intelligence, from the perspective of reality, the fact is that we can create powerful robots that can kill humans.
2. AI has the chance to destroy humankind, which means AI is lost of humans' control, either passively or actively. Though theoretically AI may destroy humankind by coincidence, it is more likely to happen gradually, which implies AI have to be able to grow and reproduce.
3. AI has the motivation to destroy humankind. To achieve this, AI algorithms should be able to learn new motivations first, and develop the motivation of destroying human next.

Reference

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (1st ed.). New York, NY, USA: Oxford University Press, Inc.

Wikipedia contributors. (2019a). *AI takeover — Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=AI_takeover&oldid=914300390

Wikipedia contributors. (2019b). *Laws of robotics — Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Laws_of_robotics&oldid=911737896

Wikipedia contributors. (2019c). *Open letter on artificial intelligence — Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Open_Letter_on_Artificial_Intelligence&oldid=911194556