ISSN: 2227-7994

FORUM A+P

INTERDISCIPLINARY JOURNAL OF ARCHITECTURE AND BUILT ENVIRONMENT

Title: *Guidelines for Risk Evaluation in Artificial Intelligence Applications*

Author: *Luca Lezzerini, Andia Vllamasi*

Source: *Forum A+P 27 | Venturing into the Age of AI: Insights and Perspectives*

ISSN: *2227-7994*

DOI: *10.37199/F40002714*

Publisher: *POLIS University Press*

Guidelines for Risk Evaluation in Artificial Intelligence Applications

LUCA LEZZERINI
POLIS University
ANDIA VLLAMASI
POLIS University

Abstract
Artificial intelligence is becoming a common element of our times. It is becoming more and more pervading into every element of our lives. Mass applications of artificial intelligence started when it began to be used in video games, but now it is available to everyone and can help with many tasks that, up to a few years ago, could be done only by humans. Discussions about artificial intelligence began very early before it existed. Most of the science-fiction literature tried to imagine many forms of AI and the consequences, both good and evil, of its use. But now artificial intelligence is a real, concrete thing and its mass usage must be subordinated to a risk evaluation and mitigation process to make it safe. In this paper, an introduction to this risk assessment will be made and the main guidelines for it will be defined. These guidelines could be used by researchers, designers, developers and even users to validate an AI-based application before delivering it to people. The paper considers the basic concepts of risk and tailors them to provide effective support in developing risk analysis for the specific area of artificial intelligence. Then a set of typical risks are defined and methods to detect and minimize them are provided. In conclusion, a call for stricter regulation of AI and high-performance processing is issued.

Keywords
Artificial intelligence, risk assessment, risk management

Introduction

Artificial intelligence technology is today a reality. It has moved out from sci-fi novels to enter people's daily lives. As with any mass product, AI must satisfy requirements about its risks for users. And these requirements require that people will be exposed to no serious risk. No well-defined standard exists today to provide a risk assessment for AI, but it is the first step for risk management. Although no standard exists, common tools to manage it are known and this will be described later. The traditional approach to risk and impact assessment follows a six-step process that can be described as identifying risks, prioritizing them, defining mitigation strategy, defining a monitoring plan, testing and reassessing risks and, finally, clearly communicating risk and mitigation procedures. Based on this six-step process, it is necessary to analyze how artificial intelligence can become a risk. For each risk, the risk assessment procedure must be applied, and a risk quantification carried out. The heart of the process is thus precisely the identification of how an artificial intelligence can become a risk.

Research question

The questions to be answered in this article are:

- what are, in a general sense, the areas of risk that can be gen erated by an artificial intelligence system in the current state of the art?
- how should this risk be assessed?
- who is responsible for assessing this risk?
- are there guidelines for legislators to define risk mitigation policies at the regulatory level?

Methodology

The methodology followed in this article begins with a review of the scientific literature relating to the risks of artificial intelligence and then moves on to define the typical risk areas of artificial intelligence in the current state of technology. This will be followed by a presentation of the main risk assessment methodologies and then one will be applied as an example. Based on the process described above, the allocation of responsibilities to the different actors in the entire life cycle of artificial intelligence will then be analyzed, and based on this, guidelines will also be produced for the various actors and, in particular, for legislators.

Literature review

Since the publication in 1921 of "R.U.R." (also known as "Rossum's Universal Robots") the Czech novel that gave the origin to the term "robot" from the Czech word "robota", meaning "hard work", a long time has passed. Yet today, Karel Capek's words, the author of "R.U.R.", are echoing and seeming very actual "The product of the human brain has escaped the control of human hands" (Capek, 1920). Science fiction has deeply investigated the theme of AI and, passing through the famous "Colossus", the main character of the homonymous novel written by D. F. Jones in 1966 where it is stated the incredibly fast growth of its AI "Listen to me: outside, in the vast world behind those doors, there are two machines. Less than twenty-four hours ago, they were busy proving to each other that two and two make four. Now they have reached the point where we hope to be in a hundred years. They think better and faster than we do, and I think we can only keep them under control in a very precarious way... but I have no desire to try to prove that..." (Jones, 2020). Isaac Asimov approached it in a very engineered way, in the 40s, by defining the legendary "Three Laws of Robotics" that have been the basis of many of his writings. Many of these trials to forecast the future of AI through novels have concluded that they will destroy us. "Terminator" with its SkyNet is only one of a plethora of examples. But now we are starting to have, in our reality, not in novels, Ais that are enough powerful to become a danger. So, now, it is time to apply a true risk management approach that should not be limited to the engineering aspect only, but also has to explore other contexts like social, political, emotional, psychological and more.

Some scholars (Khalif Ali et al., 2023) have tried to provide a literature review about frameworks for AI risk management, balancing trust, risk and security. They evidenced risks tied to bias and discrimination, privacy invasion, society manipulation, deepfakes, lethal autonomous weapon systems (LAWS), malicious use of AI, and insufficient security measures. All these risks have been grouped into three classes AI trust management, risk management and security management but, anyway, they are all risks to be mitigated despite their names. For each type of risk, (Khalif Ali., 2023) have also defined the possible types of damages. Then they defined a framework for such risk mitigation. According to their analysis, it is possible to apply a risk assessment and mitigation procedure to assess any of these risks and then find a set of countermeasures. Other scholars (Matloob et al., 2021) have developed a similar approach to specific cases, like in the cited paper, the application of AI to coal mining. Or in the financial risk management (Sheth, 2023). In the last two cited works, the application to a specific context has introduced a deeper detail but lost, in the perspective of this paper, the general view and the abstraction process needed to perceive the big picture behind AI risk management. But even starting from this research, it is possible to define a general set of AI-related risks.

Risk areas for AI

Analysing the various types of danger that artificial intelligence can pose, both in general research work and in specific ones applied to well-defined cases, it emerges that risks can be grouped into four well-defined general areas. (I. A. E. M. E., 2023, Weerts et al., 2023, GWAI-90,1990, Ghoz et al., 2023)

The first area is that of artificial intelligence which can produce specific damage such as emotional, physical or psychological. Other areas are those related to information generated by AI and violation of human rights impacts in the social or socio-political context. Damage in this first area is essentially physiological as it corresponds to risks arising from the use of artificial intelligence in general. An example of this first area could be the social isolation resulting from the humanisation of

the artificial intelligence system with which one moves to have a relationship that is no longer a human-machine but human-human, the relationship that generally becomes morbid or toxic.
A second area of risk is related to specifics and technologies that may bring with them risks that are not present in other technologies. In this second area, we find, for example, the risks connected with deepfake technologies, which allow the perception of reality to be altered through the creation of images, films or audio recordings that are extremely realistic but completely false, and which can mislead their users by providing them with misleading information. Another very important area of risk is the one related to the use scenarios of artificial intelligence systems. In this case, there are all the more engineering-type problems involving malfunctions, misuse and variations in the operating environment. This area of risk is very similar to that of any other device created by human ingenuity, and therefore risk assessment structures typical of systems engineering processes already exist for it. Analysis methodologies such as FMEA (Failure Mode Effect Analysis) (Schoitsch, 2014, Fregnani, 2022) or FMECA (Failure Mode Effect and Criticality Analysis) can be directly applied to specific cases, with the important point not to forget that, since artificial intelligence normally reacts actively and proactively to the context in which it is used, this context must be considered among the factors that can cause the failure modes of the aforementioned analyses.
An example could be that of an autonomous driving system of a vehicle that, in the face of a sensor failure, no longer has a correct perception of space and therefore becomes dangerous. As already mentioned, this type of problem is already widely considered in current engineering processes. The impact of context on the operation of artificial intelligence, on the other hand, is something in the making, as failure modes resulting from changes in context are not yet part of the average engineer's experience. A very interesting example can be that of the always self-driving systems that perceive the presence of vehicles as they detect images posted along the road or on surrounding vehicles, images that represent vehicles, or they incorrectly calculate the speed of movement and thus incorrectly classify people as vehicles or vice versa. The latter case is one of the possible cyber attacks that can be perpetrated against a self-driving vehicle.
Another type of risk is associated with specific application domains in which artificial intelligence can be used. In this risk category, the dangerousness of artificial intelligence is considered to depend on the domain to which it is applied. For example, an object classification applied to a home alarm system has a completely different dangerousness than an application in the automotive or aeronautical fields. The same can be said for classification systems applied in the financial domain rather than in the occupational safety domain. In conclusion, we can classify these risks into four categories such as risks associated with typical AI hazards, risks associated with the particular technology, risks associated with operational scenarios, and risks associated with the specific application domain of artificial intelligence.

Risk assessment and mitigation methods

Genuine standards for risk assessment in connection with the use of artificial intelligence are still being developed. To date, numerous attempts have been made, both in the academic sphere and in the national or supranational legislative sphere, to define specific risk methodologies to be able to certify the non-hazardousness of specific artificial intelligence systems. The first consideration is that these risk analysis models should be able to anticipate dangers before they materialise, i.e. be based on a preventive rather than a corrective approach through, for example, monitoring mechanisms or, even less preferably, through protection systems capable of intervening to reduce the damage when it occurs. As happens in any risk management system, we, therefore, have three possible approaches: the preventive approach based on the prediction of the risk and its elimination through prevention, the predictive approach based on the monitoring of symptoms that make it possible to predict the occurrence of the damage and act in time when it is not yet there or is at a tolerable level, and finally, the protective approach that implements protection mechanisms, active or passive, capable of protecting the object of the damage, whether human, economic, social or other, from the damage itself, either by avoiding any impact or by greatly reducing its effect. These three approaches, as already mentioned, are characteristic of all risk management methodologies and must apply to the entire life cycle of artificial intelligence, from its conception to its decommissioning.
In some states, directives or laws have been issued that attempt to standardise risk assessment and management methods in both the artificial intelligence and data governance sectors. One example of such a model is the European Commission's 2021 model (EU Commission, 2021), another is the so-called General Data Protection Regulation or GDPR of 2016. The first aims to reduce the dangerousness of the application of artificial intelligence by placing constraints on risk assessment. The second is only concerned with prohibiting the application of automatic decision-making or automatic profiling systems if these impact the rights and freedoms of individuals (Art. 22) and subject to certain special cases. Wanting to develop guidelines for the analysis of risks arising from the application of artificial intelligence, given the lack of standard frameworks both at the national and supranational level, it might be convenient to follow a strategy based on the analogy with the current systems of risk assessment and management derived from the electromechanical engineering sector and the cyber security sector. Wanting to derive a typical outline of a risk assessment process and its management for its mitigation, one may consider, analysing specific techniques such as the aforementioned FMEA/FMECA or information security frameworks, that six basic steps are necessary for this outcome. The first step is risk identification, which typically consists of drawing up a list of risks that the system presents. In the traditional engineering approach, these risks are seen as the loss of system functionality (failure modes), whereas in the information security approach, these risks are seen as the loss of one or more of the security

requirements for an information asset by a threat, i.e. an external internal agent, either intentional or unintentional, that may compromise, for example, its confidentiality or availability or integrity. In both approaches, assets are defined as those elements that one wishes to protect, be they system functionality or data, and for each one, one sees how they may be compromised. This potential "compromised state" is precisely the risk that one wishes to identify.

The second step consists of defining a priority ranking among the identified risks to deal first with the highest ones and then gradually move down to the less important ones. There are various methods of prioritisation, the two most common being one based on the product of the probability of occurrence and the severity of the occurrence of the risk, and the other on the further multiplication by the so-called detectability, i.e. the ability to detect a premonitory symptom of the occurrence of the risk.

The first approach emphasises only the preventive mode as a mode of risk management, while the second approach implicitly admits the possibility of predictive risk management as well. Both situations also allow the protective approach to be used, but this should be considered as a last resort. In the approach based on the product of probability and severity, one obtains an indicator called magnitude, which is used to prioritise the risk because the higher the magnitude, the more the risk is prioritised. In the other approach, the one also based on the possibility of detecting premonitory symptoms, an additional indicator will be produced, usually referred to as RPN or Risk Priority Number. Regardless of which approach you have chosen, once the prioritisation indicator has been calculated, you can proceed to sort the list of risks according to priority. At this point, you can begin the third step, which concerns the design of risk mitigation. In this activity, starting with the most prioritised risks, mitigation strategies are defined, which could be those already mentioned of prevention, prediction and protection. In some cases, a fourth modality is also considered, which consists in transferring the risk, either by taking out insurance to cover the economic aspect of the damage that may be generated by the occurrence of a risk. For each of these mitigation options, a cost-benefit analysis will have to be carried out, and at the end of the day, based on this analysis, one or more mitigation strategies will be defined for each risk, which will then have to be implemented.

The next step is the implementation of the mitigation strategies. At this stage, all mitigation strategies that need to be practically deployed in the field through design changes, through the acquisition of additional devices, or through training are implemented and their outcome validated. For all aspects of design modifications or the acquisition of additional devices, it may be that the risk assessment must be repeated or that specific worst-case tests have to be carried out to highlight unacceptable residual risks or new risks introduced by the devices that should mitigate the risk. The next step is the monitoring of the risks, and this monitoring can be carried out in two ways. The first method is to repeat the risk assessment periodically. Periodically may be on a time basis, i.e. setting a certain time interval (such as six months), and then repeating it, or it may be based on the occurrence of events. These events may be of an organisational or other nature, or they may relate to the occurrence of failures or, finally, to evidence from indicators that form part of the second type of monitoring. The second type of monitoring consists of continuously calculating indicators that assess the validity of the risk analysis and its management. These indicators can also be used to trigger risk review activities. The two monitoring methods are rather integrated. Both should be applied during the development phase of the AI as well as during its operational phase. The purpose of monitoring is twofold in that it is concerned both with highlighting the emergence of any risks that threaten to become real and with assessing the soundness of the risk analysis carried out and the resulting risk mitigation work.

At all stages of designing the risk management system and designing the artificial intelligence, not only the various indicators that will be the subject of operational monitoring will have to be defined, but also the indicators that will be used during the AI development phase to assess its riskiness. Another step is testing and validation. In this phase, the artificial intelligence is subjected to a series of tests that verify its correct implementation according to the design (verification), but at the same time provide information on whether the artificial intelligence is mature for the task that will be assigned to it. Both verification and validation activities are partly based on the traditional metrics of artificial intelligence (accuracy, precision, sensitivity, mean absolute error, ...) and partly on the risk indicators developed during the design phase of operational monitoring. Consequently, during the verification and validation phases, we will not only look at the performance of the artificial intelligence in functional terms but also consider how effective the risk mitigation system is. For instance, malfunctions may also be simulated during this phase to verify how the artificial intelligence will behave in such cases. Unfortunately, this type of test is extremely complex because most of the problems of artificial intelligence lie in the intelligence matrix, i.e. in that part where knowledge resides is the set of rules that artificial intelligence applies, a part that very often if not always and in the eyes of humans is practically an opaque black box. (Avin, 2021) A good example of an indicator for assessing the risk of an artificial intelligence that is already widely used and easy to implement is the so-called confusion matrix. This is a table in which, in the case of a binary classification, for example, both the exact results and the false positives or negatives are shown. From this confusion matrix, it is possible to calculate the cost of error and assess the probability of a certain risk, understood as the probability that a certain unwanted error will occur. Although very simple, this indicator makes it possible to roughly estimate, in probabilistic terms, the risk of committing misclassifications and to associate the cost, i.e. the severity of the damage, with each of them. Consequently, the confusion matrix is an excellent tool for calculating the magnitude of risk as a product of probability times severity and is therefore extremely practical in use and also highly automatable. The

same tool can also be easily extended to non-binary classification cases. Once the confusion matrix has been calculated in the AI development phase, it can be periodically recalculated a posteriori, in the operational phase, having the exact answers available, it is compared with the design matrix to check for any significant deviations, which could indicate the emergence of a risk considered mitigated or acceptable. (Ramli et al., 2022) A final step in the risk assessment and mitigation process is communication and information. In practice, this involves defining information, often common to relatively different systems, to be transferred to users, be they specialists or masses. This type of transparency allows both users to make safer use of the AI system and to highlight anomalous behaviour.

Policies and Responsibilities

Based on the risk analysis and mitigation process outlined above, it is possible to define responsibilities in the identification and management of risk. As in any engineering activity, the first actor in a risk assessment is precisely the engineering team that designs and develops the artificial intelligence system and that, like all other engineering systems, will have to go through a hazard review based on risk assessment and mitigation.

Allocating the responsibility for risk management to the engineering team alone is not enough because, in addition to having to consider those who will have to use and maintain the AI system, it is necessary to standardise risk assessment and risk management procedures. Without standardisation, it becomes difficult to define whether an AI system is safe enough to be placed on the market or not, leaving the assessment to the engineering element alone, an element that could be pressurised by the manufacturer or that could make mistakes that would not be detected until too late. Thanks to the use of standardised policies, AI products might have to undergo a risk assessment phase before they can be placed on the market, and thus be subject to a standardised evaluation, which can be improved over time based on experience, and can also improve the risk assessment phase of the engineering team, by placing it in a transparent scheme common to all. Standardisation would also lead to savings in design and development costs while increasing the safety of the AI system, as has been amply demonstrated in other cases involving engineering products.

Emerging standards for AI risk management

At present, a variety of AI-specific risk assessment methods exist, virtually all of them evolving rapidly and with considerable levels of detail. Unfortunately, none of them is a widely accepted standard, most of them having applicability at the level of a single nation. These standards are all rather immature although it is worth noting that the International Organisation for Standardisation itself has moved forward with its ISO/IEC 23894:2023 Information Technology - Artificial Intelligence - Guidance on risk management (ISO/IEC 23894:2023, 2023). Although an important contribution to the standardisation of risk analysis and management practices for AI, this standard is still very young and is meant to be applied in conjunction with the ISO standard for risk management systems ISO 31000:2018 (ISO 31000:2018 , 2018), of which it is seen as an extension to a specific sector. Other standards (actually national regulations or guidelines) are the EU AI Act (EU Commission, 2021), the UK Online Safety Bill (UK, 2023) and the US Algorithm Accountability (Mökander, 2022), the Italian White Paper on Artificial Intelligence (IT, 2020). Other standards have been developed or are being developed in other countries, for example in Brazil (Uechi et al., 2023).

All these standards have, as a common feature, the attempt to prevent the damage that an AI could potentially cause, either by acting at the development level and before use or during use. They all also try to take into account the fact that AIs evolve during use and that, therefore, their behaviour may change with the arrival of new training data, in a phenomenon very similar to that of regression in software releases. Many of these methods are beginning to become legislative requirements for the development and deployment of AI systems, thus becoming real general policies that can be used to define certification schemes on the safety of AI systems. In addition, a process very similar to the one described in Section 5 above can be found in these standards. Precisely because of the immaturity of these systems, there is now a high risk of artificial intelligence products being placed on the market that may present unacceptable risks.

Conclusions

Standard risk analysis methodologies (e.g. ISO 31000) can be applied to artificial intelligence systems, but considering risk profiles that are in part extremely innovative and different from those typical of other systems. Given the ever-increasing pervasiveness of artificial intelligence in today's widely used systems, but also in systems of limited use but of high criticality, the need to define reliable risk analysis and mitigation methodologies and to introduce mandatory legislative requirements for their application before the approval of artificial intelligence systems for use outside development laboratories becomes increasingly urgent.

These legislative requirements are the responsibility of legislators who, given the current bureaucratic slowness in defining mandatory certification policies, may even have to impose a temporary suspension of the use of artificial intelligence in various sectors, such as defence, mobility, medical applications and all those cases where there is a direct risk of damage to health, the environment or the economy of significant magnitude. The purpose of this suspension would be to allow the regulatory evolution path to be completed before allowing the use of AI applications in critical contexts. It is believed that ISO could be a good framework for technical standardisation, but it is suggested that a specific task force for artificial intelligence be set up at a global level with the task of guiding the various nations and the authority to select best practices, methodologies, guidelines and national policies, facilitating their sharing with other nations and supranational bodies, to pool the best results emerging at a local level.

Reference List

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., … Zilberman, N. (2021). *Filling gaps in trustworthy development of AI. Science*, 374(6573), 1327–1329. https://doi.org/10.1126/SCIENCE.ABI7176

Čapek, Karel. R. U. R. (2020) Rossum's Universal Robots: Kolektivní Drama O Vstupní Komeddii a Třech Aktech. *Aventinum*.

EU Commission (2021), Regulatory framework proposal on Artificial Intelligence. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai last accessed 30 October 2023

Fregnani, J. (2022). Safety Analysis Methods for Complex Systems in Aviation. *ArXiv* (Cornell University).

Ghoz, L., & Hendawy, M. An Inventory of AI ethics: Tracing 100 documents. *MSA Engineering Journal*, 2(2), 647–675. https://doi.org/10.21608/MSAENG.2023.291907

GWAI-90 14th German Workshop on Artificial Intelligence. (1990). *Informatik-Fachberichte*. https://doi.org/10.1007/978-3-642-76071-6

ISO 31000:2018 (2018) Risk management — Guidelines. Retrieved from https://www.iso.org/standard/77304.html last accessed 30 October 2023

ISO/IEC 23894:2023 (2023) Information technology - Artificial intelligence - Guidance on risk management. Retrieved from https://www.iso.org/standard/77304.html last accessed 30 October 2023

IT (20200) *Libro Bianco sull'Intelligenza Artificiale*, Retrieved from https://commission.europa.eu/system/files/2020-03/commission-white-paper-artificial-intelligence-feb2020_it.pdf last accessed October 2023

Jones, D. F. (2020) *Colossus*. Gollancz.

Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications*, 240, 122442.

Matloob, S., Li, Y., & Khan, K. Z. (2021). Safety Measurements and Risk Assessment of Coal Mining Industry Using Artificial Intelligence and Machine Learning. *Open Journal of Business and Management*, 09(03), 1198–1209. https://doi.org/10.4236/OJBM.2021.93064

Mökander, J., Juneja, P., Watson, D.S. et al. (2022) The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?. *Minds & Machines* 32, 751–758 (2022). https://doi.org/10.1007/s11023-022-09612-y

Paul, R. K., & Sarkar, B. GENERATIVE AI AND ETHICAL CONSIDERATIONS FOR TRUSTWORTHY AI IMPLEMENTATION. *Journal ID*, 2157, 0178.

Ramli, N. E., Yahya, Z. R., & Said, N. A. (2022) Confusion Matrix as Performance Measure for Corner Detectors. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 29(1), 256–265. https://doi.org/10.37934/ARASET.29.1.256265

Schmittner, C., Gruber, T., Puschner, P., & Schoitsch, E. (2014). Security application of failure mode and effect analysis (FMEA). In *Computer Safety, Reliability, and Security: 33rd International Conference, SAFECOMP 2014, Florence, Italy, September 10-12, 2014. Proceedings 33* (pp. 310-325). Springer International Publishing.

Sheth, N. (2023). Future of Artificial Intelligence and Machine Learning Systems in financial risk management and regulatory compliance. *International Journal of Software &; Hardware Research in Engineering*, 11(8). doi:10.26821/ijshre.11.8.2023.110808

Uechi, C. A. S. Moraes, T. G. (2023) Brazil's path to responsible AI. Retrieved from https://oecd.ai/en/wonk/brazils-path-to-responsible-ai last accessed October 2023

UK. (2023) Online Safety Act 2023. Retrieved from https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted last accessed 30 October 2023

Weerts, H., Xenidis, R., Tarissan, F., Olsen, H. P., & Pechenizkiy, M. Algorithmic Unfairness through the Lens of EU Non-Discrimination Law. 2023 *ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3593013.3594044