STUDIA IURIS

Jogtudományi Tanulmányok / Journal of Legal Studies

2024. I. évfolyam 4. szám

Károli Gáspár Református Egyetem
Állam- és Jogtudományi Doktori Iskola

A folyóirat a Károli Gáspár Református Egyetem Állam- és Jogtudományi Doktori Iskolájának a közleménye. A szerkesztőség célja, hogy fiatal kutatók számára színvonalas tanulmányaik megjelentetése céljából méltó fórumot biztosítson.
A folyóirat közlésre befogad tanulmányokat hazai és külföldi szerzőktől – magyar, angol és német nyelven. A tudományos tanulmányok mellett kritikus, önálló véleményeket is tartalmazó könyvismertetések és beszámolók is helyet kapnak a lapban.
A beérkezett tanulmányokat két bíráló lektorálja szakmailag. Az idegen nyelvű tanulmányokat anyanyelvi lektor is javítja, nyelvtani és stilisztikai szempontból.
A folyóirat online verziója szabadon letölthető (open access).

Kiadó: Károli Gáspár Református Egyetem Állam- és Jogtudományi Doktori Iskola
Székhely: 1042 Budapest, Viola utca 2-4
Felelős Kiadó: Tóth J. Zoltán

A tipográfia és a nyomdai előkészítés Csernák Krisztina (L'Harmattan) munkája
A nyomdai munkákat a Robinco Kft. végezte, felelős vezető Gembela Zsolt

Honlap: https://ajk.kre.hu/index.php/jdi-kezdolap.html
E-mail: doktori.ajk@kre.hu
ISSN 3057-9058 (Print)
ISSN 3057-9392 (Online)
URL: KRE ÁJK - Studia Iuris
https://ajk.kre.hu/index.php/kiadvanyok/studia-iuris.html

THE REGULATION OF ARTIFICIAL INTELLIGENCE OUTSIDE EUROPE. SECURE AI SYSTEM DEVELOPMENT, GUIDELINES FOR A BETTER FUTURE

Gergely Rideg[1]

ABSZTRAKT ■ A mesterséges intelligencia fejlesztése közkedvelt téma a 21. században. Ez a zabolázatlan technológiai újdonság olyan erővel és letaglózó hatékonysággal érkezett meg a modern gazdasági viszonyok közé, hogy az egyes tényezők hatását igazán még felmérni sem tudják. Egyre gyakrabban jelennek meg újabb és újabb kutatási eredmények azokról a kockázatokról, amelyek a mesterséges intelligencia rendszerek használatával járó különböző kockázatokat elemzik. A jelen tanulmány Nick Bostrom gondolatait is bemutatva és elemezve azokkal a szabályozási problémákkal foglalkozik, amelyek megfejtése a megbízható mesterséges intelligencia rendszerek működtetéséhez nélkülözhetetlen. A tanulmány megteremti a diskurzus alapjait azzal, hogy a kockázat, mint fogalom fundamentumait és határait részletezi, illetve kontextusba hozza és összekapcsolja a mesterséges intelligencia rendszerek kockázataival. Szabályozási kérdéseket vet fel, miközben olyan jó gyakorlatokra és iránymutatásokra hívja fel a figyelmet, mint a „Guidelines for secure AI system development for ensure the secure artificial intelligence development" elnevezésű dokumentum, amely többek között a UK National Cyber Security Centre szervezet részéről került kidolgozásra. A tanulmány olvasásakor betekintést nyerünk a Blueprint for an AI Bill of Rights dokumentumba és felépítésébe, illetve egyéb szakmai iránymutatásokba.

ABSTRACT ■ The development of Artificial Intelligence is a hot topic in the 21st century. This unbridled technological novelty has arrived in the modern economy with such force and staggering efficiency that the impact on individual actors cannot even be truly measured. More and more research are being published on the various risks associated with the use of AI systems. This paper, which also presents and analyses the ideas of Nick Bostrom, addresses the regulatory issues that are essential to the operation of reliable AI systems. The paper lays the foundations for the discourse by detailing the foundations and boundaries of risk as a concept and contextualising and linking the risks of AI systems. It raises regulatory issues, while pointing to good practices and guidelines such as the "Guidelines for secure AI system development to ensure the secure artificial intelligence development", developed

[1] PhD student, Doctoral School of Law and Political Sciences, Károli Gáspár University of the Reformed Church in Hungary.

by, among others, the UK National Cyber Security Centre. Reading the paper will provide insights into the Blueprint for an AI Bill of Rights document and its structure, as well as other professional guidelines.

Keywords: artificial intelligence systems, risk analysis, guidelines, cybersecurity, trusted AI, strategies, regulatory problem

1. Introduction

On 13 March 2024, the European Parliament approved a legislation on Artificial Intelligence (AI), which will help the technological innovation of AI while building safeguards to protect our security and fundamental rights.

The European legislator has, in our view, taken a giant step towards taming the technological monster that is now a daily topic of debate around the world.[2]

Why are we looking at regulating the use of artificial intelligence for business purposes? IBM's Global AI Adoption Index 2022 Index[3] found the following: "Today, 35% of companies reported using AI in their business, and an additional 42% reported they are exploring AI. AI adoption is growing steadily, up four points from 2021." In relation to trusted AI, the document highlights the following: "The majority of organizations haven't taken key steps to ensure their AI is trustworthy and responsible, such as reducing bias (74%), tracking performance variations and model drift (68%), and making sure they can explain AI-powered decisions (61%)".

It is clear that companies will pay much more attention to the use of artificial intelligence systems in the future. These technologies will be integrated into their operations. Whose interests will be served by these AI systems and what guarantees do we see?

The focus of the current research is artificial intelligence regulation outside Europe. The research questions are: How is artificial intelligence regulation developing outside the European Union? What are the cornerstones of AI regulation? What are the results, documents, research groups and initiatives that have been carried out so far on the development of artificial intelligence.

[2] Artificial Intelligence Act: MEPs adopt landmark law, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law (downloaded: 16.03.2024).

[3] IBM Global AI Adoption Index 2022, https://www.ibm.com/watson/resources/ai-adoption (downloaded: 10.02.2024).

The current research is part of a larger research which is now designed to outline the turning points reached by non-European actors.

However, in the following chapters, we will discuss what we consider to be risks and will mention the risks that AI can pose. And through the reflections of Nick Bostrom, we will discuss whether a superintelligence explosion is manageable or problematic.

2. Research method

The research questions will be answered through normative analysis, *interpretatio systematica* and contextual analysis. The analysis will also use an interpretation according to fundamental constitutional rights, as well as an interpretation according to the ethical values behind the law. In articulating the questions, we have kept in mind that the focus is not only on AI as a regulatory subject, but mainly on the role that AI will play/has played in each society, i.e. the social role of AI as a milestone.

3. The historical background of artificial intelligence and the control problem

There are written records from the early days of humankind that humans created fantasy stories about inventing a machine with properties beyond humans. Hephaestus built a bronze structure in the shape of a man, according to mythology. The machine, called Talos, was given by Zeus to King Minos to protect the island of Crete from invaders. The bronze giant, with physical strength beyond human limits, appears in Apollonios Rhodios' Argonautica[4].

Humans create machines so that they can perform an activity faster and more efficiently. Think of the invention of the printer. But in the case of artificial intelligence, there is one more important circumstance. Humanity is building a machine that is smarter and more intelligent than itself. What a paradox follows from this idea, when humans, although they wish to build a machine more intelligent than themselves, do not let its operation out of their hands and wish to exercise their power over that machine by expecting it to operate only according to their will.

[4] Dóra Peszlen: Apollónios Rhodios. Argonautika 3. *Studia Litteraria*, 1-4/2017, 37–41.

It is also important to mention these phenomena because, in the relationship between artificial intelligence and humankind, an important quality or characteristic must be highlighted in order to understand the complexity of the regulatory problem. It relates to the AI-control problem, which Nick Bostrom, Director of the Future of Humanity Institute at Oxford University, describes in his book "Superintelligence: Paths, Dangers, Strategies"[5].

Bostrom puts it bluntly that superintelligence is probably "the most important and greatest challenge humanity has ever faced."

Here, however, we note that what Bostrom is talking about is a superintelligent artificial intelligence, which in his interpretation means a consciousness beyond the smartest human mind. We can take it as a fact that today this kind of artificial intelligence does not exist in this form, and scientists are divided on whether such intelligence could ever exist. Nonetheless, with appropriate abstraction around some of the characteristics of AI and the interpretation of regulatory challenges, Bostrom's thoughts are interesting and useful for the present research questions.

Many experts around the world support Bostrom's claim when they report on the various risks of using artificial intelligence in their analyses. We will not go into a more detailed analysis of the risks of using AI here, but it is necessary to highlight a few ideas in the context of what this paper has to say.

4. The relationship between artificial intelligence and risk

The risk of using AI is a concern for legislators in Europe and beyond. The European legislator has taken explicit steps to identify the scientifically substantiated risks and has started to look for the associated safeguards and guarantees.

On the one hand, the European Commission has identified the benefits of AI in terms of technological improvements for citizens, businesses and public services, such as fewer machine breakdowns, safer transport, better healthcare.[6]

In addition to the benefits, AI also comes with several risks. On the one hand, automatic surveillance, which can be useful, for example for crime prevention purposes[7], can easily lead to a violation of citizens' autonomy. Although the example above also results in a serious violation, people often associate larger scale and more sinister scenarios with irresponsible and financially driven AI

[5] Nick Bostrom: *Superintelligence. Paths, Dangers, Strategies*. Oxford University Press, 2014.

[6] COM(2020) 65 final 2.

[7] In Mannheim, an automated system reports hugs to the police, https://algorithmwatch.org/en/mannheim-system-reports-hugs-police/ (16.02.2024).

development.[8] The science fiction literature and film industry played no small part in this. Autonomous weapon systems that cause death are often mentioned in this context. The development of these is currently shrouded in obscurity, but a large part of the international legal community is strongly in favour of a prior ban on such weapons. An example of such a ban is the 1995 ban on the use of laser weapons that cause permanent blindness.[9] Autonomous weapons systems, also known as killer robots, raise serious ethical concerns and could be a catalyst for a dangerous arms race. The number of potential victims is yet scientifically incalculable, which is why it is a fact that they pose immense risks at the level of society as a whole.

What do we mean when we say that using artificial intelligence is risky? What do we mean by risk? Nowadays, almost every day, everyone can see the use of artificial intelligence on the front pages of the press in various areas of our daily lives. From the artificial intelligence solutions in medicine[10] to the facial recognition applications lurking on our mobile phones, we see artificial intelligence analysing millions or billions of personal data almost every hour. What everyone probably already knows is that artificial intelligence is a risky entity. However, there is concern that society is only aware of this assumption about AI and that it is not being looked at with a sufficiently complex vision.

The risk of AI has been mentioned above, but we have not addressed what exactly we mean by risk. In the following, I will make some basic observations in this context, which will be used to assess the results of the European entities in the following chapters.

According to general risk theory, risk is obtained by considering (multiplying) two factors, the probability of a negative event occurring on the one hand, and the magnitude of the negative consequences of the event on the other hand.[11]

[8] Ethical Guidelines for Trusted Artificial Intelligence prepared by the High Level Expert Group on Artificial Intelligence.

[9] Protocol on Blinding Laser Weapons, https://fas.org/nuke/control/ccw/text/protocol4.htm (downloaded: 17.11.2020).

[10] In recent years, there has been considerable research into the use of AI in the medical field. According to a study published in the Journal of the American Medical Association, AI-based systems can, among other things, accurately diagnose common skin conditions with an accuracy comparable to that of human dermatologists. Another study published in the journal Nature showed that AI can analyse genetic data and identify personalised treatment options for patients with rare diseases.

[11] József Kindler: Általános kockázatelmélet és -módszertan. Egyetemi jegyzet. 1983. cited by Tamás Fleischer: Innováció, növekedés, kockázat. In: Miklós Bulla – Pál Tamás (eds.): *Fenntartható fejlődés Magyarországon. Jövőképek és forgatókönyvek.* Budapest, Új Mandátum Könyvkiadó, 2006. 275-284. http://real.mtak.hu/3973/1/fleischer_innovacio-novekedes-kockazat_fefemao06.pdf (01.03.2021).

Also, different disciplines have different interpretations of risk.[12] In economics, for example, loss is accompanied by the optional occurrence of gain. In general, however, it is associated with the possibility of danger, injury, damage or death. From the above equation, it can be deduced that both higher frequency and more severe consequences lead to an increase in risk. There are different classifications of risk, so we can and should distinguish between, for example, individual and global risk, direct and latent, controllable and uncontrollable, voluntary and involuntary, subjective and objective risk, etc. Objective risk, which is more relevant to our topic, is estimated based on a large number of repeated experiments, while subjective risk estimates are based on a small number of observations or possibly on conjecture. In between the above two is synthetic probability, where the probability of an event occurring is not measured directly but modelled. In modelling, the event is estimated based on similar objective probability systems.[13]

However, we also see that people tend to base their expectations and decisions on subjective estimates of risk, so there is a large discrepancy between social perception and the outcome of a scientific, empirically measured probability estimate,[14] which in turn determines their confidence in the risky thing. This concept leads us to the regulatory side of risk analysis and risk assessment.

Trust is key to the regulation of artificial intelligence, as András Tóth argues in relation to the paradox of regulating artificial intelligence; to fulfil the purpose of AI for the benefit of people, trust must be instilled in the technology.[15] Guarantees must be built into the legislation that is being developed to ensure human rights and ethical principles in AI applications. The regulation must therefore build in safeguards to ensure that when AI is used to serve people, it does not violate fundamental human rights, it is transparent in its operation and the decision-making process can be monitored. For example, the autonomous heavy machinery installed in the Mohács iron foundry should not be able to harm its operators in the event of a malfunction.

When identifying the risks posed by AI applications, it is therefore appropriate to use objective estimates, which should be based on many repetitive cases identified as a result of scientific studies (precaution principle).

The context of the analysis is defined by our view – in agreement with Ulrich Beck's ideas on the subject – that in the 21st century we live in a risk society.

[12] Fedja Netjasov – Janic Milan: A review of research on risk and safety modelling in civil aviation. *Journal of Air Transport Management*, 4/2008, 213–220.

[13] Ádám Havas: Kockázatelemzés-mágia vagy tudomány? *Iskolakultúra*, 23/1993, 21–28.

[14] Fleischer 2006.

[15] András Tóth: A mesterséges intelligencia szabályozásának paradoxonja és egyes jogi vonatkozásainak alapvető kérdései. *Infokkumunikáció és jog*, 73/2019, 3–9.

Advanced industrial society is itself an extractor of social risks, and "differences in education, skills, access to information and income determine the risk burden of different social groups".[16]

In the context of the above, Bostrom's book[17], in which he analyses the dynamics of the explosion of super intelligence that does not yet exist, contains some very exciting ideas. It looks at what happens once this intelligence is in place, and how we can create the initial conditions to drive this particular explosion towards positive outcomes.

Among many other factors, Bostrom sees the risk of AI in the agent principal problem as follows. He poses the question, "how can the sponsor or promoter of a project to develop superintelligence ensure that the superintelligence created by a successful project will serve the sponsor's goals?" Putting the question in a slightly more common law context, we think of a company whose owners wish to create an artificial intelligence system with a particular function, for which they set out the ethical, moral, economic, and functional guidelines they consider important. How can they be sure that the system they create will meet all their guidelines? The owners entrust the company's chief executive officer (agent) with the task. The principal-agent problem, well known to economists, is encountered. This is an important case of incomplete information games. Here, the following factors are at play: an agent has choices; he is expected to make decisions in the best interests of the principal; but the principal cannot observe the choices he makes, the choices he has made, the alternatives he has chosen, or whether he has chosen the best choice.[18] The agent can make good decisions and bad decisions. It can be a bad result in the light of the fact that the agent has otherwise made good decisions. By bad decisions we mean when he has made a decision because, as a trustee, he does not bear the consequences of his decision, because if he had, he would have made a different decision. Of course, we can take security measures to ensure that the principal carries out the task entrusted to him in the most favourable way for the principal. But this obviously comes at a price. It is a question of cost-effectiveness. The project owner must consider what resources he can allocate to motivate the agent, the software developers, to make the desired decisions. It may also decide to introduce more stringent controls or a stricter screening of developers.[19] In these cases, however, the potential damage to the risk side of AI systems must always be considered.

16 Fleischer 2006, 279.

17 Bostrom 2014.

18 Ákos Szalai: *Közgazdaságtani fogalmak és módszerek jogászoknak*. Budapest, Pázmány Press, 2020. 117-118. https://mek.oszk.hu/21800/21884/21884.pdf (13.03.2024).

19 Bostrom 2014, 188.

On the legal side, questions will arise as to whether the project owner has taken all reasonable measures to mitigate or manage the known risks.

Bostrom breaks down the above agent problem in the development of artificial intelligence into two cases. One is the first problem mentioned above, where the agent and the developer are both humans. Then, in his view, the problem mainly arises in the development phase. The author also predicts that in this case the usual management techniques can be applied to deal with the problem. It is pointed out, however, that the characteristics of artificial intelligence should be considered to a greater extent in the specific development methodology. Without going into a detailed analysis of the techniques, as this is beyond the scope of this research, we believe that the application of management techniques alone is not a sufficient guarantee.

In the other case, the principal and the agent are the superintelligence. He believes that this is a problem at the operational stage. To solve this problem, new techniques are needed.

The paper analyses the problems of managing artificial intelligence as follows.

He sees the methods for managing the potential explosion as falling into two broad categories. On the one hand, we can talk about the control of capabilities, and on the other hand, the selection of motivations. With the former, we can place limits on the scope of the AI, and with the latter, we can control what it strives to do. The first of the capability control methods to be mentioned is the box method, which is reminiscent of the 'sandbox" methods that are prevalent in AI research today.[20] In this method, we distinguish between physical and informational restriction methods. And the essence is nothing other than to lock in artificial intelligence. In this situation, an attempt is made to prevent the AI from interacting with the outside world outside the channels provided by the researchers. The method of restricting information from the outside works by trying to control what information can enter the box.

Among the methods used to regulate ability are restraint, incentives and traps. Traps, as a mechanism, work by having a mechanism independent of the AI run diagnostic tests on the system itself and stop it if it detects dangerous threat signals. This method may be suitable for use as a temporary safeguard during the development phase.

Among the methods of motivation selection, Bostrom mentions direct specification first. This method brings us to the main problem for lawyers. How do we regulate artificial intelligence systems? Motivation selection methods seek to shape the will of the superintelligence. In this way, we might be able to

[20] AI Sandbox, https://huit.harvard.edu/ai-sandbox 20.03.2024).

prevent unwanted outcomes. We then seek to control the system's motivations and ultimate goals. The direct specification tries to define the artificial intelligence in a rule-based and consequence-based way using some rules or values. In the book, we find examples of how machines can be induced to follow Asimov's laws or to obey the rules of different countries' legal systems. The main problem is that in all cases the rules must be precise, applicable in all situations and translatable into machine language. It is also difficult to determine what to assign value to, or even how to define a concept.

To add to Bostrom's thoughts, we would like to nuance the problem and draw attention to the complexity of the issue in the following. We also run into a hurdle in defining the principles used to regulate artificial intelligence when we consider that each statement is a matter of relativity. In fact, if we want to specify that the AI system should take care of environmental sustainability and, in this context, water quality adequacy, that it should focus on adequate water quality, and we plant this as a kind of principle in its codes, we are faced with the following problem. When considering the derived value of 'water quality' for environmental sustainability, we find two contradictory criteria. Here, the "quality of drinking water" and the "quality of the food chain of fish populations" are in conflict. The comparison is based on whether we are part of the population using the lake as a drinking water reservoir or whether we are anglers or conservationists, for whom the latter factor is more important. The right 'phosphate level' is cardinal, as lower phosphate levels are better for drinking water quality but worse for fish populations."[21] Therefore, to make a proper assessment, we need to determine the case-by-case order of the values.

Bostrom's book concludes with the question of what we should do to properly manage the explosion of super intelligence. As well as drawing attention to the need to assess our strategic situation and build capacity, he points to the need to take specific measures. He mentions developments in the field of technical problems of machine intelligence security as just such a specific measure. It also has a specific objective to help spread "good practices" among AI researchers. He believes that any progress on the problem of governance should be communicated to all researchers.

A series of good practice documents have been published around the world in recent years. Various organisations have set out ethical and moral lines and frameworks[22] for the safe and trustworthy use of AI in specific industry sectors.

[21] From Principles to Practice, An interdisciplinary framework, https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf 17. (26.11.2022).

[22] UNESCO Recommendations on the Ethics of AI, November 2021. or UNICEF Policy Guidance on AI for Children, November 2021.

One such guideline will be described in detail below, followed by a document containing similar good practices and guidelines.

5. Guidelines for secure AI system development to ensure the secure artificial intelligence development

The timeliness of regulating artificial intelligence (AI) is beyond debate. In addition to the opportunities offered by new technologies, there is now a detailed mapping of the risks surrounding AI systems.

Highly advanced AI is predicted by prominent figures from different disciplines as a technology with gigantic risks.[23] Risk factors are linked to a broad spectrum of fundamental human rights, highlighting the potential dangers of using AI.[24]

On 11/27/2023, artificial intelligence regulation reached another milestone. On this day, the National Security Agency (NSA), UK National Cyber Security Centre (NCSC-UK), U.S Cybersecurity and Infrastructure Security Agency (CISA) and other partners released their global guidelines, which are guidelines for secure AI system development to ensure secure artificial intelligence development.

23 partners from 18 countries have joined the document and agencies from all corners of the globe have contributed to the document from Chile to France and Japan. This reflects the cross-national challenges of AI systems from a cybersecurity perspective.

Of course, this document is not without precedent, as the National Cyber Security Centre previously published another document[25] in August 2022, entitled "Principles for the security of machine learning", which also addressed fundamental principles to address and prevent the additional risks inherent in machine learning systems.

As digitalisation becomes more and more prevalent in our lives and we spend more and more time online, the security of these technologies becomes crucial. Besides expecting IT hardware to be secure, it is equally important that software provides the right level of security. Public administrations, among others, are

[23] Around April 2023, it was reported in various media that decisive people such as Elon Musk, Yuval Noah Harari and Steve Wozniak, among others, are calling for an immediate halt to the development of certain types of artificial intelligence systems. Die Rückkehr des Wunderglaubens, https://www.spiegel.de/wissenschaft/kuenstliche-intelligenz-die-rueckkehr-des-wunderglaubens-kolumne-a-d53eb350-b5b5-4888-9bf8-8fc510d018b8 (15.04.2023).

[24] Sándor Udvary: Az önvezető gépjárművek egyes felelősségi kérdései. *Pro Publico Bono – Magyar Közigazgatás,* 2/2019, 146–155.

[25] Principles for the security of machine learning, https://www.ncsc.gov.uk/files/Principles-for-the-security-of-machine-learning.pdf (10.09.2023).

also trying to get on board this digitalisation trend and many services are now available either optionally or exclusively in digital form. There is no doubt that serious problems can arise when some public services become dysfunctional, for example when they stop working, as a result of cyber-attacks. Failures in software and hardware products increase the attack surface on which cybercriminals can cause damage.

The importance of these guidelines is also underlined by the fact that several non-governmental organisations have contributed to their development. It is encouraging to note that the list includes several major global IT companies such as Google, Microsoft, Amazon, IBM, etc., which are playing a key role in the digital development process.

Tech giants such as Microsoft, which are pioneers in the development of various AI systems at many points – think of their partnership with OpenAI – naturally have their own AI security protocols. [26] In the document "AI security risk assessment framework", published on 9 December 2021, they explicitly address machine learning security assessment, within which they specifically address the secure storage, access and integrity of the data used, the types of sensitive data and they elaborate on the security criteria for models. It includes professional guidelines such as that the source of the data collected should be verified before use, the source should be stored with the data and documented. In addition to these, it is also more specific in its cultivation and includes criteria specifically for model teaching and development.[27]

Looking more closely at the NCSC published guidelines document we discuss, we see that, as in the Microsoft document, artificial intelligence is specifically defined as machine learning applications, and within that, all types of machine learning AI are included. In line with the technological developments and trends of the time, AI systems are typically based on machine learning models.

To be clear from an application perspective, the document also provides a definition of machine learning applications. "MI applications are applications that:

- involve software components (models) that allow computers to recognise and bring context to patterns in data without the rules having to be explicitly programmed by a human
- generate predictions, recommendations, or decisions based on statistical reasoning."

[26] Best practices for AI security risk management, https://www.microsoft.com/en-us/security/blog/2021/12/09/best-practices-for-ai-security-risk-management/ (10.09.2023).

[27] Microsoft Security AI Security Risk Assessment, Best practices and guidance to secure AI systems,https://github.com/Azure/AI-Security-Risk-Assessment/blob/main/AI_Risk_Assessment_v4.1.4.pdf (10.09.2023).

As to the reasons for the creation of this document, the document declares the following. Just like in other areas of our lives, new tools and new techniques in programme development provide opportunities for new abuses. Before the advent of car use, it was not natural for an accident to be caused by deliberately damaging parts of a car. Artificial intelligence systems contain new vulnerabilities that can be exploited by prepared malicious actors, both on the hardware and software side. The paper draws attention to this by stressing that attackers can induce unintended behaviour in machine learning systems by using so-called adversarial machine learning, leading to the problem repeatedly mentioned by lawyers that the output of the system cannot be predicted. In the case of machine learning AI systems, this "black box effect" is present anyway, in the case of such cyber-attacks this unintended behaviour is deliberately induced.

There can also be cases where users are allowed to perform unauthorised operations, or data poisoning, where the training data as a data domain is corrupted.

The structure of the document follows the 4 phases in the lifecycle of AI systems development, namely secure design, secure development, secure deployment, and secure operation and maintenance.

This life cycle includes the requirement that once the software containing the AI systems is created, it is monitored during its use, with each update meeting cybersecurity criteria.

In fact, in each chapter, the document suggests considerations and measures that will help reduce the overall risk of the organisational AI system development process.

One such suggestion is to consider the security benefits and trade-offs of each model when selecting an AI model at the design stage of development.

By integrating well-established principles like "security-by-design" and "security-by-default", the publication outlines the existing vulnerabilities specific to AI and suggests ways to consider them during the development process. Typically, end users lack the understanding to grasp the risks associated with AI. Additionally, cybersecurity authorities emphasize the importance for AI system operators to educate users about potential risks and provide guidance on the secure utilization of these systems.[28]

While the document certainly describes the guidelines in sufficient detail for its intended purpose, the agencies emphasize that the measures and adherence to such guidelines are not a substitute for the development of a proper cybersecurity practice and risk management program or protocol. Such research itself should

[28] Internationale Cybersicherheitsbehörden veröffentlichen Leitfaden zur Entwicklung sicherer KI-Systeme, https://www.bsi.bund.de/DE/Service-Navi/Presse/Pressemitteilungen/Presse2023/231127_Leitfaden-sicher-KI-Systeme.html (13.01.2024).

be used in conjunction with established cybersecurity risk management and incident response best practices. The guidelines set out in this document are closely aligned with the good practices for software development lifecycle practices that have already been identified in subsequent documents:

- the NCSC's Secure development and deployment guidance
- the National Institute of Standards and Technology (NIST) Secure Software Development Framework (SSDF)6

The document is not binding legislation that would impose a strict obligation on companies developing AI systems, and thus cannot be used to enforce its provisions. However, it is noted that the use of new technologies and the success of AI systems are based on trust and confidence in them, which cannot be achieved by legislation alone. This document can be successful and can be considered a milestone because of the significant international partnership and contributors.

Increasing user awareness creates the need for the system to be used to comply with cybersecurity recommendations. This kind of user confidence can be achieved by applying a standard, by obtaining a certificate, with which an operator can not only be successful, but also help build a more secure digital future for its users.

6. Blueprint for an AI Bill of Rights

In late 2022, the White House proposed a Blueprint for an AI Bill of Rights. "The Blueprint for an AI Bill of Rights is a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence."[29] The five principles are the followings: safe and effective systems; algorithmic discrimination protections; data privacy; notice and explanation; human alternatives, consideration, and fallback.

The document provides concrete guidance on the principles to be applied to address the identified risks, which can provide appropriate safeguards to ensure that the design, development and operation of AI systems do not cause any harm. The document stresses that it has been drawn up following appropriate public consultation and that the conclusions drawn therefrom are included.

[29] Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf (22.12.2023).

This framework provides a national values statement and toolkit that is sector-agnostic to inform building these protections into policy, practice, or the technological design process.

The document also integrates itself into the legal order by expressing that "where existing law or policy – such as sector-specific privacy laws and oversight requirements – do not already provide guidance, the Blueprint for an AI Bill of Rights should be used to inform policy decisions."

What makes the document unique is that, in addition to the general principles, it provides the reader with concrete guidance and a toolbox by answering the following questions for each principle. Why is this principle important? What should be expected from automated systems? How can these principles move into practice? These are the questions that market players who want to prepare for the use of AI are also asking about the practical application of AI and risk management. However, let's see how far this document really succeeds in answering these questions by means of a concrete example.

In our opinion one of the most interesting principles is the "human alternatives, consideration, and fallback".

We find this important and interesting because the system is designed to bypass humans to perform a task faster and more efficiently. On the other hand, it is also linked to the realisation that we know the exponential nature of artificial intelligence and, as mentioned earlier, its ability to cause enormous damage in a very short time. For this reason, the document stresses that "you should be able to opt out from automated systems in favour of a human alternative, where appropriate."

In response to the question why this principle is important, the document details the following:

"No matter how rigorously an automated system is tested, there will always be situations for which the system fails."

This principle is essentially nothing more than a so-called "kill switch"[30], which serves the purpose in the IT sector of switching off a program or device if it starts to malfunction. This is not only necessary when using artificial intelligence systems, it should be part of any program that may malfunction. However, this technique is, by definition, a system in operation and not a preventive technique. It makes sense to talk about guarantees built in and applied during the development of the programme and guarantees during the operation of the programme, both for AI systems and for other IT solutions. The principle under analysis here

[30] Will There Be A 'Kill Switch' For AI?, https://www.forbes.com/sites/cognitiveworld/2020/03/05/will-there-be-a-kill-switch-for-ai/ (12.01.2024.).

focuses on the monitoring of the program once it is running. The amount of damage that occurs in this case depends on when the human supervisor detects the error and the time that elapses between the detection of the error and the pressing of the *off* button.

It also uses negative examples to illustrate the damage that can occur if the principle is ignored. In a Colorado unemployment benefit scheme, claimants were required to have a smartphone to prove their identity. Understandably, those who did not have a mobile phone could not identify themselves due to a lack of human means.

After identifying, why the principle is important for the community, the document also gives examples of how the principle can be put into practice. Examples include systems to help employees choose the right health insurance for their needs in the marketplace, and customer service systems to help answer common problems and questions. Perhaps a shortcoming of the document is that these examples are not very detailed and numerous. Nevertheless, they are properly referenced so that specific cases that have occurred can be traced.

7. Other examples of good practice

While the Blueprint may be a milestone in terms of good practice, we nevertheless believe that there are more sophisticated and useful documents for users who are new to AI.

Such documents have been produced under the auspices of the OECD. One of these is "The state of implementation of the OECD AI principles four years on"[31] which shares with the reader in a detailed way the practices of implementing the principles promoted by the OECD. It gives the quality seal created by the German AI Association to promote the use of human-centred and human-serving AI as an example. This seal identifies a common set of values and validation processes to express the ethical compatibility of products. The key criteria are ethics, impartiality, transparency, security and privacy.

31 The state of implementation of the OECD AI Principles four years on, https://www.oecd-ilibrary.org/docserver/835641c9-en.pdf?expires=1712094731&id=id&accname=guest&checksum=65B325A7C953BC3F 7BE8B89128BE9F6E (12.01.2024.).

Documents[32] and webpages[33] published by Ernst & Young Global Limited showcase the potential of AI for business through real-world, authentic examples.

The document "The Artificial Intelligence (AI) global regulatory landscape" has the great advantage of outlining regulatory trends, but also provides recommendations on what steps individual companies and policy vendors can take to ensure the safe use of AI.

We find examples for leading practices to create a trusted AI ecosystem. It is necessary to have Al ethical design policies and standards for the development of Al, including an Al ethical code of conduct and Al design principles. The Al ethical design standards should define and govern the Al governance and accountability mechanisms to safeguard users, follow social norms and comply with laws and regulations.

There is a need for related strategies, so that artificial intelligence and its control develop in the right direction.

Regulating artificial intelligence is very important, according to which various non-European countries included their artificial intelligence strategies, which they formulated, starting in 2017. These strategies outline the main regulatory directions and ethical directives around which the regulation is intended to be built. Without analyzing these strategies in more detail, we note that, for example, in March 2017 Canada published the Pan-Canadian Artificial Intelligence Strategy. South Korea also published its Artificial Intelligence Strategy for Innovative Growth in December 2018.

Country	**Name of the document**	**Date of issue**
United States:	Executive Order on Maintaining American Leadership in Artificial Intelligence	February 2019
Canada	Pan-Canadian Artificial Intelligence Strategy	March 2017
China	New Generation Artificial Intelligence Development Plan	July 2017
South Korea	Artificial Intelligence Strategy for Innovative Growth	December 2018
United Arab Emirates (UAE)	National AI Strategy	October 2017,
India	National Strategy for Artificial Intelligence	June 2018
Russia	National Strategy for AI	October 2019
Saudi Arabia	SDAIA Strategy	August 2019

[32] The Artificial Intelligence (AI) global regulatory landscape, https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/ai/ey-the-artificial-intelligence-ai-global-regulatory-landscape.pdf (20.02.2024).

[33] AI Use cases, https://www.ey.com/en_gl/services/ai/use-cases#tabs-ca1ee0a390-item-1c236c7145-tab (28.01.2024).

8. Conclusion

Summarizing the above, we can see that there is no lack of theoretical foundations, ethical guidelines and researched philosophical arguments regarding the development of artificial intelligence. Research on the regulation of artificial intelligence is a very popular field in twenty-first century law and engineering. The question arises, then, what is the obstacle to the regulation of AI and the kind of positive intelligence explosion that Nick Bostrom has predicted.

In our view, the common intelligence, the kind of subjective risk measurement discussed in the previous chapters and the development of all these are the goal to put this technological progress into operation.

We have seen examples of good practices that can help companies to implement AI applications safely. Some of the papers point to the important role that standards will play in promoting good technical documentation. The relevance and usefulness of good practices will be seen over time, and time-proven good practices will certainly contribute to the low-risk operation of AI systems. As pointed out in the Ernst & Young paper cited above in relation to existing regulatory trends, regulators are seeking to link AI guarantees with other areas such as data protection. In this respect, we believe that it will be important in the future not only to understand how AI works, but also to successfully link the regulation of AI with existing regulation and different regulatory areas. In this context, it is essential to understand the nature of artificial intelligence and the risks of artificial intelligence systems. It is important to stress that this task requires staff with multidisciplinary knowledge, as is the case in this area.

We think it is important to underline our view that, although good practices in a market context can certainly prove useful, as the consumer will choose a higher quality and safer product, the truly reassuring thing would be the regulation of artificial intelligence that can be embedded and enforced in the relevant legal system. That is why we welcome the efforts of the European legislator in this direction.

Bibliography

Ádám Havas: Kockázatelemzés-mágia vagy tudomány? *Iskolakultúra*, 23/1993, 21–28.

András Tóth: A mesterséges intelligencia szabályozásának paradoxonja és egyes jogi vonatkozásainak alapvető kérdései. *Infokkumunikáció és jog*, 73/2019, 3–9.

Artificial Intelligence Act: MEPs adopt landmark law, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law (downloaded: 16.03.2024)

Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf (22.12.2023)

Dóra Peszlen: Apollónios Rhodios. Argonautika 3. *Studia Litteraria*, 1-4/2017, 37–41.

Fedja Netjasov – Janic Milan: A review of research on risk and safety modelling in civil aviation. *Journal of Air Transport Management*, 4/2008, 213–220.

IBM Global AI Adoption Index 2022, https://www.ibm.com/watson/resources/ai-adoption (downloaded: 10.02.2024)

József Kindler: *Általános kockázatelmélet és -módszertan. Egyetemi jegyzet*. 1983.

Nick Bostrom: *Superintelligence. Paths, Dangers, Strategies*. Oxford University Press, 2014.

Ákos Szalai: *Közgazdaságtani fogalmak és módszerek jogászoknak*. Budapest, Pázmány Press, 2020. https://mek.oszk.hu/21800/21884/21884.pdf (13.03.2024)

Tamás Fleischer: Innováció, növekedés, kockázat. In: Miklós Bulla – Pál Tamás (eds.): *Fenntartható fejlődés Magyarországon. Jövőképek és forgatókönyvek.* Budapest, Új Mandátum Könyvkiadó, 2006. 275-284. http://real.mtak.hu/3973/1/fleischer_innovacio-novekedes-kockazat_fefemao06.pdf (01.03.2021)