11TH INTERNATIONAL SCIENTIFIC CONFERENCE
ON DEFENSIVE TECHNOLOGIES
OTEH 2024
Serbia, Tara, 9-11 October 2024

RELIABILITY OF ARTIFICIAL INTELLIGENCE

SLAVKO POKORNI
Information Technology School, Belgrade, slavko.pokorni@its.edu.rs

Abstract: *The goal of this paper is to show the importance of reliability for artificial intelligence (AI), which is in nowadays applying in almost every area of human life. Everything can fail, and AI is not an exception, as any other product or devices. AI also need to be reliable. The definitions of AI, reliability and some reasons for unreliability such as quality of data, small or insufficient number of training and test data, training on unreal data, nature of learning algorithms and user incompetence are analysed in this paper.*

Keywords: *reliability, artificial intelligence.*

1. INTRODUCTION

Everything can fail, and artificial intelligence (AI) is not an exception. If AI is an attempt to replace human intelligence with machine intelligence, and human reasoning can sometimes fail, so AI can fail in a similar way [1]. If something can fail, we speak about reliability, and if it can be repaired, we speak about maintenance. Reliability is related to failures, and failures can cause losses from little to very big, for example economic and sometimes loss of life. So, reliability is important, and sometimes critical, depending of area where AI is applicated (airplanes, autonomous vehicles, medical devices, etc.). It implies that reliability of AI is important in order to use it with confidence and safety, for the benefit of all people, without misuse.

Artificial intelligence is mainly realized as software, but it involves hardware, data and human. So, when we speak about AI reliability, it obviously includes hardware, software, data and human reliability.

2. DEFINITION OF RELIABILITY

In general, reliability is defined as the ability of an item to perform a required (expected) function under stated conditions for a stated period of time [1]. Instead of term item, in this paper we will use term product as general term, and this will include device, element or system realized in hardware, software or both. AI can also be treated as product. Reliability of AI can be defined in similar way as reliability of a product.

In [2] Reliability of Artificial Intelligence[1] is defined as the probability, at a certain level of confidence, that the System will successfully, without failure, perform the function for which it is intended, within the specified performance limits, during the specified duration of the tasks, when it is used in the prescribed manner and for the purpose for which it is intended, under defined load levels, taking into account the previous system usage time.

Quantitatively, reliability is expressed in probability, and are very important in reducing downtime and operational and maintenance cost of a product. Reliability and maintenance are mutually connected. Higher reliability means less costly maintenance. There are many factors which can influence reliability. For example, reliability of hardware can change if environment changes (temperature, mechanical stresses, etc). Similarly, reliability of AI changes if data and user behaviour changes.

High reliability is very important, especially in professional equipment as military, medical or driverless cars, and it comprises hardware reliability, software reliability, and human reliability [3], and it can include AI, so reliability of AI is also important.

Calculating reliability, we essentially calculate prognostic reliability. That is why we are not talking about determination, but about reliability assessment or reliability prognostic. Reliability prognostic has been done for almost 60 years using MIL-HDBK-217, mostly for electronic hardware [4]. Reliability assessment of software is more complicated.

Reliability is building in during designing of a product, provided in production and supported in use. It is also connected with cost. More reliable product is more expensive, but more reliable product is also cheaper for maintenance.

Artificial intelligence is in essence a software but it involves hardware, data and human and reliability of AI can be defined as in [5].

Because AI is dominantly software, it is connected with software reliability. Software errors are dominantly cause of software unreliability. Software can contain errors, and errors can produce faults. Errors in software are produced by programmers, so it is important to be familiar with these errors.

[1] In [2] it is called System or Artificial Intelligence System.

Corresponding author: S. Pokorni;
E-mail: slavko.pokorni@its.edu.rs
DOI: 10.5937/OTEH24118P

AI can learn from human, and human knowledge can be erroneous, so it is connected with human reliability.

AI can learn from data, and data can be accidentally or intentionally corrupted, so AI can fail. Now we can speak about data reliability, and AI reliability is connected with data reliability.

Reliability is not easy to calculate or evaluate. In calculation of reliability input data is the biggest problem. Not because there is too much data, but because, sometimes, there is too little data. When we are dealing with maintenance, also the problem is more often that we do not have enough data. How many data we have about AI reliability, especially if it is new application?

3. DEFINITION OF AI

There is no generally accepted definition of artificial intelligence [6]. According to the Encyclopaedia Britannica dictionary [7] artificial intelligence is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. According to the Merriam-Webster dictionary [8], AI is a branch of computer science dealing with the simulation of intelligent behaviour in computers, or the capability of a machine to imitate intelligent human behaviour.

Government of the Republic of Serbia ([6, 2] accepted next definition of AI, also used in [9]: "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals."

Artificial intelligence-based systems can be purely software-based and operate in a virtual world (for example: virtual assistants, photo analysis software, web browsers, recommendation systems, speech and face recognition systems) or they can be embedded in devices - hardware (for example: advanced robots, autonomous vehicles, drones, etc.) [2, 9].

Artificial intelligence was founded as an academic and scientific discipline in the middle of fifties of the last century, and since then development has gone in different directions, being divided in sub-fields. Therefore, it is not a surprise that the definition of AI has been changed during time.

In the history of AI, there have been ups and downs, starting with the logic-based approach (during the 1950s and 1960s), the knowledge-based expert systems approach (1970s and 1980s), and the data-based approach (since 2000) years onwards - with periods of disappointment and reduced investment [6]. In the last decades, AI is defined as a study of intelligent agents - any device that precepts its environment and takes actions (by learning or using knowledge) to achieve its goals [1].

Essentially, the use of AI is an attempt to replace human intelligence with machine intelligence. Because of that, sometimes, AI is called machine intelligence. But it does not mean that AI learn just copying human. Nowadays AI can learn in its own way. And it, maybe, gives IT a potential to outperform human (brain).

4. RELIABILITY OF AI

As it is mentioned before, everything can fail, and Artificial intelligence (AI) is not an exception. If AI is an attempt to replace human intelligence with machine intelligence, and human reasoning can sometimes fail, so AI can fail in a similar way. So, this is the reason of erroneous reasoning (erroneous concluding, decisioning) or wrong learning? Having that in mind, can we raise the question about the reliability of AI, or how to avoid AI to fail? [1].

If AI fails, it is unreliable, so can we trust in it?

4.1. Trustworthiness

This is an important question which attracted the attention of ISO/IEC. In [10], there are surveys of topics related to the so-called trustworthiness in AI systems, including the following: (1) approaches to establish trust in AI systems through transparency, explainability, controllability, etc.; (2) engineering pitfalls and typical associated threats and risks to AI systems, along with possible mitigation techniques and methods; and (3) approaches to assess and achieve availability, resiliency, reliability, accuracy, safety, security, and privacy of AI systems. In that document, trustworthiness is defined as an ability to meet stakeholders expectations in a verifiable way, including the characteristics of trustworthiness such as reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, and usability.

It seems that trustworthiness is broader term than reliability.

In [11] under the headline "Enhancing the reliability of artificial intelligence" is stated that "Computers that learn for themselves are with us now. As they become more common in 'high-stakes' applications like robotic surgery, terrorism detection and driverless cars, researchers ask what can be done to make sure we can trust them. "So, are they reliable? Or, can they fail? Or can we fool them?

In [12] is discussed why deep learning AI is so easy to be fooled. An example is in a self-driving car application in a real situation. It is, however, said that it can happen in the case of sabotage as well.

There are examples of erroneous AI. From these examples we can derive some reasons for AI to fail, such as: quality of data, small or insufficient number of training and test data, training on unreal data, nature of learning algorithms, user incompetence.

4.2. Quality of data

Quality of data can have significant impact on the reliability of AI. What if AI learns from corrupted data or data which are corrupted on purpose (for example by an enemy)? The consequences can be different, from insignificant to catastrophic. So we need some kind of

control on the data which AI uses to learn. This problem doesn't imply not to use AI, or not to allow AI to learn from data, but to be aware of the problem in other to prevent difficult consequences.

Some examples can be found in [13], AI failures from IBM, Microsoft, Apple, and Amazon. The example from IBM happened in 2013, when IBM partnered with the University of Texas MD Anderson Cancer Center which developed a new "Oncology Expert Advisor" system with the goal to cure cancer [14].

Obviously, quality of data influence reliability. That is the reason to speak of data reliability. Data reliability refers to the completeness and accuracy of data as a measure of how well it can be counted on to be consistent and free from errors across time and sources. The more reliable data is, the more trustworthy it becomes [15].

There are authors who ask questions about potential risks, such as whether AI will pose an existential threat to humanity, or whether AI technology will be concentrated in the hands of the few [13]. But is not only question of reliability.

4.3. Nature of the algorithm

There is also a question raised whether AI can fail to function as expected, and the reason is because of the nature of the machine-learning algorithms on which modern AI techniques are commonly built. These algorithms are capable of learning from massive amounts of data, and once that data is internalized, they are capable of making decisions experientially or intuitively like humans. This means that for the first time, computers are no longer merely executing detailed pre-written instructions but are capable of arriving at dynamic solutions to problems based on patterns in data that humans may not even be able to perceive. This new approach comes at a price, because many of these algorithms can be black boxes, even to their creators [13]. Algorithm of AI is good if data from which this algorithm learns is good. So, again we can speak about data reliability.

4.4. Number and nature of data

One of the important questions is whether AI can work on a small number of data, for example, the number of failures of a product. In [16], it is concluded that model's predictive accuracy depends on the relevancy, sufficiency, and quality of the training and test data. Two questions are commonly asked with regard to failure history data: (1) How many failure events are required to train a model? And (2) How many records is considered as "enough"?

In July 2018, StatNews reviewed internal IBM documents and found that IBM's Watson was making erroneous, downright dangerous cancer treatment advice. In [12], it is concluded that, probably, the reason is because the software is trained on a small number of hypothetical cancer patients, rather than on real patient data.

Reliability is also important for security of AI (for example protection of unauthorized access).

4.5. User competence

It seems to me that user competence or incompetence in using AI can also be the reason for inadequate results which AI can delivered. Important question is if human can control AI.

As a conclusion for AI reliability, or trustworthiness (which is a broader term), let us use the statement from [17]: „Having the capability to generate tremendous benefits for individuals and society, AI also gives rise to certain risks that should be properly managed„ and „It is known that humans are biased in their decision making. Since AI systems are designed by humans, it is possible that humans inject their bias into them, even in an unintended way. "

Analysing failures of AI, can help to improve AI reliability. But on the other hand, use of AI can help to improve its reliability, and evaluate of other products, and also improve maintenance (methodology and training).

4.6. AI and human reliability

As we stated before AI can learn from human, and human knowledge can be erroneous. Because we can speak about human reliability, and if human is involved in AI, obviously AI reliability is connected with human reliability, because human action can influence reliability of AI.

I considered human reliability important from the beginning of my work in reliability, so human reliability is included in my textbooks and lectures for students [18, 19].

5. CONCLUSION

Artificial intelligence (AI) is in nowadays applying in almost every area of human life, and can has big impact. Everything can fail to function properly or as expected, and AI is not an exception, so as any other product AI must be reliable.

Some reason why AI is not reliable are inadequate quality of data used to learn, small or insufficient number of training and test data, training with unreal data, change of data, nature of learning algorithms, changes of user behaviour.

Someone must have in mind that AI is mainly software and uses data to learn.

In order to avoid situations when AI becomes unreliable, it is important to be familiar with causes of its unreliability.

To build successful AI, there is a need to be familiar with cases when AI failed in order not to make the same mistakes. Analysing failure of AI can help to improve AI reliability. Designers who want to apply AI must have in mind reliability.

If humans are biased in their decision making, this can

happen also with AI if AI learn from human.

AI is dominantly software, but can involve hardware and human, and of course data, so reliability of AI is related with software reliability, hardware reliability, human reliability and data reliability. Software reliability is usually more complex problem than hardware reliability, and AI reliability is more complex than any of these included human and data reliability.

References

[1] POKORNI S. *Current State of the application of Artificial Intelligence in Reliability and Maintainability*. Vojnotehnički glasnik/Military Technical Courier, 2021, Vol. 69, Issue 3, pp. 578-593, DOI: 10.5937/vojtehg69-30434, https://doi:10.5937/vojtehg69-30434, ISSN 0042-8469, UDC 623 + 355/359

[2] Ethical guidelines for the development, implementation and use of reliable and responsible artificial intelligence. "Sl. glasnik RS", br. 23/2023. Available at: http://demo.paragraf.rs/demo/combined/Old/t/t2023_03/SG_023_2023_007.htm. (In Serbian)

[3] POKORNI, S. *Reliability of Internet of Things*. In: 8th International Scientific Conference on Defensive Technologies OTEH 2018, Belgrade, pp.567-570, 11-12 October [online]. Available at: http://www.vti.mod.gov.rs/oteh18/elementi/rad/027.htm. ISBN 978-86-81123-88-4.

[4] POKORNI, S. *Reliability prediction of electronic equipment: problems and experience*. In: 7th International Scientific Conference on Defensive Technologies OTEH 2016, Belgrade, pp.695-700, 6-7 October. ISBN 978-86-81123-82-9.

[5] POKORNI S. *Data-driven reliability and availability of electronic equipment*, Vojnotehnički glasnik/Military Technical Courier, 2023, Vol. 71, Issue 3, pp. 769-782, DOI: 10.5937/vojtehg71-43474, https://doi:10.5937/vojtehg71-43474, eISSN 2217-4753, ISSN 0042-8469, UDC 623 + 355/359

[6] Government of the Republic of Serbia. *Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period 2020-2025*. 2019. [online]. Available at: https://www.srbija.gov.rs/tekst/en/149169/strategy-for-the-development-of-artificial-intelligence-in-the-republic-of-serbia-for-the-period-2020-2025.php

[7] COPELAND, B.J. *Artificial intelligence* In: Encyclopedia Britannica. London: Encyclopedia Britannica, Inc. 1998. [online]. Available at: https://www.britannica.com/technology/artificial-intelligence

[8] Merriam-Webster. Artificial intelligence. In: *Merriam-Webster.com dictionary*. 2020. [online]. Available at: https://www.merriam-webster.com/dictionary/artificial%20intelligence

[9] European Commission. 2019. *A definition of AI: Main capabilities and scientific disciplines.* Brussels: European Commission Independent High-Level Expert Group on Artificial Intelligence [online]. Available at: https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

[10] ISO. 2020. *ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence* [online]. Available at: https://www.iso.org/standard/77608.html?browse=tc

[11] University of Cambridge. *Enhancing the reliability of artificial intelligence*. 2016. [online]. Available at: https://phys.org/news/2016-10-reliability-artificial-intelligence.html

[12] HEAVEN, D. *Why deep-learning AIs are so easy to fool*. Nature, 09 October 2019. [online]. Available at: https://www.nature.com/articles/d41586-019-03013-5

[13] BATHAEE, Y. *The Artificial Intelligence Black Box and the Failure of Intent and Causation*. Harvard Journal of Law & Technology, 2018. 31(2) [online]. Available at: https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf

[14] BLIER, N. *Stories of AI Failure and How to Avoid Similar AI Fails.* Lexalytics, 30 January 2020. [online]. Available at: https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020

[15] IBM What is data reliability? Available at: https://www.ibm.com/topics/data-reliability

[16] -Microsoft. *Azure AI guide for predictive maintenance solutions.* 2020. [online]. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/predictive-maintenance-playbook

[17] European Commission. Draft Ethics Guidelines for Trustworthy AI. 2018. Available at: https://digital-strategy.ec.europa.eu/en/library/draft-ethics-guidelines-trustworthy-ai

[18] POKORNI S. Reliability and maintenance of technical systems. Military academy. Belgrade. 2002. (in Serbian)

[19] POKORNI S. Reliability of information systems. Information Technology School. Belgrade. 2014. (in Serbian)