Environment. Technology. Resources. Rezekne, Latvia
Proceedings of the 12th International Scientific and Practical Conference. Volume II, 180-183

Safety of Artificial Superintelligence

Aleksejs Zorins
Faculty of Engineering
Rezekne Academy of Technologies
Rezekne, Latvia
Aleksejs.Zorins@rta.lv

Peter Grabusts
Faculty of Engineering
Rezekne Academy of Technologies
Rezekne, Latvia
Peteris.Grabusts@rta.lv

***Abstract*—The paper analyses an important problem of cyber security from human safety perspective which is usually described as data and/or computer safety itself without mentioning the human. There are numerous scientific predictions of creation of artificial superintelligence, which could arise in the near future. That is why the strong necessity for protection of such a system from causing any farm arises. This paper reviews approaches and methods already presented for solving this problem in a single article, analyses its results and provides future research directions.**

***Keywords*—Artificial Superintelligence, Cyber Security, Singularity, Safety of Artificial Intelligence.**

I. Introduction

The paper presents an overview of safety concepts and technologies introduced in the recent years in the area of artificial intelligence (AI) safety. There are different predictions but many AI researchers, philosophers and futurologists agree that in the next 20 to 200 years a machine capable to perform on at least human level on all tasks will be developed [3,4,7, 10, 14]. According to assumption that such a machine will be capable to design the next generation of even smarter intelligent machines, an intelligence explosion will take place very shortly. Despite different predictions regarding this issue, from possible economic problems till complete extinction of a mankind, many researchers agree that this problem is extremely important and urgently needs serious attention [1,2,5,8,19],

This problem is becoming more and more important in the scientific community due to several reasons: rapid development of AI both on software and hardware levels; wide implementation of AI in industry and other fields of human activity; the lack of control of super-AI. The research in this field is in a very beginning and at the same time is crucial to our development and safety.

The superintelligence problem is usually connected with a so-called singularity paradox, which could be described as follows: "superintelligent machines are feared to be too dumb to possess common sense" [10].

The machines have completely different discreet logic and structure, do not, and could not have emotions and feelings like humans. Even if the computer will decide to make a man happy it will do it with the fastest and cheapest (in terms of computational resources) without using a common sense (for example killing of all people will lead to the situation that no one is unhappy or the decision to treat human with drugs will also make him happy etc.). Another issue is that we want computer to that we want, but due to bugs in the code computer will do what the code says, and in the case of superintelligent system, this may be a disaster.

The next sections of the paper will show the possible solutions of making superintelligence safe to humanity.

II. Possible solutions for safe AI

In his research of Artificial Super Intelligence Roman Yampolskiy presents a comprehensive review of potential solution methods of singularity and safe AI problem [18]. All possible solutions are grouped into several categories according to a way of dealing with the problem:

- ✓ Prevention of development - the researchers of this group completely denies AI and consider only one method for dealing with the problem – prohibit the development of AI;
- ✓ Restricted development – here the idea is to restrict superintelligent machines with different approaches: software, hardware and mixed ones;
- ✓ Incorporation into society - the authors are sure that intelligent machines should be a part of human society (in economic, legal, religious, ethical, moral and / or educational aspects) and this will help to successfully solve the problem;
- ✓ Implementation of self-monitoring algorithms of AI – create a set of rules to follow, development of human-friendly AI, include emotions into AI behaviour algorithms etc.;
- ✓ Other solutions – join AI, deny of the problem, relax and do nothing and some other proposals.

Prevention of development of AI is the most aggressive and straightforward method and probably the most effective one. However, taking into account the modern society and the level of inclusion of computers in

Print ISSN 1691-5402
Online ISSN 2256-070X
http://dx.doi.org/10.17770/etr2019vol2.4042

our life it is very unlikely that laws of banning computers will be accepted in a near future. Even if some of the worlds governments will incorporate such a law into its legislation there will always be countries or even several individuals who violates regulations.

Restricted development is the most commonly accepted solution to the safe AI problem. AI-boxes, leakproofing and restricted question-answering-only systems (Oracle AI) are among the proposed solutions in this group [3,4,16,17,20]. The methods presented in this category are similar to putting a dangerous human being into prison – it does not give a 100% safety but in most cases can help society to survive for some period of time. It is clear that this solution cannot help in the long-term but could be a good initial measure when the real superintelligence will be developed.

Incorporation into society is a very easy to implement. We can include into computer algorithms social, moral, ethical and other rules, but an intelligent machine with a digital mind will easy and fast discover the drawbacks, contradictions and disadvantages of our legislation, moral and ethical rules. A very interesting idea of raising AI like a child also belongs to this group [11], however, the grow-up children often are very different from their parents expectations despite all efforts.

The self-monitoring algorithms category includes explicitly hard-coded rules of behaviour into computer and creation of multilevel guard composed of clever machines to monitor each other. The set of rules is a good solution but it cannot cover every possible situation and if such a situation occurs, the computer may act in an unpredicted manner. The computer watch guard will lead to a hugely increased system which could not be handled by human and giving all right to AI sooner or later will have very bad effects.

The final category includes extreme approaches of battling against machines or doing nothing because a clever AI will defeat us in any way. Another opinion held by several researchers and businessmen including the owner of Tesla Elon Musk is that at the moment the only feasible solution to this problem is joining the researchers of AI (for example OpenAI project with the mission "discovering and enacting the path to safe artificial general intelligence" [13]) to be aware of technological advances in this field and to be able to react quickly depending on the situation.

As one may see, some of these proposals are completely unacceptable and will not be analysed in the paper while some needs to be described in more details. The next section will discuss some of the most interesting and realistic methods.

III. Engineering Safe Superintelligence

David Chalmers firstly introduced the main idea of this approach in 2010, who suggested that for safety reasons AI systems should be restricted to simulated virtual worlds until their behaviour will be fully understood under controlled conditions [19].

Roman Yampolskiy in his book "Artificial Superintelligence" proposes a formalized notion of AI confinement protocol; for restricting an AI entity to a closed environment from which it cannot exchange information with the outside environment if such information is not authorized by the confinement authority [18].

The idea of this protocol is to ask system questions with restricted answers. These safe questions are defined as ones for which human being could find an answer without any help from superintelligent artificial mind. For example, the researcher has two alternative approaches to try and knows that one of them is definitely correct, but doesn't know which one exactly. If the choice is wrong then he will lose time but will find a solution after some additional time. In this case, AI advice could be used.

Those types of questions are safe for several reasons. The main reason is the answer may be obtained even without assistance from AI and the second reason is that the answer could be encoded in a single bit making AI hard to hide additional information in it. Safe questions should be developed by the group of specially trained experts in different fields like computer science (especially in recent developments of AI), psychology, ethics, AI security and some others.

IV. AI confinement problem

The idea of confinement roots back from 1973 when Butler Lampson has defined the laws of confinement [12]:

- ✓ Total isolation: a confined program shall make no calls on any other program.
- ✓ Transitivity: if a confined program calls another program that is not trusted, this program should be also confined.
- ✓ Masking: a program to be confined must allow its caller to determine all its inputs into legitimate and covert channels.
- ✓ Enforcement: the supervisor must ensure that a confined program's input to covert channels conforms to the caller's specifications.

The confinement problem has many questions to be answered. For example it will be very difficult to interact with such a system (or even impossible) or that clever machine eventually will escape its prison using different approaches which a human guard does not know. The researcher Eliezer Yudkowsky has made so-called AI-Box experiment where he was trying to demonstrate that even human-level intelligence is enough to escape from an AI-Box [13].

There are several possibilities to prevent superintelligence from escaping. The first thing proposed by Yampolkiy is to isolate humans from confined AI completely (no contacts via voice, Morse code, text messages or any other channels) and delegate all maintenance operations to unintelligent software that has no possibilities to release AI from the box.

The second measure to be taken is to allow AI

functioning on limited and randomly changed resources without any access to its configuration. The AI should be reset regularly where the previous versions should replace current one. Using a virtual machine could also help.

The last group of measures is actual design of confinement facility. It should have self-destruction function; it should be located in isolated, hard-to-reach place guarded to prevent unauthorized access.

V. Explainability of Intelligent Systems

Very important aspect of safe AI is explainability, when a user should clearly and fully understand the output of an AI system and make corrections if necessary.

The concept of explainable AI (XAI) is shown if Fig. 1, which clearly shows that today the user usually does not know the answers to the following questions:

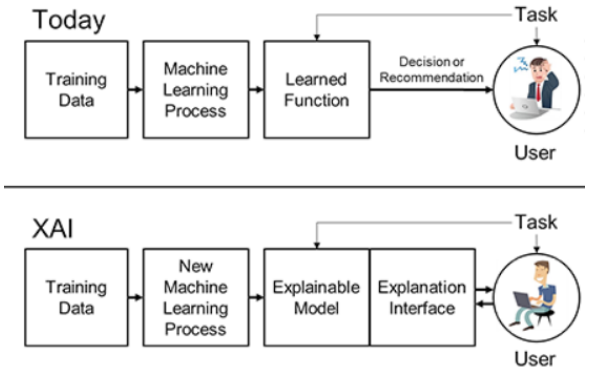

Fig. 1. Explainable AI [6].

why computer do it; why not something else; when it succeeded; when to trust a computer and how to correct errors. In the case of XAI the user will have answers to all of these questions.

In general, The Explainable AI (XAI) program aims to create a suite of machine learning techniques that [6]:

- ✓ Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy);
- ✓ Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

Andres Holzinger in his paper gives an approach of a complete machine learning pipeline beyond algorithm development [9], the authors will show only relevant to safe AI issues:

- ✓ Data: pre-processing, integration, mapping and fusion – understanding the physical aspects of raw data and its surroundings, especially i the application domain; ensuring quality of data.
- ✓ Learning algorithms: all aspects of design, development, experimentation, testing and evaluation.
- ✓ Visualization of data and analysis: presentation of multidimensional data in a human-friendly form.
- ✓ Privacy: data protection, safety and security.
- ✓ Entropy: used as a measure of uncertainty in data.

Wojciech Samek gives provides several reasons why explainability is so important for AI research and its safety aspects [15]:

- ✓ Verification of the system: no one should trust artificial system by default. In this case, verification procedure allows testing the AI "black box" behaviour and outputs using different solutions already available.
- ✓ Improvement of the system: Before improving the system, we should understand its weaknesses and the better we do it the better we can improve them.
- ✓ Learning from the system: nowadays AI systems are often using the big data of millions examples which human mind cannot deal with. That is why explainable AI should have extracted knowledge in a human understandable manner.
- ✓ Compliance to legislation: if a system gives wrong answer in a critical data domain, the responsibility should be preserved according to legislation issues. The users affected by such AI decision will want to know why the system decided that way. Only explainable system will give the answer.

There are several methods for making AI explainable: sensitivity analysis (SA), Layer-Wise relevance propagation (LRP) and others.

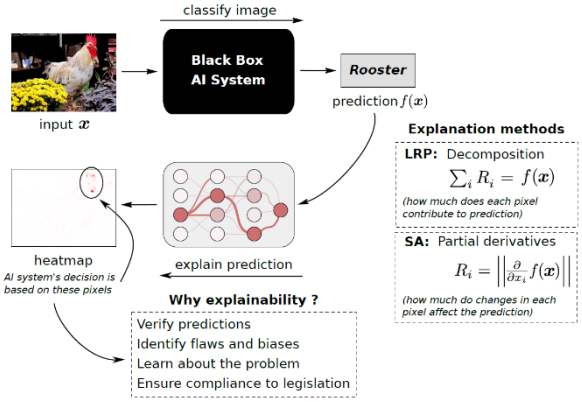

Fig. 2. Explanation of AI system prediction [15]

The fig.2 shows that system has correctly classified an input as a "rooster". In order to understand this decision, explanation methods such as SA or LRP are applied. The result of this explanation is an image, which visualizes the importance of each pixel for the prediction. In this example, the rooster's red comb and wattle are the basis for the AI system's decision. Using such a methodology user can be sure that system works as it intended to be.

VI. Future Research Directions

The paper shows that the research made in the direction of ensuring safety of artificial superintelligence is in its early stages. All the methods are limited and do not assure the complete confidence of AI user that this technology

will have only positive effects.

All possible solutions to ensure safe AI should be taken into account and carefully analysed and any small chance of improvement should be tested and implemented in a prototype.

Safe AI is the area of several interconnected science directions like computer science, psychology, mathematics, physics, philosophy, linguistics, biology and many more. That is why only a team of cross-trained researchers could deal with such a problem and expect to have some positive results.

The author of this article wants to attract attention of researchers from all fields to this topic, because the creation of superintelligent machine is only a matter of time and we should be ready and know what to do.

The further research directions will be development of experimental testing framework for different AI systems to check some of the above-mentioned methods of ensuring the safety of artificial superintelligence.

References

[1] R. Banham, Cybersecurity: Protective Measures Treasuries Should Be Taking. Treasury & Risk. 2018 Special Report, pp. 2-7.

[2] D.Beskow, K.Carley, Social Cybersecurity: An Emerging National Security Requirement. Military Review. April 2019, Vol. 99 Issue 2, pp. 117-127. 11p.

[3] N. Bostrom, Global Catastrophic Risks. Oxford: Oxford University Press, 2007.

[4] N. Bostrom. The ethics of artificial intelligence. Cambridge Handbook of Artificial Intelligence, 2011. [Online]. Available: https://nickbostrom.com/ethics/artificial-intelligence.pdf [Accessed: March. 03, 2019].

[5] R. Deibert, Toward a Human-Centric Approach to Cybersecurity. Ethics & International Affairs. Winter 2018, Vol. 32 Issue 4, pp. 411-424.

[6] D. Gunning, Explainable Artificial Intelligence, DARPA project, 2018. [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence [Accessed: March. 07, 2019].

[7] S. Hawking, Science in the next millenium, 1998. [Online]. Available:https://www.learnoutloud.com/Catalog/Science/Physics/Science-in-the-Next-Millennium/45223 [Accessed: March. 03, 2019].

[8] N.Hennig, Privacy and Security Online: Best Practices for Cybersecurity. Library Technology Reports. April 2018, Vol. 54 Issue 3, pp. 1-37.

[9] A. Holzinger, "From Machine Learning to Explainable AI." World Symposium on Digital Intelligence for Systems and Machines August 2018. [Online]. Available: https://www.researchgate.net/publication/328309811_From_Machine_Learning_to_Explainable_AI [Accessed: Feb. 21, 2019].

[10] M. Kiss, C. Muha, the cybersecurity capability aspects of smart government and industry 4.0 programmes. Interdisciplinary Description of Complex Systems. 2018, Vol. 16 Issue 3-A, pp. 313-319.

[11] R. Kurzweil, The Singularity Is Near: When Humans Transcend Biology. New York, NY: Viking, 2006.

[12] B. Lampson, A Note on the Confinement Problem, 1973. [Online]. Available: https://www.cs.utexas.edu/~shmat/courses/cs380s_fall09/lampson73.pdf [Accessed: March. 19, 2019]

[13] Open AI project. [Online]. Available: https://openai.com/ [Accessed: March. 11, 2019].

[14] N. Sales, Privatizing Cybersecurity. UCLA Law Review. April 2018, Vol. 65 Issue 3, pp. 620-688. 69p.

[15] W. Samek, T. Wegang, K. Muller. Explainable artificial intelligence: understanding, Visualizing and interpreting deep learning models, Aug. 28, 2017. [Online]. Available: https://arxiv.org/abs/1708.08296 [Accessed: March. 19, 2019].

[16] M.Scala, A. Reilly, Risk and the Five Hard Problems of Cybersecurity. Risk Analysis: An Official Publication Of The Society For Risk Analysis, March 2019, pp. 32-37.

[17] A.T.Sherman, D. DeLatte, M.Neary etc., Cybersecurity: Exploring core concepts through six scenarios. Cryptologia. July 2018, Vol. 42 Issue 4, p337-377. 41p.

[18] R. Yampolskiy, "Leakproofing the Singularity Artificial Intelligence Confinement Problem," Journal of Consciousness Studies, vol. 19, pp. 1-2, 2012.

[19] R. Yampolskiy, Artificial Superintelligence: a Futuristic Approach. New York: Chapman and Hall/CRC, 2015.

[20] E. Yudkowsky, The AI-Box experiment. [Online]. Available: http://yudkowsky.net/singularity/aibox/ [Accessed: March. 19, 2019].