*Information Technology and Management Science*
ISSN 2255-9094 (online)
2020, vol. 23, pp. 64–68
https://doi.org/10.7250/itms-2020-0009
*https://itms-journals.rtu.lv*

1862
RIGA TECHNICAL UNIVERSITY

# Developing Ecological Safety of Artificial Intelligence in Human Society

Aleksejs Zorins[1], Peter Grabusts[2]
[1,2] *Rezekne Academy of Technologies, Rezekne, Latvia*

***Abstract* – The paper presents cyber systems especially based on artificial intelligence (AI) from a perspective of ecological safety for humanity. The study provides a definition of ecological safety of AI and discusses its relevance to a modern science and society, as well as reviews risks of smart AI systems.**

***Keywords* – Artificial intelligence, cybersecurity, risks of artificial intelligence, safe artificial intelligence.**

## I. Introduction

The main difference among AI systems is a goal or task they can solve. Most modern AI systems are designed to solve narrow tasks with many limitations or assumptions: translation systems, face recognition, medical diagnostic systems, autopilots etc. Some of such neural network systems try to imitate some of human brain functionality aspects [1], [2].

The majority of AI researchers agree that in the future (from 20 till 100 years ahead) a computer will have intellectual abilities like a human or beyond [3], [4], [5], [6], [7]. In this case, a computer will be able to create easily as many self-replications as it will consider necessary. It is also possible that these replicas will be much smarter than their creator. In this case, rapid exponential growth of artificial intelligence will be inevitable.

The AI safety problem becomes more and more important due to the rapid development of computer hardware and software; its broad implementation in many spheres of human activity such as finance, medicine, education, entertainment etc.; the lack of understanding and control of smart computer system development and usage; the number of in-code bugs and errors in the final release of intelligent systems. The research in this area is still at its early stage and, at the same time, is critically important to development and safety of humankind. By ecological safety of artificial intelligence, the authors imply such an artificial system that is safe, clean, friendly and sustainable to both humanity and nature.

Intelligent computers are totally different from human logic and structure; they do not have and, in the authors' opinion, will never have emotions and feelings like humans. Even if a smart computer for some algorithmic reason will decide to make humans happy, it will use its digital sense and logic that may be implemented in a very strange and cruel way (for instance, a decision to reduce the size of humanity will lead to better living conditions or treating people with drugs will also make them happy etc.). The errors in a source code of intelligent software may lead to catastrophic results [8].

We may define artificial superintelligence (sometimes called general AI) as AI systems that are able to outperform humans on most real-world tasks. The next chapter of this paper will present some attempts of making artificial intelligence safe to humanity.

## II. Attempts of Making AI Safer

Yampolsky presents a comprehensive review of potential solution methods of making artificial intelligence safer in his papers [9], [10]. These solutions are divided into several categories with respect to a way they are dealing with this problem:

- ✓ Prevention of development – this solution considers complete destruction of AI systems because humans are not able to fully control them, especially in the case of superintelligence. The main method for dealing with the problem – complete ban of the development of AI.
- ✓ Restricted development – the researchers of this group offer different ways of AI development restriction on different levels: software, hardware and both.
- ✓ Incorporation into society – "adepts" of this idea want to fully incorporate AI machines into society and give them full access to all areas: economics, legislation, religion, ethics, education etc.) and are sure that this will help get rid of the problem.
- ✓ Implementation of self-monitoring algorithms of AI – creation of rules to follow, development of human-friendly AI, including emotions into AI behaviour algorithms etc.
- ✓ Other solutions – some authors propose joining AI in different ways, including biotechnology, denial of the problem and some other approaches.

Restricted development is one of the most popular ones: AI-boxes, leak proofing and restricted question-answering-only systems (Oracle AI) are the main representatives in this group [3], [4], [11], [12], [8]. The methods of this category are similar of putting a dangerous human being into a prison – it will not give total safety but in most cases could help society survive for some period of time. The above-mentioned solution is not

suitable in the long term but could be a good starting point before the real superintelligence is created.

The category of self-monitoring algorithms presents inclusion of algorithmically coded rules into computer behaviour and creation of multilevel security, including smart machines that will monitor each other. The set of behaviour rules may be a good solution; however, it is impossible to predict and describe all possible events and if such an event occurs, this computerized guard may act in an unpredicted manner [8].

David Chalmers introduced his idea of safety AI engineering in 2010, which included a suggestion that for safety reasons AI systems should be restricted to simulated virtual worlds until their behaviour would be totally and clearly understood by humans [10].

The book of R. Yampolsky "Artificial Superintelligence" presents a clear definition of AI confinement protocol; which implies a restriction of an intelligent machine to a closed environment from which it cannot exchange any information with the world outside if such information is not approved by the confinement authority [9].

The main principle of this approach is asking a system questions with restricted answers. Such questions could be answered by humans without help of artificial intelligence and the answers may be compared later.

"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" [8].

Another approach is to put AI in a prison, both digital and physical one and monitor its activity from a safe distance. The idea of such a prison dates back to 1973 when Butler Lampson defined the laws of confinement [13]:

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

This isolation approach has several tasks to be solved. One of such tasks is interaction with such a system or that intelligent program may escape its prison by using methods and tactics, which a human guard does not know. Eliezer Yudkowsky has made the so-called AI-Box experiment (often criticised by many other researchers, however) where he was demonstrating that in some situations using even human-level intelligence allowed escaping from an AI-Box [14].

## III. Goal-Based Approach to AI Safety

Legg and Hutter give a formal definition of intelligence based on algorithmic information theory [15]. According to their definition, "intelligence measures an agent's ability to achieve goals in a wide range of environments". This definition is broad enough to use it for description of human and artificial intelligence.

The core thing is a goal, which can fully describe human or system behaviour, and that is why it is so important to define the goals of AI systems very precisely and clearly at the earliest stage of its development. Such an issue could prevent many possible negative outcomes during implementation and use.

Most present-day AI systems are mainly focused on quite narrow goals; however, a trend towards greater generality can be observed. The more intelligent an artificial system is, the more control it will have over the key factors of the environment according to its goals. When two artificial agents have conflicting goals in the same environment, then typically the more intelligent agent will have success and the less intelligent will fail. If the goals of the artificial system are not aligned with ours, then we should consider situations when the goal of AI will be achieved, and the goals of humanity will be ignored [16].

Apart from a specific problem or task-based goals, each clever AI system might have the so-called instrumental goals [17]:

- ✓ Self-improvement: using adaptation algorithms the clever artificial agent becomes better in achieving its goals.
- ✓ Goal-preservation and self-preservation: this principle ensures that future versions of AI will be programmed to achieve the same goals, thus leading to achievement of the final wished result.
- ✓ Resource acquisition: the more resources AI could collect, the faster it may get to its end goals.

The main idea behind this theory is to develop a goal-based framework, which could help in developing human safe AI systems. The main issues are moral characteristics of persons in charge for such projects.

## IV. Explainability of AI

Explainability is a very important issue of safe AI when a user should clearly understand the output of a computer system and make corrections, if necessary.

The Explainable AI (XAI) software goal is to create such machine learning techniques that [18]:

- ✓ 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 describes a complete development process of a machine learning algorithm, considering explainability issues [19]:

- ✓ Data: pre-processing, implementation, and usage – understanding the structure and features of initial data and its sources and ensuring quality of data.
- ✓ Learning algorithms: including all stages of design, development, testing, implementation, and evaluation.
- ✓ Visualization of data and analysis: possibility of presenting multidimensional data in a user-convenient visual form.
- ✓ Privacy: data protection and security issues.
- ✓ Entropy: used to measure and describe the level of uncertainty in data.

Wojciech Samek describes several reasons why explainability is crucially important for making AI safe both in design and implementation [20]:

- ✓ Verification of the system: nobody should trust artificial system by default. Verification procedure allows testing artificial intelligence "black box" behaviour and outputs using different methods.
- ✓ Improvement of the system: it should be based on proper analysis of system strong and weak points before making improvements; the better this process is, the better results we obtain with the improved version of AI.
- ✓ Learning from the system: today intelligent systems use big data consisting of billions of examples, which human intellect cannot deal with. Therefore, explainable AI should have extracted knowledge in a manner fully trusted and understandable for a human.
- ✓ Legislation issues: if a computer system makes a mistake in critical data, the responsibility should be ensured according to legislation. The answer why the system behaved this way should always be received and only explainable system can help in this situation.

Sensitivity analysis (SA), Layer-Wise relevance propagation (LRP) and other methods make AI more explainable [17].

## V. Risks of Intelligent AI Systems

The risk management of AI systems should be the key aspect of its development and implementation phase. AI should be 100 % safe and error free; of course, this is a very hard to achieve. However, if the scientific and business society ignores safety even in a smallest possible fraction, the results of such a behaviour may lead to serious negative consequences, i.e., to a point of no return.

These risks are raised due to such factors as [21]:

- ✓ Source data. Ingesting, sorting, linking, and properly using data have become more and more difficult due to the increased amount of unstructured data being received from different and sometimes contrary sources, such as the internet, social media, smartphones, sensors, and the Internet of Things. Sometimes it is easier to use or reveal sensitive information from anonymised data. Another issue is intentionally or unintentionally presented faulty and erroneous data, which should never be used for training of intelligent machines.
- ✓ Technology issues. Technology and process factors on all levels of development and implementation can have a negative impact on performance of AI systems.
- ✓ Security considerations. Another factor is the potential to exploit seemingly non-sensitive marketing, health, and financial data that companies collect to use in AI systems. If security precautions are insufficient, it is possible to merge these threads together creating false identities.
- ✓ Model misbehaving. Artificial intelligent systems themselves can run into problems when they output biased results, become unstable, or give conclusions for which there are no true reasons. Consider, for example, the potential for artificial systems to discriminate unintentionally against protected classes and other groups by using together a zip code and income data to create targeted offerings.
- ✓ Interaction considerations. The user interface between people and computers is also a key risk area. Among the most obvious are problems in manufacturing, automated transportation, and infrastructure systems. Accidents and injuries can happen if operators of vehicles or other equipment do not recognise when systems should be overruled (e.g., self-driving cars). Human judgment can also prove faulty in overriding system results.

There are a number of other risks such as lapses in data management, scripting errors, and misjudgement in model-training data, which can compromise fairness, privacy, security, and compliance. Besides, we do not consider intentional efforts to compromise AI structure and use it in criminal applications.

## VI. Ecological Safety of AI

At the beginning of computer and AI era, scientists were trying to create computers like human brains. There are thousands of publications on this topic and hundreds of applications able to solve complicated tasks. Still no real progress has been achieved using this paradigm. Computers were solving very narrow and specific tasks not even close to a general ability of human brain.

In the recent time, scientists have stopped idealising human brain as an ideal computational mechanism and the best model for clever AI creation. Homo sapiens ideal has been doubted. One of the most famous experiments in neuroscience was conducted by Benjamin Libet in 1983 when he demonstrated that our sense of free will might be an illusion, a controversy that had only increased ever since [22]. Stanford state prison experiment performed in 1971 by a group of psychologists led by Philip Zimbardo showed that behaviour of people was largely defined by a role ("guard" and "prisoner") they were put into [18]. Elizabeth Loftus proved that a human under certain

circumstances could easily form false memories and be sure that these false memories were completely true [23].

Computers have reached a certain threshold in performance when very simple calculations allow getting extremely impressive results using such technologies as Deep Learning, Big Data etc. Google was attempting to put a human-like acting robot into their self-driving car Waymo and succeeded only when this robot "was thrown out" and let a car system incorporate all information from its sensors and cloud and act accordingly. Similar stories of other companies may be found in [24].

The Google Translate Service started to show progress when the company took all UN official translations in all languages and uploaded them in their system. The system only finds connections between words and phrases in different languages from a database and provides a result. Now Neural Machine Translation technology has reached near human translation quality [25].

All these examples and many others show that AI has outperformed humans in many fields and has become our manager. AI does not need to be as smart as human; it only incorporates all necessary information and decides. YouTube and Google Search give us relevant search results adapting to our browsing history and we do not even think that there could be some other information which these or other systems simply do not show due to some programmed limitations or rules, or simply mistakes in the source code. The society is becoming a servant of computer systems designed by only few people in the world. Is it wise to trust this small group of people that cannot be checked?

AI development team should consider and accomplish several goals to ensure near 100 % safety of intelligent system defined in [22]:

- ✓ ensuring that AI project conforms with ethical principles taking into consideration the impact it may have on affected users, stakeholders, and environment;
- ✓ ensuring that AI project is non-discriminatory and fair assuming its potential to have discriminatory effects on humans, by reducing biases that may corrupt system outputs, and by being aware of the factors surrounding fairness that come into account in each phase of software life cycle;
- ✓ ensuring that AI project is worthy of public trust by providing guarantee to society that the product is safe, accurate, reliable, secure, and robust;
- ✓ ensuring that AI project provides possibility of correction by prioritizing both the transparency of design and testing process, and the transparency and interpretability of its decisions and behaviours.

The FAST Track Principles could help in achieving these goals. Implementing this concept by all members of AI project delivery team ensures better support of a responsible environment for data innovation.

Principles of fairness, accountability, sustainability, and transparency at every level of project development stage are crucial and require joint attention and active involvement of experts possessing technical skills, domain knowledge, project/product management skills, and policy competence. Ethical AI development and implementation procedure is a team effort on all levels.

The main FAST Track Principles are as follows:

*Fairness*. All intelligent systems that compute social or demographic data must be designed to have a near-zero possible negative impact on the persons involved:

- ✓ the datasets should be equitable;
- ✓ the system architectures should only include reasonable features, processes and analytical structures;
- ✓ they should not have inequitable impact;
- ✓ they should be implemented in a correct and unbiased way.

*Accountability*. All intelligent systems must be developed to facilitate end-to-end answerability and auditability. This process requires responsibility from humans involved in the design and implementation process as well as the use of activity monitoring protocols that ensures review and explanation at all stages.

*Sustainability*. Programmers, software engineers and users of intelligent systems must be aware of the fact that these technologies have a transformative effect on individual humans and society. They should always act with constant sensitivity to response from the real world. They also should keep in mind that the technical sustainability of artificial systems totally depends on their safety.

*Transparency*. AI system development team should be able to:

- ✓ explain to its users in an understandable language how and why a system performed the way it did;
- ✓ justify the ethical permissibility, ensure that the system is not discriminatory and could be trusted by public regarding its outputs and performance mechanisms and principles.

All in all, the AI system development often lacks the above-mentioned recommendations and principles, which is why the efforts to expand and introduce this knowledge are extremely important in the world that mostly depends on computers driven by AI applications.

The authors of the present study define ecological safety of AI as a framework of ethical, moral, technical, legislation and other rules, techniques and procedures that will help humans control and use AI systems without any harm to their mental and physical health [26].

This framework should strictly and clearly define:

- ✓ Global analysis with case studies of any possible harm of AI systems to a human (especially children), both in mental and physical aspects.
- ✓ Improvement of cyber legislation by a team of trusted IT and law professionals with the highest responsibility level and full public transparency.

- ✓ Ethical and moral code for AI system developers with the highest possible level of control of its implementation (including special AI development control committees).
- ✓ Improvement of modern education system in a way that allows children to obtain early education (at least till 14 years) without any help of computer technologies.

It is necessary to adapt the first four principles in IT education courses and study programmes on all levels.

As the first part of future research, the authors consider to perform a global analysis with case studies of any possible harm of AI systems to humans. The results of it should be presented to wider audience to make a real impact and further progress in understanding the ecological safety of AI. Only if this problem is well explained and understood, we can expect real positive changes in the situation within the AI sector, which at this moment is quite pessimistic.

## VII. Conclusion

The authors understand that this ecological framework needs to be discussed, improved, developed and implemented as soon as possible and will need joint efforts at all society levels. There are different attempts in this area, but there is no unity among specialists who are dealing with this problem.

The main problem is motivation. The main goal of the present study has been to motivate scientific community to push this process faster and start acting before it is too late. The authors really hope that this idea will raise interest and awareness and will be fully developed and accepted.

## References

[1] S. Schliebs, N. Kasabov, "Evolving spiking neural networks: A Survey", 2013. [Online] Available: https://www.zora.uzh.ch/id/eprint/75356/1/Schliebs_Kasabov_Evolving_spiking_neural_networks.pdf [Accessed: May. 9, 2020].

[2] A. T. Sherman, *et al*. "Cybersecurity: Exploring core concepts through six scenarios," *Cryptologia*, July 2018, vol. 42, no. 4, pp. 337–377. https://doi.org/10.1080/01611194.2017.1362063

[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: Feb 22, 2020].

[5] S. Hawking, *Science in the next millennium*, 1998. [Online]. Available: https://www.learnoutloud.com/Catalog/Science/Physics/Science-in-the-Next-Millennium/45223 [Accessed: Feb 19, 2020].

[6] M. Kiss, and C. Muha, "The cybersecurity capability aspects of smart government and industry 4.0 programmes," *Interdisciplinary Description of Complex Systems*, vol. 16, no. 3-A, pp. 313–319, 2018. https://doi.org/10.7906/indecs.16.3.2

[7] N. Sales, "Privatizing Cybersecurity," *UCLA Law Review*, April 2018, vol. 65, no. 3, pp. 620–688, 2018.

[8] E. Yudkowsky, The AI-Box experiment. [Online]. Available: http://yudkowsky.net/singularity/aibox/ [Accessed: Feb. 9, 2020].

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

[10] R. V. Yampolskiy, *Artificial Superintelligence: A Futuristic Approach*. Chapman and Hall/CRC, 2015. https://doi.org/10.1201/b18612

[11] M. Scala, and 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, 2019.

[12] A. Tavanaei, *et al*. Deep Learning in Spiking Neural Networks, 2019. [Online] Available: https://arxiv.org/pdf/1804.08150.pdf [Accessed: May. 9, 2020].

[13] B. Lampson, *A Note on the Confinement Problem*, 1973. [Online]. Available: https://www.cs.utexas.edu/~shmat/courses/cs380s_fall09/lampson73.pdf [Accessed: Feb. 19, 2020]

[14] Open AI project. [Online]. Available: https://openai.com/ [Accessed: Feb. 11, 2020].

[15] S. Legg, and M. Hutter, "Universal Intelligence: A definition of machine intelligence," *Minds & Machines,* vol. 17, pp. 391–444, 2007. https://doi.org/10.1007/s11023-007-9079-x

[16] S. J. Russell, "Should We Fear Supersmart Robots?" *Scientific American*, vol. 314, no. 6, pp. 58–59, 2016. https://doi.org/10.1038/scientificamerican0616-58

[17] T. Everitt, Towards Safe Artificial General Intelligence. PhD thesis, Australian National University, 2018.

[18] D. Gunning, Explainable Artificial Intelligence, DARPA project, 2018. [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence [Accessed: February 23, 2020].

[19] A. Holzinger, "From Machine Learning to Explainable AI," *World Symposium on Digital Intelligence for Systems and Machines,* August 2018. https://doi.org/10.1109/DISA.2018.8490530

[20] W. Samek, T. Wegang, and K. Muller, Explainable *artificial intelligence: understanding, Visualizing and interpreting deep learning models,* 2017. [Online]. Available: https://arxiv.org/abs/1708.08296 [Accessed: Feb. 7, 2020].

[21] B. Cheatham, K. Javanmardian, and H. Samandari, "Confronting the risks of artificial intelligence," *McKinsey Quarterly*, 2019. [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/Confronting-the-risks-of-artificial-intelligence [Accessed: Sept 22, 2020].

[22] L. Benjamin, C. A. Curtis, E. W. Wright, and D. K. Pearl, "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential) - The Unconscious Initiation of a Freely Voluntary Act," *Brain,* vol. 106, no. 3, pp. 623–642, 1983. https://doi.org/10.1093/brain/106.3.623

[23] E. F. Loftus, and J. E. Pickrell, "The formation of false memories," *Psychiatric Annals*, vol. 25, no. 12, pp. 720–725, 1995. https://doi.org/10.3928/0048-5713-19951201-07

[24] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Rev*., vol. 1, pp. 187–210, 2018. https://doi.org/10.1146/annurev-control-060117-105157

[25] A. Way, "Quality Expectations of Machine Translation," in *Translation Quality Assessment. Machine Translation: Technologies and Applications,* Moorkens J., Castilho S., Gaspari F., Doherty S. (eds), vol. 1. Springer, Cham. 2018. https://doi.org/10.1007/978-3-319-91241-7_8

[26] A. Zorins, and P. Grabusts, "Safety of Artificial Superintelligence," *Environment. Technology. Resources, Proceedings 12th International Scientific and Practical Conference*, Rezekne, Latvia, 2019. https://doi.org/10.17770/etr2019vol2.4042

**Aleksejs Zorins** received his Mg. sc. ing. degree in Information Technology from Riga Technical University in 2001. Since that time, he has working as a Lecturer at Rezekne Academy of Technologies. From 2009 to 2016, he was the Director of the study programme "Electronic Commerce" at the Department of Computer Science. Since 2015, he has been a Doctoral student at Rezekne Academy of Technologies, Faculty of Engineering. His research interests include artificial intelligence and its safety aspects.
E-Mail: Aleksejs.Zorins@rta.lv
ORCID iD: https://orcid.org/0000-0003-0929-8352

**Peter Grabusts** received his Dr. sc. ing. degree in Information Technology from Riga Technical University in 2006. Since 1996, he has been working at Rezekne Academy of Technologies. Since 2014, he has been a Professor at the Department of Computer Science. His research interests include data mining technologies, neural networks and clustering methods. His current research interests include ontologies and techniques for clustering.
E-Mail: peter@rta.lv
ORCID iD: https://orcid.org/0000-0002-9627-9901