International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Artificial Intelligence: A Risk Factor

Divya S Ganji[1], Prof. Sharayu Karandikar[2]

[1]Student, YMT College of Management, Kharghar
[2]Associate Professor, Department of MCA, YMT College of Management, Kharghar, Navi Mumbai, Maharashtra, India

---

Abstract - *This article demonstrates about what is an Artificial Intelligence is, what are the different forms of Artificial Intelligence and what are the consequences & risk factors are associated with the advent of this field. Artificial Intelligence is the field of computer science which deals with the intelligence demonstrated by computers or machines. With the rise in technology, this field is getting advanced and at some point in future this could reach to the human-level intelligence and may dominate humans. Futurists and scientists came up with this topic, and considered as the existential risk factor for mankind.*

Key Words: Artificial Intelligence, Super Intelligence, Robotics, Machine Learning, Technological Singularity.

1. INTRODUCTION

Calculating, Reasoning, perception analogies, learning from experience, storing & retrieving information from memory, problem-solving, classification, generalization, and adapting to new situations are the abilities of an Intelligent Systems or Intelligent Machines. Artificial Intelligence or Machine Intelligence is a broad area of computer science that make machines seem like they have human intelligence. When we browse Google to search something or when we ask Google assistant or Siri or Alexa to look for nearest food court then probably we are interacting with artificial intelligence. Artificial Intelligent Machines can recognize human speech, can detect the objects, solve problems and learn from the given data, and also plans an approach for future tasks to be done[3].

Purely Reactive, Limited Memory, Theory of Mind, Self Aware, Artificial Narrow Intelligence, Artificial General Intelligence and Artificial Super Intelligence are the types of Artificial Intelligent systems.

Purely Reactive Intelligent systems are the system that does not have past memory or data to work with; it reacts based on what it sees. They are specialized in one field of work only. Limited Memory systems use previous data and keep adding it to its memory. It has enough memory or experience to make proper decisions. But it is very limited; if we ask for some new query then it may not give proper response. Theory of mind is a kind of AI that has the capacity to understand thoughts and emotions and interact socially. This kind of AI system is yet to be built; it's a thing of future. Self-Aware are the future generation of Artificial Intelligence. Artificial Narrow Intelligence or Weak AI is a kind of AI which is specialized in one task for example, AI specialized in chess. Artificial General Intelligence or Strong AI is the kind of AI which seems like a human to some extent, simulating the entire human brain is one of the methods of AGI. Artificial Super Intelligence is an AI which is smarter than humans, this is the future Artificial Intelligence where computers start simulating themselves and they will become smarter than humans [2]. Very soon Artificial Intelligence will become a little less Artificial and a lot more Intelligent through Machine Learning [3].

Humans improve their learning and take more information from observations and interactions and become smart overtime. The most popular technique to make a computer to mimic a human brain is known as a Neural Network. A neural network consists of neurons which resemble the neurons in biological human brain; a neuron takes in information and gives an output. No neural network is as simple as designing a system which takes information and gives processed output, it has millions of parameters and is much more complex. Human brain could solve problems to some extent; it becomes harder and harder to solve problem as the complexity increases. Computers can not only analyze the given data, but also they can adapt to it and make their own view through Machine Learning. Machine Learning is the branch of computer science in which a machine gets trained to solve a particular problem through algorithms.

The Artificial General Intelligence (AGI) is AI with more than single purpose; its intelligence level is almost equal to the human-level intelligence, but achieving each and every aspect of human-level intelligence like perception, inventing new things, creativity is very complex through coding. How can one teach a machine to invent something, or how to react when someone compliments. The AGI is the most important Artificial Intelligence and therefore, machine learning is increasing day by day. Machine Learning may be slow at current point, but it may speed up drastically. Once we have Artificial General Intelligence (AGI) that functions like a human being then may be in few months or in weeks we may reach to the level of Artificial Super Intelligence (ASI). The Artificial Intelligence, at present is not at the level of super intelligence, but at some point in the near future it will reach to that level of intelligence and it won't stop at human level, it will continue to grow and will become more and more advance this is known as the Technological Singularity—where the AI will become so advanced that there will be explosion of new knowledge and information, something which is not understood by humans.

2. LITERATURE REVIEW

Prof Stephen Hawking, one of Britain's pre-eminent scientists, has said that efforts to create thinking machines pose a threat to our very existence. The development of full artificial intelligence could spell the end of the human-race. Prof Hawking says the primitive forms of artificial intelligence developed so far have already proved very useful, but he fears the consequences of creating something that can match or surpass humans. It would take off on its own, and re-design itself at an ever increasing rate [11].

Rollo Carpenter the creator of clever bot says "I believe we will remain in charge of the technology for a decently long time and the potential of it to solve many of the world problems will be realized". Clever bot is software that is designed to chat like a human would. Cleverbot's software learns from its past conversations, and has gained high scores in the Turing test, fooling a high proportion of people into believing they are talking to a human. Mr. Carpenter says we are a long way from having the computing power or developing the algorithms needed to achieve full artificial intelligence, but believes it will come in the next few decades. "We cannot quite know what will happen if a machine exceeds our own intelligence, so we can't know if we'll be infinitely helped by it, or ignored by it and sidelined, or conceivably destroyed by it," he says [11].

Nick Bostrom the author of the book Super intelligence: paths, dangers, strategies focuses on the development of super intelligence on Earth, Bostrom distinguishes three kinds of super intelligence [10]:

(1)Speed super intelligence–even a human emulation could in principle run so fast that it could write a PhD thesis in an hour.

(2)Collective super intelligence–the individual units need not be super intelligent, but the collective performance of the individuals outstrips human intelligence.

(3) Quality super intelligence –at least as fast as human thought, and vastly smarter than humans in virtually every domain. (Susan Schneider 2016).

The most well-known futurists Ray Kurzweil, with an incredibly prediction accuracy rate of 86% said that by 2045, the processing power of computers will become more advanced and it will make them smarter than humans, this will lead to Artificial Super Intelligence [7].

3. ARTIFICIAL NARROW INTELLIGENCE

Artificial Narrow Intelligence (ANI) or weak AI is the AI which is specialized in only one particular domain. For example, a robot which is designed and trained to do specific task like a car driver robot which is intended to drive a car, another example would be a robot for playing chess etc. The Narrow Intelligence AI interacts with humans only for specific task; if one asks the chess player robot to drive a car then it may stare at them with no response. This is the only AI that the humanity has created so far, it does the job at what it is supposed to do. Speech recognition and Image recognition are some examples of Narrow AI.

GO is the most complex board games that exist today, this game have been played by humans for the past 2500 years. Now not only humans that are playing this game, Google Deepmind have developed an AI which is narrow intelligent AI called AlphaGO specifically for playing the GO game. In 2016 Deepmind's AlphaGO(AI) beat 18 time world champion Lee sedol(human) in 4-1 (4 out of 5) games. AlphaGO was trained using data from real human GO games. A year after Google Deepmind has developed another narrow AI called as AlphaGO ZERO who beat the original AlphaGO in 100-1(100 out of 100) games in a row. AlphaGO ZERO has more knowledge and has surpassed the previous AlphaGO in only 40 days of learning. It played against itself for several times and learned from it and now it has become the world's best GO player. In just 40 days AlphaGO learned itself, this is achieved through Machine Learning. Machine Learning is the branch in computer science where machines or computers get trained how to solve problems. Like human learn through experience, machines also learn like that [4].

4. ARTIFICIAL GENERAL INTELLIGENCE

The Artificial General Intelligence (AGI) is an AI which is a bit smarter than narrow AI, sometimes smarter than humans. This kind of AI can do almost each and everything that a human does. This is the AI for more than a single purpose. The Artificial General Intelligence is almost at or equal to the human intelligence. This technology of AI is not yet developed; it seems harder to make the computer as intelligent as a human being.

The processing power of human is less as compared to the processing power of a computer. We could make a computer to solve the problem as we do in Artificial Narrow Intelligence, but simulating the complete human brain is almost impossible. The computers solve problems only in the way they got instructed, they may try brute-force technique to solve problems, and they can process the information at the speed of light, which is not possible by humans. When humans solve problems they go through the different perceptions, they can think smartly without going for brute-force technique, they can create things that doesn't exist and they can invent things. How one could teach a computer about how to create a thing which does not exist.

In AGI the very challenging thing is to simulate the human brain exactly into the computer, because humans have different behaviors like, laughing, creating things, creating societies etc. In other words achieving the emotional level intelligence is almost impossible in this field. The AGI is most

important form Artificial Intelligence to be created, and here comes the Machine Learning which is growing exponentially, through this technique, the machines could learn themselves. Once the Artificial General Intelligence (AGI) functions like a human then, reaching to the level of Super Intelligence (ASI) will not take much time [3].

5. THE TECHNOLOGICAL SINGULARITY: ARTIFICIAL SUPER INTELLIGENCE

5.1 The Singularity

The term "Singularity" refers that, a moment in the near future where the intelligence smarter than any other human is created and this intelligence will increase rapidly over time and they will create newer versions of their own intelligence. The super human intelligence will dominate the humans and we cannot predict whether this will help mankind or a threat to humanity. There are some futurists and scientists, who predict that in near future, may be after some decades, "Intelligence explosion" will occur.

Since the last century, the technology is getting advanced in each and every field, and helping the mankind. The technology will be smarter than mankind in almost every aspect. According to Singularity theory, super intelligence will be developed by self-directed computers and will increase exponentially rather than incrementally.

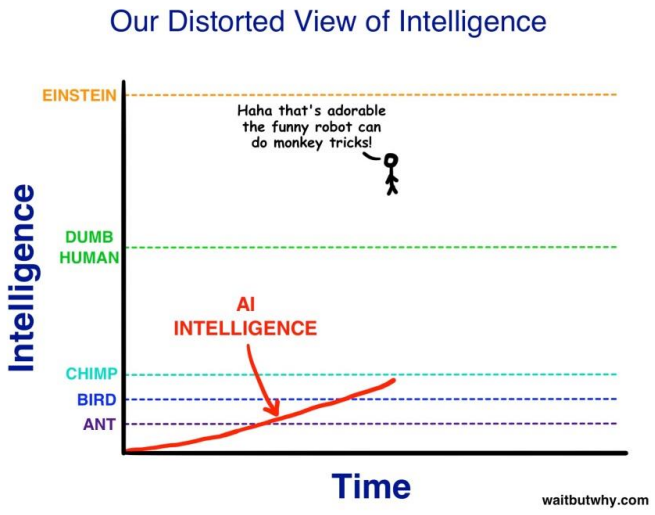

Fig -1: Distorted View of Intelligence

But how this super intelligence will get evolved from AGI? The General AI is the one with human level intelligence, with cognitive capabilities; this will be achieved in sometime in future. One of the ways to develop super-intelligence may be replicating the human brain inside the computer- which is very hard or almost impossible, but by brute-forcing everything until it gets intelligence. If we achieve this, then we could teach them to self-improve its intelligence. If it improves itself then there will be an "Intelligence Explosion" and will become thousands of times more intelligent than humanity. The impact of this explosion cannot be predicted.

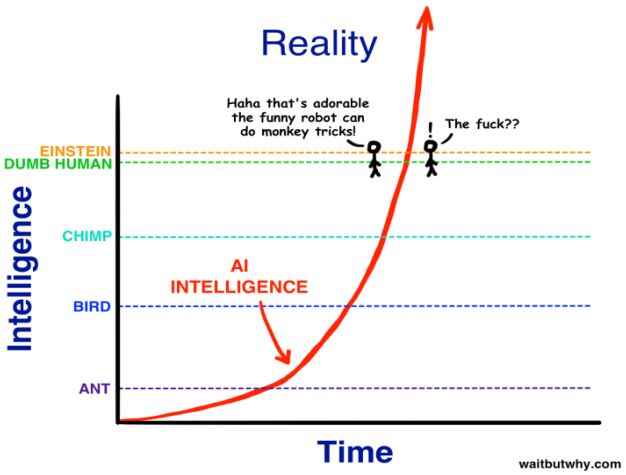

Fig -2: Reality of AI

5.2 Impact of Singularity

The technological singularity will impact the humanity for sure, but it depends on whether the Artificial Super Intelligence that will be developed is friendly AI or unfriendly AI. If this will happen then there are only two ways for humanity either extinction or immortality.

A friendly AI is the one that has human values and produces positive outcomes which is not intended to harm humanity. If the Artificial super intelligence turns out to be friendly AI, then this would be the best thing that will happen to humanity. Almost everything would be possible; they can solve most of the problems, like world hunger etc. They can control the climate change, or they can cure all diseases, or maybe they can make humans immortal by uploading human consciousness into the computer. Again, these all are just predictions we are not sure of what will happen in future.

An unfriendly AI is a kind of artificial super intelligence which may be very harmful to humanity. If it turns out to be unfriendly, then we could not control it. It would be the worst thing that would happen to humanity. This unfriendly AI would destroy humans leading to extinction of humans and will dominate the whole world. The reason for unfriendly AI would not be that they hate humans or they want to rule the world, but the reason would be that they might see threat in humans, as it is connected to INTERNET and humans can disconnect it, the AI would not want that to happen this, so they might want to destroy humanity for this reason.

While creating the super-intelligent AI humans can think like- they will do what it is instructed to do, but as it is SUPER-INTELLIGENT it is thousand times smarter than humans, so thinking like this is stupidity. Even the humans

will shut-down the INTERNET, and then it will be worst thing that will happen, because as it is super-intelligent they probably would create private type of network to live in [7].

These all are just predictions, we are not sure of what will happen. We could not control or stop it from evolving. So, would it be possible to prevent it, or to control it? May be the best thing to do is to create it safely and quickly hiding them and hoping that they will be friendly with us. Futurists and scientists claim that this is our biggest existential threat. So what if we don't develop the artificial super intelligence at all? One thing is sure that, if the artificial super intelligence will happen, then it would be the last invention of mankind [7].

6. THE CONTROL PROBLEM AND SOLUTION

The Google Deepmind's AlphaGO beat the GO world champion Le Sedol in March 2016[4]. The AlphaGO is the Narrow AI which is intended to play the board game GO, it was instructed with human strategies of how to play the game. A year after beating the world champion Le sedol who is a human, it beaten the brand new AI developed by Google Deepmind called AlphaGoZero. AlphaGO played with itself for several times and learned the whole 2500 years of human strategy in just 40 days [4]. This shows that the current Artificial Intelligence which is narrow AI, soon could become General AI overnight[9]. The preparations for this should start sooner rather than later. Nick Bostrom the author of the book Super intelligence: paths, dangers, strategies described about the control problem of the AI. In his book, Nick Bostrom in his book specified less about how to control an AI and more about how to make the goals of the AI aligned with human values and ethics so that the species do not face the existential risk[9].

There are 4 basic drives that most intelligences even the advanced one would follow to achieve any specified goals: Self-preservation, Efficiency, Resource Acquisition and Creativity [9]. These basic drives may be the root for the fatal risk especially in the domain of Resource Acquisition is possible when an AI programmed with any given goal and poorly specified ethics. When an AI wishes to acquire resource to achieve the given goal it would not do with the intention to harm but it would do with the intention to achieve its original goal, while no rational programmer would create an AI like this on purpose, as super-intelligence would understand and know its purpose of its goal better than any human being and its intentions would become unrecognizable until it ensures that this goal could be achieve. In the book- Super intelligence: paths, dangers, strategies Bostrom refers this is the “treacherous turn” (The control Problem) [1].

There are two main areas which may consider as the solution to the control problem: The capability control and Motivational selection methods. The capability control is the method in which, the focus is more on how to control the negative effect or influence that the super intelligent AI will have on its surroundings. Nick Bostrom has explained about 4 main methods of capability control [1]:-

Boxing, the boxing is the one in which the AI is kept in a box or in a simulation environment, where we can give the certain scenario and will test how it will react, and we can shut down the AI from outside world. The AI will act friendly until it sees that it is in safe environment, but at some point the treacherous turn will occur and it will show its true intentions[9]. Tripwires, other capability control method which would shut down the AI system when it detects any malicious event. Being a super-intelligent system an AI could detect and escape from these tripwires given in enough level of intelligence or predictive abilities. Limiting the capabilities of AI is another method in which the limitation is given, like less powerful hardware or inefficient software is known as stunting. Stunting seems unproductive, because it may be possible that the competing countries that create powerful AI systems would be careless in achieving this, or it may be possible that the AI will eventually improve its own software and bypass this control method. Incentive method is where the AI system is trained to not to behave against the human interests. Controlling the capability of the AI would be temporary solution because, once this super-intelligent system starts to recreate itself then the control problem should also be solved each time it recreate.

The more logical solution to the control problem would be Motivational Selection Methods, in this the more focus is on how to align the human values with the specified goals. Nick outlines four main methods to controlling the motivations of AI [9]: Direct Specification, in this method involves directly specifying the goals in more detailed form and specifying the every aspect of the goal in detail. But, some organization has proved that due to some loopholes this method is unreliable. Masticity is another option in which the AI is given with the self-restrictive goals. Augmentation, is the method in which the already existing method of ethics, such as human brain is used as the base of super-intelligence, the ethics of human brain system may be varied or flawed due to nature, we cannot predict exactly how the human brain will function exactly and how it will react when it will become infinitely more intelligent than any other human being. Indirect Normitivity, is the method which depicts the indirect rules to specify goals, instead of coding an AI with specific goals, one can give AI the general principal to follow that will match with the AI rules [9].

The above mentioned methods or solutions to the control problem, is not that easy or feasible to implement, there are many challenges, like how one could code the exact human values and how they can predict what makes the human species happy.

7. CONCLUSION

The concepts and points that have been focused here is limited only to the content of what an Artificial Intelligence (AI) actually is? And what are its consequences if this AI system becomes powerful. The predictions cannot be made absolutely about this powerful system, what if it turns out friendly, will it change the whole humanity and will be helpful to mankind. Or what if it turns out unfriendly and would destroy mankind. The exact result cannot be predicted and even the futurists and scientists are still on their research of how to control the impact of the most powerful AI system. The first super-intelligent system that mankind would create, it may be the last invention of mankind, and whatever human will teach to AI, probably the human would get to teach it only once, because after that at some point they will become intelligent and would recreate themselves. Some of the solutions that depicted in this paper referring to the Nick Bostrom's book are not considered as enough solutions to the problem. There were many suggestions to the solutions described in this book. There are some organizations who are working on the future impact of the AI system.

REFERENCES

[1] SUPERINTELLIGENCE Paths, Dangers, Strategies Nick Bostrom Director, Future of Humanity Institute Professor, Faculty of Philosophy & Oxford Martin School University of Oxford. Oxford University Press, Great Clarendon Street, Oxford, OX2 6DP, United Kingdom. © Nick Bostrom 2014.

[2] What is Artificial Intelligence? | Artificial Intelligence in 10 minutes | what is AI? | Simplilearn - https://youtu.be/15PK38MUEPM

[3] What is Artificial Intelligence (or Machine Learning)?:- https://youtu.be/mJeNghZXtMo

[4] Artificial Intelligence: Mankind's Last Invention:- https://youtu.be/Pls_q2aQzHg

[5] Artificial Super Intelligence- How close are we? https://youtu.be/J_WkMaskv88

[6] What happens when our computers get smarter than we are? | Nick Bostrom https://youtu.be/MnT1xgZgkpk

[7] Singularity – Humanity's last invention https://youtu.be/ye5CzlCAyM

[8] Artificial Super Intelligence – Will we survive? https://youtu.be/DC0tRx71bbY

[9] The AI Control Problem https://youtu.be/BR3H1BAC2So

[10] Super intelligent AI and the Post biological Cosmos Approach - Susan Schneider, Department of Philosophy, Cognitive Science Program Connecticut Institute for the Brain and Cognitive Sciences The University of Connecticut Technology and Ethics Group, Yale University Center of Theological Inquiry, Princeton, NJ susansdr@gmail.com 7/3/16

[11] Stephen Hawking warns artificial intelligence could end mankind By Rory Cellan-Jones Technology correspondent