ETHICAL GUIDELINES

FOR DEVELOPMENT, IMPLEMENTATION AND USE OF ROBUST AND ACCOUNTABLE ARTIFICIAL INTELLIGENCE

Belgrade, February 2023

Contents

1. INTRODUCTION

1.1 The reason for adopting the Guidelines

The purpose of the *Ethical Guidelines for Development, Implementation and Use of Robust and Accountable Artificial Intelligence* (hereinafter: the Guidelines) is to enable science to evolve and progress, especially in the field of artificial intelligence, but at the same time not to allow human beings to become threatened or neglected as the centrepiece of all processes that affect them and in which they are a direct or indirect factor. Moreover, the artificial intelligence systems that are developed must be in harmony with the wellbeing of humans, animals and the environment.

Artificial intelligence is one of the pillars of the Fourth Industrial Revolution. As a branch of computer science and engineering, the development of artificial intelligence began several decades ago, with periods of advancement and stagnation. In recent years, thanks to breakthroughs in the field of deep neural networks, the increasing amount of available data suitable for machine learning, and the growing availability of microprocessors capable of large-scale numerical computations, we are witnessing a sudden development and spread of the use of artificial intelligence in areas such as healthcare, finance, education, energy, natural language processing, speech technologies, computer vision, etc.

The development of artificial intelligence (systems) will aim to create solutions that comply with the relevant standards throughout their lifecycle, and on the basis of which it will be possible to characterise these systems as reliable and accountable. In general, a robust and accountable artificial intelligence is one that is technically robust and safe, in compliance with the law, and in compliance with the adopted ethical principles and values. Each of the three components mentioned above is considered separately; the requirements and the results of the assessment of one component do not depend on the requirements and the results of the assessment of another component. The above components must be in harmony with each other, and only if all three are met can a given artificial intelligence be assessed as robust and accountable.

The main purpose of adopting these Guidelines is to prevent processes involving artificial intelligence systems from endangering or marginalising humans and human agency, and to prevent the infringement of the freedom of action, thought and decision-making to such an extent that the rights and frameworks protecting these values become meaningless, trivialised or forgotten. This is primarily about creating ecosystems that will use artificial intelligence to increase human productivity, optimise the use of resources for work and the functioning of society in general, and improve the quality of human life.

1.2 The basis for adoption

The basis for the adoption of the Guidelines is primarily the *Strategy of Development of Artificial Intelligence for the Republic of Serbia* for the period 2020–2025[1], which identified the ethical and safe use of artificial intelligence as one of its five goals. This activity is included in the Action Plan for the period 2020-2022. [2]

[1] *Official Gazette of the Republic of Serbia,* No 96/2019.

[2] *Official Gazette of the Republic of Serbia,* No 81/2020.

To achieve this goal, mechanisms must be developed and put in place to ensure the responsible development of artificial intelligence and verify that these systems meet the highest ethical and security standards. These Guidelines define such standards and ways to verify their implementation in the development and use of artificial intelligence systems.

In November 2021, UNESCO adopted the *Recommendation on the Ethics of AI*[3]. The representatives of the Republic of Serbia participated in the drafting of this document.[4] The principles set out in the Recommendation are reflected in these Guidelines.

In accordance with Article 72 of the *Stabilisation and Association Agreement* between the European Communities and their Member States, of the one part, and the Republic of Serbia, of the other part (hereinafter: the Stabilisation Agreement), the Republic of Serbia has undertaken to gradually harmonise its existing and future legislation with the EU *acquis communautaire*, which also includes the Community legislative acts. This commitment has also been reiterated in the provisions of the Constitution of the Republic of Serbia, in particular Article 194.

In April 2021 the European Commission submitted to the European Union a proposed regulatory framework for artificial intelligence – the proposed Regulation of the European Parliament and the Council laying down harmonised rules on artificial intelligence[5] and amending certain Union legislative acts (hereinafter: proposed EU Regulation on AI). With this act, the European Union aims to establish a legal framework for the development and use of artificial intelligence, to facilitate and increase investments and innovations in this area, and to create a single market for the legal, safe and robust use of artificial intelligence systems.

Previously, the Council of Europe Commissioner for Human Rights has issued the *Recommendations on Artificial Intelligence and Human Rights*, which consist of 10 steps. These Recommendations build on the Council of Europe's previous work in this area, in particular the *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems*, the *Guidelines on Artificial Intelligence and Data Protection*, the *Declaration by the Committee of Ministers on the Manipulative Capabilities of Algorithmic Processes*, the *Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications*, and the *Report of the UN Special Rapporteur on Freedom of Opinion and Expression*, which discusses the human rights implications of artificial intelligence technologies in the information society.

In order to ensure the gradual harmonization of the Serbian legal framework with the EU *acquis communautaire* and to establish a legal framework for the development and use of ethical artificial intelligence systems in the Republic of Serbia, the Government has issued a decree adopting these Guidelines. It is recommended that the Guidelines be applied by all state authorities and organisations, authorities and organisations of the Autonomous Province, authorities and organisations of local self-governments, public enterprises, special regulatory bodies, as well as legal entities and natural persons performing public functions, when they establish artificial intelligence systems and use them in their work.

[3] UNESCO, (2021), *Recommendation on the Ethics of AI*, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi

[4] UNESCO, Artificial Intelligence, available at: https://en.unesco.org/artificial-intelligence/ethics

[5] European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

The Guidelines aim to cover the widest possible range of actors in the artificial intelligence ecosystem, in order to establish a horizontal approach to the application of rules. The Guidelines include the following actors:

- persons working on the development and/or implementation of artificial intelligence systems;
- persons who use artificial intelligence systems primarily for their work, which includes interaction with other persons (e.g. market participants);
- persons using artificial intelligence systems who are:
    - directly affected by the systems (e.g. they use systems to access public services)
    - indirectly affected by the systems (e.g. they are part of a rare disease research group where medical data is processed as part of the Republic of Serbia's strategy to improve public health);
- the general public in the broadest sense.

The Guidelines do not address the ownership structure, the contractual and legal obligations, and other legal issues related to the specific work products and research in the artificial intelligence ecosystem.

1.3 The implementation of the Guidelines

The implementation of the Guidelines consists of: 1) organising public consultations and training to familiarise the public, including professionals, with the concept and importance of artificial intelligence, 2) monitoring and evaluation of the implementation of the Guidelines in the public and private sectors.

The competent minister shall lay down detailed rules on how the Guidelines are to be integrated and implemented.

1.4 The legal framework

Given the importance of the development and use of artificial intelligence in the Republic of Serbia, the need to regulate this area by law is recognised.

2. GLOSSARY OF TERMS AND DEFINITIONS

2.1 Ethics

Ethics is the science of morality. It studies the purpose and goals of moral norms, the main criteria for moral evaluation, and the foundations and sources of morality in general. Ethics studies human behaviour that is considered acceptable and moral from certain points of view, where this behaviour affects other humans, animals that may feel pain, suffering, fear and stress, and ecosystems.

Key ethical concepts include: morality, good, evil, conscience, freedom, happiness, love, virtue.

2.2 Artificial intelligence system

Artificial intelligence system or artificial intelligence is a term that is defined in many different ways. Most definitions take it to mean software (i.e. software model) that is trained on a dataset to perform specific tasks (e.g. recognising certain patterns and similar). The defined terms have the following meanings in these Guidelines:

The Independent Expert Group of the European Commission proposed the following definition: "**Artificial intelligence** refers to systems that display intelligent behaviour by analysing their environment and taking actions – with a certain degree of autonomy – to achieve specific goals. AI systems can be purely software-based, acting in the virtual world, (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones and the like)." [6] These systems include the machine learning systems and expert systems. In these Guidelines, "artificial intelligence system", "System", "system", "artificial intelligence" or just "AI" shall have the same meaning.

Bias describes systemic and repeated errors in the computer system that lead to unfair results, such as favouring one category over another in a way that differs from the intended algorithm functions.

Autonomous system is a System that acts or performs tasks with a high degree of autonomy, i.e. without outside influence.

Security – the fundamental concepts of security of information systems are based on information security, which consists of three characteristics: confidentiality, integrity and availability, also known as the "triangle", the "trinity" or the "triad" of security.[7] When protecting information and data, it is necessary to allow controlled access to those who are authorised to see it, which ensures confidentiality. Integrity means that the data has the intended meaning and has not been unintentionally or intentionally altered. Availability means that anyone who is authorised to do so can access and modify information and data within a specified time frame.

System architect is the person who designs the System.

System architecture is a set of activities that design the System; System architects are persons who design the System.

System exploitation is the use of the System by authorised users.

System application is the creation of a suitable System that meets the functional and project specifications.

Persons responsible for monitoring are the persons authorised to check if the System is used correctly.

Multi-cloud means the use of several public clouds.

System reliability is the probability, with a certain level of confidence, that a system will perform its intended purpose correctly, within the specified performance limits, for a specified period of time, without failure, when used in the correct manner and for the intended purpose, under the specified loading parameters, and taking into account its previous time of use.

Poka-yoke means avoiding unintended errors, "mistake-proofing".

[6] A definition of AI: Main capabilities and scientific disciplines, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, 2018.

[7] See Bourgeois, D., (2014), *Information Systems for Business and Beyond*, Saylor Foundation, pp.64-65.

Risk is the state of the System caused by the lack of proper safeguards, and is a source of risky events that can lead to quality changes and losses in the System.

Self-learning system is a System that autonomously recognises patterns in the data it has been trained to use, without the need for monitoring.

System testing is a phase in the development of a System in which its correctness and reliability are tested; the phase in which all errors in the System are identified and corrected.

Training data is data used to train the System.

Input data are variables, i.e. parameters that provide results (output data) after they have been processed by the System.

Human-in-the-loop means that human intervention in the System is possible at all stages of decision-making.

Human-of-the-loop means that intervention is possible in the development phase and during monitoring of the System's work.

Human-in-command means that it is possible to control the work and all the activities of the System, including its wider economic, social, legal and ethical impacts. Decisions about when and how to use the System are also controlled, including decisions in which situations the System will not be used.

2.3 High-risk artificial intelligence systems

A high-risk system is a System that has the tendency to directly or indirectly violate the principles and conditions set out in these Guidelines, without necessarily doing so.

For the purposes of these Guidelines, high-risk systems are not considered undesirable. However, due to the importance of the areas of life in which they are applied and the extent of their potential impact on humans and their integrity, they need to be specifically analysed and their impact needs to be assessed.

For the purposes of these Guidelines, a high-risk system is a system which:

- is part of the safety (security) system of a product or is itself a product that has the function of or operates as a safety (security) system and as such requires a third-party assessment of compliance with the legal standards for the use of artificial intelligence systems;
- is listed and designated as a high-risk system in the Guidelines (hereinafter: the list of high-risk systems).

For classification as a high-risk system, it is not relevant whether the System is in use (it is sufficient that it was created as such), nor is it relevant whether it constitutes a stand-alone product and/or service, or is an integral part of another product/service.

The Guidelines do not apply to Systems that have been prohibited by laws regulating artificial intelligence systems.

High-risk systems are artificial intelligence systems in the following areas:

- biometric identification and classification of persons: this includes, in particular, systems for real-time remote biometric identification of persons and systems for post remote biometric identification of persons;
- management of critical infrastructures and their operation: this includes, in particular, systems for the management of roads and transport, water supply, gas supply, heating and electricity supply, or security systems of the above-mentioned systems which are an integral part of these systems;
- education, vocational training and qualification: this includes, in particular, systems for determining whether a person has access to a particular educational or training institution, or systems for distributing persons to such institutions, as well as systems for evaluation and scoring of persons attending such institutions, including systems for scoring of exams (entrance exams) required for enrolment in such institutions;
- employment, human resource management and access to self-employment: this includes, in particular, systems for recruiting and hiring people, including systems that advertise vacancies, screen, filter and evaluate candidates for a particular job (in the context of interviews or tests) and make the final decision on hiring; systems that make decisions on labour relations of workers (promotion, bonuses, dismissal, changes in job descriptions, specific tasks); and systems that carry out monitoring and performance appraisal of workers, on the basis of which employment-related decisions are made. [8]
- healthcare: this includes, in particular, systems that analyse genetic and health data;
- access to and use of public and social services and basic private services: this includes, in particular, systems that decide whether a person is entitled to public services and social benefits and that decide whether to authorise, reduce or terminate such benefits, as well as the conditions under which such decisions are made. This also applies to systems that determine the creditworthiness of individuals and their credit rating, except where such systems are used for personal and non-commercial purposes. Finally, this also applies to systems designed for use in emergency medical services or other emergency services (fire brigade, police and the like), provided that these systems determine priorities for the provision of such emergency assistance;
- law enforcement: this includes, in particular, systems for law enforcement agencies that perform risk assessment of individuals who may commit or repeat crimes based on their personal characteristics, attributes, or past criminal conduct; systems used as lie detectors or means of detecting the emotional state of persons; systems used to assess the truthfulness of evidence in pre-trial, trial and court proceedings; systems used to assess individuals; systems used to analyse criminal offences concerning individuals which enable the search of complex, interconnected or unconnected large datasets from different sources or in different formats to detect unknown patterns in data or hidden links between data;

[8] This provision is not limited to the recruitment of persons under an employment relationship, but also to other forms of employment or labour (service contracts, intermittent and temporary work agreements and the like, in accordance with Labour Law) where such employment includes the selection procedure and its maintenance (performance appraisal, remuneration, sanctions, termination, etc.)

- management of human migrations, asylum-seeking and surveillance of state borders: this includes, in particular, systems used as lie detectors or means of detecting the emotional state of persons; systems used by the competent authorities/institutions to assess the risks (including security risk, risk of illegal migration or health risk) posed by persons seeking to enter or who have entered the territory of the Republic of Serbia; systems used by the competent authorities to verify the authenticity of travel documents, in particular by checking the security aspects of such documents; systems to assist the competent authorities/institutions in the examination of applications for asylum, visas, residence and work permits and related procedures (this includes checking the criminal and misdemeanour record of the person and whether any complaints have been lodged against such applications and the nature of the complaints etc.) in order to decide on such applications;
- judiciary and democratic processes: this includes, in particular, systems to assist the judicial authorities in analysing and interpreting circumstances, facts and legal norms in order to apply the relevant legal norms to the specific sets of circumstances and facts.

The list of high-risk systems also includes AI-based recommendation systems on platforms used by a large number of people, such as various social networks. Based on various data and set goals, these systems decide which content to forward to an individual or group. For individuals, this can influence their attention, desires, thoughts, opinions, creativity, feelings, decisions and activities. In a society, this can lead to social bubbles and polarisation on important social issues, ultimately affecting the democratic capacity of the society. The list also includes personal data processing systems used by state authorities and organisations, provincial authorities and organisations, local self-government authorities and organisations, institutions, and public enterprises when making decisions related to their delegated tasks and mandate.

The list is not exhaustive. The development of artificial intelligence forces us to keep this list open and to understand it as an overview of representative examples of high-risk systems that can be added to or modified.[9]

[9] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.

2.4 Protection of personal data

Personal data means any data relating to a natural person whose identity is or can be identified, directly or indirectly, on the basis of identity characteristics such as name and identification number, location data, electronic communications network identifiers, or one or more characteristics of his or her physical, physiological, genetic, mental, economic, cultural and social identity.

Processing of specific personal data means the processing of data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, as well as the processing of genetic data, biometric data for identification purposes, data concerning health or data concerning a natural person's sex life or sexual orientation.

A **data processing impact assessment for the protection of personal data** must be carried out prior to processing where certain types of data processing, in particular those using new technologies, are likely to present a high risk to the rights and freedoms of natural persons, taking into account the nature, scope, circumstances and purposes of the data processing. A data processing impact assessment must be carried out in the following circumstances: 1) when systematic and comprehensive assessments of the condition and characteristics of a natural person are carried out by automated processing of personal data, including profiling, and are used to take decisions which have an impact on the legal position of the person or affect the person in a similarly significant way; 2) in the case of processing of specific personal data or personal data relating to criminal convictions and offences, on a large scale; and 3) in the case of systematic monitoring of publicly accessible areas, on a large scale. The list of types of data processing that require an impact assessment is contained in the Decision of the Commissioner for Information of Public Importance and Personal Data Protection.

The person responsible for the protection of personal data is the person who performs the relevant duties and tasks in accordance with the Law on Personal Data Protection. It is mandatory to designate such a person 1) if data processing is carried out by the public authorities, with the exception of data processing carried out by the court under its judicial mandate; 2) if the basic activities of the data holder consist of data processing which, by virtue of its nature, scope and purpose, requires regular and systematic monitoring by a large number of persons to whom such data relate; and 3) if the basic activities of the data holder consist of processing of specific personal data or personal data relating to criminal convictions and offences, on a large scale.

3. PRINCIPLES

Without prejudice to other concepts and principles, the following principles have been identified as crucial starting points for the development, implementation and use of artificial intelligence systems that deserve human trust because of their reliability and accountability:

3.1 Explainability and verifiability

One of the most important characteristics of the human mind is to perceive its surroundings and to look for answers and explanations as to why and how things are the way they are.

This characteristic has influenced human evolution and the development of science, and therefore artificial intelligence. The human need to understand and see things clearly is reflected in the principles of explainability and verifiability.

For the purposes of these Guidelines, explainability means that all processes - development, testing, commissioning in real environments, operation, monitoring and shutdown - must be transparent. The potential and purpose of the system must be explainable, while the decisions (recommendations) made by the AI system (to the extent that it is effective) must be clearly explained to anyone affected by the system (either directly or indirectly). If certain aspects of the system's work cannot be explained, the system must be called a "black box" model. [10]

Verifiability is a complementary element of this principle, ensuring that the system can be checked at any time, in any process and at any stage of its lifecycle. Verifiability includes measures and procedures for checking an AI system during testing and implementation, and for examining the short- and long-term impact of such a system.

3.2 Dignity

All members of society have the duty to mutually respect and protect the right to dignity as one of the fundamental and incontestable rights of every human being. Every human being has the right to the protection of his or her dignity. Violation or disregard of this right is punishable by law.

Human dignity (hereinafter: dignity) should be understood as the baseline principle of the protection of human integrity. With this in mind, those to whom these Guidelines apply should place human integrity at the centre at all times and regardless of the stage a particular AI solution is at (development, implementation or use). It is therefore necessary to develop systems that give priority to respect for human personality, freedom and autonomy at all stages.

Respect for human personality means creating systems that respect the cognitive, social and cultural characteristics of all individuals. Artificial intelligence systems that are being developed must be attuned to all these aspects. Therefore, it must be ensured that they do not subordinate humans to the functions of the system and that they do not endanger the dignity and integrity of humans.

To ensure respect for the principle of dignity, artificial intelligence systems must not be such that they grossly disregard the autonomy of human decision-making in their processes.

[10] There are various definitions of "black box" systems, but they all emphasise one aspect. These are artificial intelligence systems based on a model created directly from data using a developed algorithm. This means that the people who developed the system cannot understand how the model variables interact to make certain predictions, and the system itself does not show how these data/results were generated. Even if a person had a list of input variables, the black box models can be such complex combinations of variables that no human can understand how the variables are connected to make the final prediction. Some even interpret "black box" systems as models so complex that humans simply cannot interpret them.

The Constitution of the Republic of Serbia states that “human dignity is inviolable and everyone shall be obliged to respect and protect it. Everyone shall have the right to free development of his personality if this does not violate the rights of others guaranteed by the Constitution.” [11]

The Convention of Human Rights states the following: “Human dignity (dignity) is not only a fundamental human right, but also the basis of human rights. Human dignity is inherent in every person.”[12]

In the Republic of Serbia, the concept of dignity is regulated in the following ways:

- “The personal dignity (honour, reputation or reverence) of the persons to whom the information relates shall be protected by law.” [13]
- “Whoever ill-treats another or treats such person in humiliating and degrading manner, shall be punished with imprisonment up to one year.” [14]
- “Community service is any socially beneficial work that does not offend human dignity and is not performed for profit.” [15]

This principle emphasises that the integrity and dignity of all persons who may be affected by the artificial intelligence system must be protected at all times. However, dignity is a general concept that can be defined differently in life as well as in law, even if its essence remains the same. It is therefore appropriate to link the concept of dignity to three key concepts: honour, reputation and reverence.

3.3 Prohibition to cause damage

An artificial intelligence system must meet safety standards and have adequate mechanisms to prevent damage to persons and their property. Should damage nevertheless occur, it must be repaired as quickly as possible and the damaged party must be compensated in a manner prescribed by law.

The Law on Contracts and Torts defines damage as the reduction of a person’s property (regular damage) and the prevention of its growth (lost benefits), as well as the causing of physical or mental pain or distress to another person (non-material damage). [16] The Law also states that every person is obliged to refrain from actions that may cause damage to another. [17]

In addition to civil law liability, the law also recognises the criminal and misdemeanour liability of both natural persons and legal entities for damage they cause to another.

The Criminal Code[18] defines a large number of offences. Worth mentioning are the offences against the life and body of persons, against property, and against the freedoms and rights of man and citizen. A separate law also regulates the liability of persons for damage caused by the commission of acts of lesser social danger – misdemeanours. [19]

[11] Constitution of the Republic of Serbia, *Official Gazette of the Republic of Serbia, No* 98/2006 and 115/2021.

[12] Convention of Human Rights

[13] Law on Public Information and Media, *Official Gazette of the Republic of Serbia, No* 83/2014… author’s interpretation - 12/2016.

[14] Criminal Code, *Official Gazette of the Republic of Serbia, No* 85/2005, 88/2005 – correction and 35/2019.

[15] Ibidem.

[16] Law on Contracts and Torts, *Official Gazette of the SFRY, No* 29/78, 39/85, 45/89 – decision of the Constitutional Court …. and No. 18/2020.

[17] Ibidem.

[18] Criminal Code, *Official Gazette of the Republic of Serbia, No* 85/2005, 88/2005 – correction 121/2012,… and 35/2019.

Particular attention should be paid to the protection of vulnerable groups (e.g. older people, persons with disabilities, children, pregnant women, etc.) and to groups that are in an underprivileged position due to certain factors (e.g. employees versus employers, consumers versus businesses, etc.).

Artificial intelligence systems must be used in a secure manner and they must be safe and reliable. Their use for malicious purposes must be prevented.

3.4 Fairness

The principle of fairness refers to the protection of rights and integrity against discrimination, especially when it comes to discrimination against vulnerable groups (e.g. persons with disabilities). The concept of fairness is multi-faceted and has different meanings in different areas of life. In healthcare[20], for example, the principle of fairness means the prohibition of discrimination in the provision of healthcare services on the basis of race, sex, gender, sexual orientation and gender identity, age, ethnicity, social origin, religion, political or other beliefs, property status, culture, language, health condition, type of illness, mental or physical disability or other personal characteristics that may give rise to discrimination.

Therefore, artificial intelligence systems must prevent discrimination when they are used.

The principle of fairness has a substantive and a procedural dimension. The substantive dimension includes the protection of individuals or groups from unfair prejudice, discrimination and stigmatisation. Artificial intelligence systems should provide equal opportunities for all – firstly in terms of access to education, goods, services and technologies, and secondly in terms of preventing persons who use AI systems from being misled when making decisions. The procedural dimension of fairness involves the ability to challenge decisions made by AI systems and humans who manage or are responsible for these systems, and to obtain effective legal protection. To meet this requirement, competences and responsibilities must be clearly defined and the decision-making process must be explained in a clear and transparent manner.

This reduces the possibility of incorrect or incomplete understanding of the purpose and objectives of using these systems, which could potentially limit the freedom of choice in selecting which system to use. The fair use of artificial intelligence systems can lead to greater fairness in society in general and reduce the differences between people in terms of their social, economic and educational status.

[19] Law on Misdemeanours, *Official Gazette of the Republic of Serbia, No* 65/2013… 91/2019 and other laws.

[20] Law on Healthcare, *Official Gazette of the Republic of Serbia, No* 25/2019-40.

4. REQUIREMENTS FOR ROBUST AND ACCOUNTABLE ARTIFICIAL INTELLIGENCE

Building and developing robust and accountable artificial intelligence requires meeting certain requirements[21] based on the principles set out in these Guidelines. The Requirements are:

1. Human agency (mediation, control, participation) and control;
2. Technical reliability and security;
3. Privacy, personal data protection and data management;
4. Transparency;
5. Diversity, non-discrimination and equality;
6. Social and environmental wellbeing;
7. Accountability.

The requirements consist of verifiable parameters, i.e. technical and non-technical methods used to confirm and demonstrate compliance with the principles.

Technical methods aim to guide the development, implementation and use of artificial intelligence systems in such a way that artificial intelligence systems behave reliably and minimise possible intended or unintended damage to humans and society in general. Technical methods are provided in the form of recommendations.

Non-technical methods are used to examine organisational and other non-technical elements important to the development and use of artificial intelligence systems. These methods are provided in the form of a Questionnaire designed to assess whether individual artificial intelligence systems meet the basic principles and requirements contained in the Guidelines. The purpose of the Questionnaire is to assess the robustness and accountability of artificial intelligence systems from the perspective of ethical standards.

The Questionnaire assists individuals and organisations that develop, market, procure, implement and/or use artificial intelligence systems to assess whether the systems meet the requirements set out in the Guidelines. The Questionnaire can be used in all social and economic spheres, and it provides the minimal horizontal framework for the establishment of safe artificial intelligence systems in the Republic of Serbia. The Questionnaire can be adapted to specific areas and sectors. To learn more about the requirements that a System must meet in order to be assessed as robust and accountable, it is advisable to study the questions from the Questionnaire already in the planning stage, before the actual work on System development begins. It is therefore recommended to complete the questionnaire in the earliest planning stage of System development, in the beta phase, but also to repeat it in all later stages, so that the System is continuously monitored throughout its life cycle.

[21] These requirements are intended as an open list; the Guidelines do not preclude the use of other requirements that may apply from the perspective of ethical principles for the development, implementation and use of robust and accountable artificial intelligence.

In the broadest sense, the Questionnaire for Assessment of Artificial Intelligence Systems contributes to raising awareness and promotes the culture of building a robust and accountable artificial intelligence system in the Republic of Serbia. The Questionnaire raises social awareness of the importance of meeting certain requirements for artificial intelligence. At the same time, the use of the Questionnaire improves transparency and strengthens society's trust in sustainable systems that meet the standards. The Questionnaire for Assessment of Artificial Intelligence Systems assists individuals and/or organisations to identify areas for improvement and encourages them to take action to address the challenges identified. Completing the Questionnaire provides insight into the established measures and identifies which measures still need to be implemented to create reliable artificial intelligence systems. Therefore, the Questionnaire is an important tool for the development of innovative solutions in the field of artificial intelligence.

The Questionnaire does not preclude the use of other tools and methods to assess whether a System meets the requirements set out in the adopted Guidelines and/or legislation. The Questionnaire is not a guide to the legal system of the Republic of Serbia, and completion of the Questionnaire does not relieve any legal obligations and responsibilities.

4.1 Human agency and control

An artificial intelligence system should provide reliable support for decision-making while being subject to constant human monitoring and control. This group of questions assesses the impact of the System on human decisions and actions, especially for systems that provide support for decision-making, risk analysis and risk prediction (systems for recommendations, predicate supervision, financial risk analysis, etc.). The assessment also explores the perspective of persons who develop or maintain the system, persons who use it, and persons who are affected by it, focusing on their perceptions and expectations of artificial intelligence and examining their preferences, trust and (in)dependence in decision-making.

4.1.1 Questionnaire

Human agency

- Is the AI system designed to:
    - communicate (interact);
    - influence decisions (make recommendations);
    - make decisions?
- Are the persons using the AI system and/or the persons affected by the AI system aware that they are interacting with the system?
    - Yes (if yes, how?)
    - No
- Are the persons using the AI system and/or the persons affected by the AI system informed that decisions, content, advice or outcomes result from the algorithm activities of the system?
    - Yes (if yes, how?)
    - No

- To what extent does the AI system impact the personal autonomy of a human being making decisions?
  - Fully
  - Significantly
  - Partially
  - Minimally
  - No impact
- Are there procedures in place to prevent the person using the AI system from relying solely on the AI system when making decisions?
  - Yes (if yes, what procedures?)
  - No
- Is there a procedure in place to prevent the AI system from unintentionally impacting the personal autonomy of a human being (through self-learning)?
  - Yes (if yes, what procedure?)
  - No
- Are there procedures in place to prevent the potentially harmful consequences of "AI addiction" in persons using the AI system and/or persons affected by the AI system?
  - Yes (if yes, what procedures?)
  - No
- Have measures been taken to reduce the risk of "AI addiction" in persons using the AI system and/or persons affected by the AI system?
  - Yes (if yes, what measures?)
  - No

Control. This part provides for the self-assessment of control measures introduced in the management mechanisms, such as:

1. *Human-in-the-loop* (HITL) - human intervention is possible at all stages of decision-making
2. *Human-of-the-loop* (HOTL) - human intervention is possible in development and during monitoring
3. *Human-in-command* (HIC) - it is possible to control the work of the AI system, including its wider economic, social, legal and ethical impacts. Decisions about when and how to use the system are also controlled, which means that in certain situations the AI system will not be used.

- Is the AI system:
  - a system without monitoring;
  - a HITL system where human intervention is possible at all stages of decision-making;
  - a HOTL system where human intervention is possible in development and during monitoring; or
  - a HIC system that allows control over all system activities?
- Are the persons responsible for control trained to carry out the control tasks?
  - Yes
  - No
- Are there mechanisms in place to detect and respond to the harmful effects of the AI system?
  - Yes (if yes, what mechanisms)
  - No
- Is there a "termination button" or procedure to safely terminate the operation of the system, when necessary?
  - Yes
  - No

- Are there special control measures to detect and record self-learning and/or autonomous behaviour of the AI system?
    - Yes (if yes, what measures?)
    - No

4.1.2 Recommendations

When designing the AI system, all users and scenarios of data processing must be taken into account, as well as the full range of variability and specificities of personal data processing. In the development of the AI system, it is necessary to document the following:

- possibilities and functionalities of the AI system;
- scenarios of use;
- operational structure and configurations that contribute to robust and accountable use of the system;
- limitations of the AI system;
- segments in which the AI system is not designed to be used;
- overview of the accuracy and regularity the AI system's work and description of the extent to which such results can be expected for generalised use in scenarios not originally considered;
- limits to the expected further development of the AI system without direct human intervention.

It is recommended to:

1. Provide technical documentation that clearly explains the design of the AI system, its subsystems and components, including the mechanisms for monitoring and control of the operation of the system.
2. Design a system that allows monitoring and control of its operation and ex-post analysis of the processing results against the input data.
3. Allow the option to choose between different processing results produced by the system when the system involves interaction with users, and when it can produce more than one processing result with different likelihoods.
4. Include early detection of harmful impacts or side effects that affect equality and human rights in the process of planning, designing and developing the system, and allow for monitoring and ex-post analysis of the work of the system.
5. Identify and document a range of models to assess whether the system is functioning properly. Different assessment methods help to identify anomalies in the work of the system more efficiently.
6. For system designers – analyse the source data used to train the system algorithm and get a comprehensive picture of whether the source data actually represents variability for all users or only for a small group of users. The dataset used to train the system model should be representative - it should be an accurate picture of the real system being modelled.
7. Identify the persons responsible for responding to problems in the work of the system, who monitor the system and control its work at all stages, from development and testing, through learning and training, to full exploitation, i.e. the day-to-day operation of the system.

It is necessary to map these persons, define their exact tasks, and determine the way in which they will be selected, trained, monitored and evaluated in terms of their capacity over time, as it is possible that the AI system will be constantly learning and evolving and, over time, will outgrow the capacity of the individuals responsible for its control.

8. Identify the elements of the system, user tools and reporting tools that the persons from point (7) should be familiar with, including the ability to understand the output results of the system, based on which certain actions can be taken (e.g. shutting down the system);
9. Establish and document criteria for the use of the system; the criteria must include the metrics and thresholds. If the assessment report shows values that exceed the thresholds, it is necessary to define and document the approach and plan to address the identified issues.

If the persons responsible for the monitoring and control of the system detect some anomalies in the behaviour of the system which, over time, may lead to the undesirable situation described in the point above, they shall have the authority, at their discretion, to temporarily shut down the system for a limited period of time, providing a detailed justification for this action. This decision shall be subject to review by the larger team that took part in designing the system or by the relevant commission.

4.2 Technical reliability and security

The key requirement to build a robust artificial intelligence system is its technical reliability and security. Technical reliability means that systems are developed under constant risk assessment and prevention, and that they behave reliably and as intended, while minimising possible unintended and unforeseen damage.

4.2.1 Questionnaire

The questions in this part cover four main aspects: 1) protection against threats and abuse, 2) security, 3) accuracy and precision, and 4) reliability, contingency plans and replicability.

Protection against threats (attacks) and abuse

- Is the AI system categorised as ICT infrastructure of special interest?
  - ○ Yes
  - ○ No
- Is the AI system certified to information security standards (such as ISO 27000 or another), or is it aligned with such standards?
  - ○ Yes, it is certified to... (please specify the standard(s))
  - ○ Yes, it is aligned with... (please specify the standard(s))
  - ○ It has not been verified that the AI system meets the requirements of the information security standards
- Have potential threats (attacks) to which the AI system could be vulnerable been identified?
  - ○ Yes (please specify, e.g. design flaws, technical flaws, etc.)
  - ○ There is no risk of an attack on the system

- Have different forms of vulnerability and possible points of attack been considered? (Tick all that apply):
    - Data manipulation
    - Modification of data classification model
    - Inverse engineering model (disclosure of model parameters)
    - Other:
- Have measures been taken to ensure the integrity and protection of the AI system from potential attacks during its lifecycle? If yes, please specify.
    - Yes, the following measures were taken in the analysis stage:
    - Yes, the following measures were taken in the development stage:
    - Yes, the following measures were taken in the testing stage:
    - Yes, the following measures were taken in the implementation stage:
    - Yes, the following measures were taken in the shutdown stage:
    - No special measures have been taken.
- Has unauthorised access to the system been tested (e.g. penetration testing)?
    - Yes, using the following tool:
    - Yes, for the following parts of the system:
    - No
- What software solutions do you use to protect the system from attacks?
    - We use the following:
    - There is no specific protection for this system
- What is the timeframe for the planned AI system updates to address any discovered security flaws?
    - No system updates are planned
    - Up to one year
    - Up to five years
    - The continuous protection of the system will be ensured by contract

Security

- Have all risks to the AI system been identified, including risk metrics and risk levels?
    - Yes (please specify the risks):
    - No
- Is there a mechanism for continuous risk monitoring and assessment?
    - Yes, the assessment is carried out continuously
    - Yes, the assessment is carried out at least once a year
    - Yes, the assessment is carried out before every update
    - No
- Are end users informed about the existing or potential risks?
    - Yes, they are informed (indicate in what way)
    - No
- Are the possible consequences of cyber-attacks (incidents) identified?
    - Yes (please specify the consequences)
    - No

- Has the impact of the system on the stability and reliability of decision-making been assessed (risk of data and algorithms being compromised)?
  - Yes (please specify what kind of assessment)
  - No
- Is the testing of the system consistent with the level of stability and reliability?
  - Yes (please specify)
  - No
- Does the system include error tolerance?
  - Yes, another system with artificial intelligence is used for this purpose
  - Yes, a system without artificial intelligence is used for this purpose
  - No
- Is there a mechanism in place to assess changes in the work of the AI system in order to evaluate the stability, reliability and safety of the system?
  - Yes, there are notifications when any kind of change in the functioning occurs.
  - Yes, the following is used (please specify)
  - No

Accuracy and precision

- What measures have been taken to ensure that the data used to develop the AI system is accurate, up-to-date, complete and representative?
  - The following measures have been taken
  - No special measures have been taken
- How is the accuracy of the AI system monitored and documented?
  - Indicate what is done and how
  - No monitoring is done
- Has a statistical analysis been carried out, and are possible scenarios or possible classification categories evenly and adequately represented?
  - Yes (please specify)
  - No
- Are end users informed about the level of precision, responsiveness and accuracy of the AI system?
  - Yes (please specify in what way)
  - No

Reliability, contingency plans and replicability

- Can the AI system have critical, contradictory or harmful consequences if the results of the system's work have low reliability and/or low replicability?
  - The system is reliable because the following steps have been taken:
  - It hasn’t been verified if the system can have harmful consequences
- Is the effectiveness of the system measured (whether it achieves its intended purpose)?
  - Yes, in the following way:
  - No, because:
- Does the replicability of the system's work depend on a specific context and conditions?
  - Yes, it depends on the following:
  - No

- Are the reliability and replicability of the AI system tested, verified and evaluated?
    - Yes
    - No
- Is the process of testing and verifying the reliability and replicability of the AI system documented?
    - Yes, in the following way:
    - No
- Is there a plan for removing errors from the AI system?
    - Yes
    - No
- Is there a special procedure for when the AI system gives results with low reliability?
    - Yes (please specify)
    - No
- Does the AI system use continuous learning?
    - Yes
    - No
- Have the potentially harmful consequences of AI learning that may affect the final result been considered?
    - Yes, the learning process of the AI system is monitored and the appropriate corrections are made
    - No

4.2.2 Recommendations

Technical reliability

To ensure the technical reliability and security of artificial intelligence systems, the general recommendations for the development of software systems must first be followed. In addition, specific methods should be introduced for systems based on machine learning or other AI development methods.

1. Identify other metrics for training and monitoring assessment

Using multiple metrics instead of just one helps to understand the relationship between different types of errors and user experiences:

- Consider metrics such as collecting user feedback through surveys, values that measure system performance at the level of the whole system, and short- and long-term validity, such as click-through rate or customer lifetime value, as well as the rate of false-positives and false-negatives, disaggregated by subgroups (categories).
- Make sure your metrics are relevant - for example, a fire detection system should have a high detection rate, even though this may occasionally lead to false alarms.

2. *Whenever possible, verify input data*

Machine learning systems reflect the data on which they have been trained. Therefore, it is necessary to continuously analyse the input data in order to ensure that it is sufficiently understood. If this is not possible (for sensitive data), the input data should be analysed by calculating aggregated, anonymised group values and statistics. Answer the following questions:

- Do the data have errors (e.g. missing values, wrong tags) that affect the data quality?
- Has the data been sampled to accurately represent the system users (e.g. the system will be used by all age groups but the training data only relates to seniors) and realistic usage scenarios (e.g. the system will be used all year round but training is only done on data collected in summer)? Is the data accurate?
- Are there discrepancies in the performance of the system between training and use? During training, try to identify any potential shifts that need to be addressed, including changes to the training datasets and/or objective function. During evaluation, try to ensure that the evaluation data reflects the usage scenario as much as possible.
- Are some model features redundant or unnecessary? Use the simplest model that meets the performance requirements.
- For systems that are trained under supervision or belong to the high-risk group, consider the relationship between the tags in the training data and the projected values.

2. *Understand the limitations of datasets and models*

- A model that is trained to detect correlations should not be used to make decisions about causality. For example, a model may learn that people who buy basketball shoes are taller on average, but that does not mean that someone who has bought basketball shoes will also become taller.
- Modern machine learning models largely reflect the regularities in the data used to train them. In order to recognise the capabilities and limitations of the model, the training procedure must define the scope and coverage of the different usage scenarios.
- Whenever possible, explain the limitations to users in a transparent way.

3. *Test, test, test*

- Perform rigorous modular testing to test each system component individually (components include the code, the data and the model itself).
- Perform integration tests to understand how individual components interact with other parts of the system.
- Proactively detect the input data drift by testing the statistical values of data entered into the system to ensure that they do not change in unforeseen ways.
- Use a “gold standard” dataset to ensure that the system works as intended. Update this dataset regularly according to changes in users and usage scenarios in order to reduce the risk of training on the test dataset.

- Perform iterative user testing to include the different needs of users in different development cycles.
- Use the poka-yoke approach: build quality assurance into the system to avoid unintended errors or prevent them from causing an immediate response (e.g. if an important feature suddenly disappears, the system will not come up with a response).

4. *Monitor the system during use*

Continuous monitoring ensures that the system works as intended, taking user feedback into account.

- Plan time intervals for fixing any problems in the system.
- Consider both short-term and long-term solutions to problems. Balance the short-term and long-term solutions.
- Before updating the model used, analyse the differences between the model used and the proposed changed model, and the impact of the new model on the overall quality of the system and the user experience

Security

Security ensures that the system functions as intended, regardless of possible attacks. It is essential to assess the security of a system before using it in areas where security is a critical parameter. There are many challenges related to system security. For example, it is difficult to predict all scenarios in advance and to develop systems that provide both security constraints and flexibility in generating creative solutions adapted to different input data.

As artificial intelligence technology evolves, there are also new attack risks that should be anticipated, such as: training data poisoning, avoidance attacks, access to sensitive training data, model theft and adversarial attacks.[22] Before developing a system, all risks and consequences of attacks should be considered in order to make the right decisions for system development.

1. *Identify potential threats*
   - Consider whether someone has an interest in forcing the system to operate in an unpredictable or harmful way.
   - Identify the undesirable consequences of system errors and assess the likelihood and severity of these consequences.
   - Develop a rigorous threat model that will predict as many attacks as possible. For example, a system that allows the attacker to change machine learning input data is more vulnerable than a system that processes metadata collected by servers because it is more difficult to change such input characteristics without direct access to the servers.
2. *Define the procedure for removing threats*
   - Test system performance with various tools such as *CleverHans* or *Adversarial Robustness 360 Toolbox* - ART.
   - Set up an internal “red team” to try to attack the system or organise an external challenge with prizes that will put your system to the test.
   - Develop the procedure for removing different types of threats.

[22] Researchers from Microsoft, the Berkman Clein Centre for Internet and Society and Harvard University have attempted to create an initial taxonomy of potential errors in AI systems that are made unintentionally or intentionally, available at https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

3. *Ongoing education*

- Provide education for your team about the latest types of threats and attacks that are emerging in the field.

4.3 Privacy, personal data protection and data management

The concepts of privacy and protection of personal data are closely linked to the principle of not causing damage. In order to prevent the violation of privacy and the right to the protection of personal data, appropriate data management is required. Such data management includes the quality and integrity of the data used, its relevance to the area of life in which the system is developed and used, data access protocols, and the ability of the system to process data in a way that protects privacy and the right to the protection of personal data.

4.3.1 Questionnaire

- Has the impact of the system on privacy, the right to physical, mental and ethical integrity, and the right to protection of personal data been analysed?
  - Yes (please specify)
  - No, there is no processing of personal data[23]
  - No
- Depending on where the AI system is used, is there a mechanism to highlight the system components that impact privacy?
  - Yes (please specify)
  - No
- Does the preparation, training, development and use of the AI system require the processing of personal data (including specific personal data)?
  - Yes (please specify)
  - No
- Is there a legal basis for the (intended) processing of personal data? [24]
  - Yes, as follows:
    - consent
    - a contract with the person;
    - compliance with the data holder’s legal obligations;
    - the protection of the vital interests of the person;
    - performance of tasks in the public interest/exercise of legal powers by the data holder/third party;
    - legitimate interest of the data holder/third party;
  - No

[23] Only the last two questions should be completed in this group of questions.

[24] Article 12 of the Law on Personal Data Protection, *Official Gazette of the Republic of Serbia, No* 87/2018.

- If specific personal data are processed, have the conditions for allowing this processing been met?
  - Yes
  - No
  - There is no processing of specific personal data
- Is there a specific (justified and lawful) purpose for processing personal data??
  - Yes (please specify)
  - No
- Is the area in which the system is used regulated by specific rules and is the (intended) processing of personal data carried out in accordance with those specific rules?
  - Yes (please specify)
  - No
- Have any of the following measures been taken (some of which are required by the Law on Personal Data Protection (*Official Gazette of the Republic of Serbia*, No. 87/2018) and laws/regulations of other countries or the EU that are mandatory in this specific case)?
  - A data processing impact assessment on the protection of personal data has been carried out;
  - A person appointed for the protection of personal data is involved in the development, procurement and use of the AI system;
  - Technical, organisational and personnel measures are in place to protect personal data (including restricted access to data by authorised persons, mechanisms for recording/reporting access to and editing of data, etc.);
  - There are mechanisms for privacy by design and privacy by default such as: encryption, pseudonymisation, aggregation, anonymisation, etc.
  - The principle of protection of personal data is ensured in relation to concrete data (including specific personal data), including the principle of data minimization.
- Does the AI system allow a person to withdraw consent to the processing of personal data, to file complaints, and to delete data at the request of the person to whom the data relates?
  - Yes
  - No
- Have the consequences of personal data processing on privacy and the right to personal data protection been considered throughout the whole lifecycle of the AI system?
  - Yes
  - No
- Does the processing of other data within the AI system, that is not personal data, have privacy and personal data protection implications?
  - Yes
  - No
- Is the AI system aligned with relevant standards[25] or other generally accepted data management protocols?
  - Yes (please specify)
  - No

4.3.2 Recommendations

Data management ensures the accuracy, security and availability of data with the aim of maintaining data quality and data protection. A data holder is obliged to protect the data in an AI system. It is necessary to ensure lawful access to data while respecting the privacy of the individual, in accordance with data protection regulations, in particular the protection of personal data.

1. *Data management includes:*

- Definition of data elements and data entry to create a common business vocabulary in a business glossary
- Identification of data attributes (metadata) and data entry
- Definition of user roles and procedures for authentication and authorisation of data access
- Data management processes
- Policies and rules that define how specific data should be managed during its lifecycle
- Management of data codebooks so that all operational and analytical systems use the same classifications (master data management)
- Technology* that enables the management of structured, multi-structured and unstructured data across all infrastructure environments

*Data management technologies consist of: data catalogue, data fabric software, data lake and master data management.

A data catalogue includes:

- a business glossary
- automated data discovery, profiling, tagging, cataloguing and glossary mapping
- automated detection of sensitive data and management classification
- interoperability with other catalogues, tools and applications for sharing metadata

Data fabric software includes:

- data sources, multi-cloud and edge data connectivity
- data stewardship tools
- data cleansing and integration
- metadata
- data protection assurance
- universal data access security across multiple data stores in a distributed data landscape

[25] For example: ISO (ISO 27001 and others), IEEE.

Data stores support data encryption, anonymisation and pseudonymisation, and integration with the data catalogue.

Non-technological tools include the following:

- Implementation of a legal framework that specifies who may process what kind of data, when and for what purpose; provision of transparent information about the purpose and functioning of the AI system, the nature of the data processing and other data protection-related issues.
- Promotion of secure operating environments such as the infrastructure of the Government Data Centre.
- Control by the individuals whose data is used, with full respect for the rights of the individual.
- Restricting access to confidential data.
- Professional management by persons trained in ethical data use, with the aim of balancing concern for the public good with the risks of data processing, in close collaboration with researchers and the professional community.

4.4 Transparency

Transparency can be defined[26] as:

- the extent to which the system discloses processes or parameters related to its operation; and
- a characteristic that makes it possible to see how and why the system has made a particular decision or acted in a particular way, taking into account its environment.

Transparency is important for at least three reasons: 1) autonomous and intelligent systems (AIS) can make mistakes or cause damage, and transparency is necessary to detect how and why; 2) AIS should be understandable to users; and 3) accountability is not possible without adequate transparency.

One of the characteristics of intelligent systems is their autonomous nature. An autonomous system can be defined as a "system that is capable of making autonomous decisions in response to specific input data or stimuli, where the degree of human intervention depends on the degree of autonomy of the system". Currently, systems show a certain degree of autonomy in interacting with their environment. Robots have been used for a number of years to classify items and for quality control, as well as to manage industrial warehouses of products. Intelligent agents are being used to provide financial services to reduce the risk of human error and thus improve trading performance on the stock exchange. However, the future development of artificial intelligence systems in some fields such as healthcare, pharmaceuticals or law will depend on the approach and ability to trace the decision-making process, on the interpretation techniques and explainability of the results, on the way users interact with intelligent systems, and on the presentation of the results to end users.

[26] A.F.T. Winfield et al., (2021), *IEEE P7001: A Proposed Standard on Transparency, Front. Robot. AI*, https://doi.org/10.3389/frobt.2021.665729.

Transparency is the key component that contributes to the development of reliable and trustworthy artificial intelligence, and it consists of three elements: 1) traceability of the AI system, 2) explainability of the AI system and the system model in particular, and 3) communication - dialogue with all stakeholders about the limitations of the AI system.

4.4.1 Questionnaire

Traceability

Traceability allows organisations to assess whether the development processes of AI systems (i.e. data, procedures and processes that affect the decisions made by artificial intelligence) are adequately documented to enable the tracing of steps, increase transparency, and strengthen public trust in artificial intelligence.

- Have measures been taken to trace the AI system throughout its life cycle?
  - Yes
  - No
- Is there technical documentation (dossier/portfolio) of the AI system that is regularly updated and are these documents archived in accordance with legal requirements? [27]
  - Yes
  - No
- Have measures been taken for the continuous quality assessment of the input data of the AI system?
  - Yes, as follows:
    - quantification of missing values;
    - investigation of gaps in the data flow;
    - detection of cases where the data is insufficient to complete the task;
    - identification of input data that contains errors, is not true, is not accurate or is not in an appropriate format;
    - other:
  - No
- Is it possible to retrospectively note which data the AI system used to make (a) particular decision(s) or recommendation(s)?
  - Yes
  - No
- Is it possible to retroactively note which model or rule the AI system used to make (a) particular decision(s) or recommendation(s)?
  - Yes
  - No
- Have measures been taken for the continuous quality assessment of the output data of the AI system?
  - Yes, as follows:
    - verification if the results are within the expected range;
    - detection of irregularities in the output results;
    - redistribution of the input data that has led to irregularities;
    - other:
  - No

[27] Law on Electronic Document Electronic Identification and Trust Services in Electronic Business, „ *Official Gazette of the Republic of Serbia,* No. 94/2017 and 52/2021.

- Does the AI system record the user’s access to the system when making decisions and recommendations?
  - Yes
  - No
- Does the AI system record metadata about the use of the system (date and time of the beginning and end of the use of the system, the database used by the system as a reference data source, etc.)?
  - Yes, as follows:
  - No
- Does the AI system record access to the system by the persons responsible for decision-making?
  - Yes
  - No
- Is there a document that clearly explains the system model and includes information on: 1) the purpose of the algorithm, 2) the dataset used for model training, 2) the data source and the manner of data collection, and 4) the characteristics of the algorithm?
  - Yes (please specify by indicating the number from the list above)
  - No

Explainability

This group of questions aims to assess the level of understanding of an AI system – specifically, the understanding of the system design and why it was designed in that way. This will increase user confidence in artificial intelligence. Decisions resulting from the use of AI need to be explained and understood by those who are directly or indirectly affected by them, so that the decisions can be challenged. However, it is not always possible to explain why the model suggested a particular decision or outcome (i.e. what combination of input factors led to such an outcome). These are the so-called "black box" models, and they require additional attention, i.e. a different set of measures to achieve explainability (e.g. traceability, external evaluation and transparent communication about the scope and capacities of the AI system), provided of course that the AI system as a whole complies with the fundamental human rights. Explainability depends on the context, in particular on the assessment of possible harmful consequences of errors/incorrect outcomes on human lives.

- Has it been explained to users how the AI system suggests decisions?
  - Yes, as follows:
    - in the user manual (video, audio, document, etc.)
    - by organising training, workshops, etc.
  - No
- Is there a mechanism to monitor the level of user understanding of the AI system (optimal level of explanation)?
  - Yes, as follows:
  - No
- Is there continuous monitoring and analysis of the level of user understanding in order to organise additional training or make appropriate corrections in the AI system?
  - Yes, as follows:
  - No

Communication

This group of questions aims to assess whether the AI system is adapted to the specific situations in which it will be used, i.e. the capabilities and limitations of the system. This may also include providing information about the accuracy level and limitations of the AI system.

- Are users of interactive systems (chat bots, robot lawyers) informed that they are interacting with an AI system and not with humans?
    - Yes, in a clear and transparent way, already in the first step, every time the system is accessed
    - No
- Are users who do not wish to communicate with an AI system offered another form of communication?
    - Yes (please specify)
    - No
- Is there a mechanism to provide information about the purpose of the system and the criteria it uses to make decisions?
    - Yes (please specify)
    - No
- Are users informed about the benefits of using the AI system?
    - Yes, as follows:
    - No
- Are users informed about the technical limitations and potential risks of the system that may affect its decision-making (e.g. the level of accuracy and/or the range of possible errors)?
    - Yes, all limitations are listed in a separate chapter
    - No
- Has training material been developed for the appropriate use of the system?
    - Yes
    - No

4.4.2 Recommendations

Traceability is a crucial requirement for creating robust and accountable AI systems. New information and communication technologies, such as the Internet of Things, cloud computing and mobile computing, have enabled the further development of approaches to Big Data processing and artificial intelligence algorithms. In order to understand and interpret the information contained in datasets, the most important facts must be filtered out and conclusions must be based on knowledge and/or probability theory. Therefore, the Guidelines distinguish between traceability at the level of:

- data origin, access and extraction,
- machine learning algorithms and models,
- processes for automated data preparation and processing, and processes for generating conclusions from identified input factors and output recommendations relevant to problem solving.

Explainability is defined as “the degree to which the internal status and decision-making processes in an autonomous system are available to stakeholders, including end users.” When an AI system has a significant impact on people’s lives, it is necessary to explain the decision-making processes adequately and in a timely manner, adapted to the level of knowledge of the stakeholders (e.g. non-experts, regulators or researchers).

If the key elements for building the system are not transparent (e.g. the learning model and process), the AI-based decision is not intuitive or explainable. For many “non-technical” users, algorithm-based intelligent machine learning programmes[28] are a “black box”, e.g. neural networks for pattern recognition.

In these circumstances, the black box phenomenon makes users question the decisions made by the system: Why did you do that? Why is this the result? When were you successful and when did you fail? When can I trust the system? This reactive scepticism directly affects user trust and the efficiency of decision-making, and thus also impacts the adoption of AI solutions, including financial and legal decisions, medical diagnoses, industrial process monitoring, security screening, employment, legal judgments, university enrolment, smart homes and driverless vehicles.

Explainability[29], then, refers to those AI techniques that help system users (AI engineers, end users, and auditors) understand the reasons why the model produces its results. In addition, explainability also refers to those techniques that provide transparency about the input data as well as the “reason” why the use of the algorithm produces a particular output. The algorithm itself does not necessarily have to be disclosed in such a case. Another step on the way to robust artificial intelligence is accountable artificial intelligence, which, in addition to explainability, also includes other principles that must be met when using a system in practical scenarios: fairness, human-centredness, privacy awareness, accountability, safety and security.

While certain models (statistical models and decision trees) can be mapped into rules, ensuring interpretability of results, this is not the case for deep neural networks, which have become widespread in recent years due to the increasing amount of data available for machine learning. Recent trends related to explainable artificial intelligence models include the training of more easily explainable models, the use of neural logic networks, the introduction of interoperable models, the use of knowledge graphs, etc. [30]

Communication

To ensure respect for fundamental rights and harmonisation with basic human rights, in particular the right to information, AI systems must be recognisable as such. Users must be informed that they are interacting with a system, and they must be able to request communication with a human being if necessary.

The utility, effectiveness, efficiency and usability of an AI system are ensured by involving end users in the design, evaluation and implementation of the graphical user interface.

To ensure an accurate and explicit link between the abstract principles that the system must comply with and the concrete decisions on implementation, the use of ethics-by-design is recommended. Various by-design concepts are already widely used, for example privacy-by-design or security-by-design. To gain trust, a system must be understandable and explainable to the people who use it, and safe in all its processes.

[28] Franco-German position paper on "Speeding up Industrial AI and Trustworthiness", https://cris.vtt.fi/en/publications/franco-german-position-paper-on-speeding-up-industrial-ai-and-tru

4.5 Diversity, non-discrimination and equality

For an AI system to be reliable and accountable, it must allow for inclusion and diversity in its lifecycle. Artificial intelligence systems may have certain shortcomings, for example because they are incomplete or have a poor management model. This can lead to discrimination or marginalisation of certain groups of people, which can result in the exacerbation of certain prejudices and additional marginalisation of vulnerable groups.

AI systems need to be user-centred and designed so that anyone can use the AI products or services, regardless of age, gender, ability and other characteristics. It is particularly important to make these technologies accessible to persons with disabilities, who can be found in all social groups.

4.5.1 Questionnaire

<u>Prevention of prejudice</u>

- Are there procedures to prevent the development or promotion of prejudices in the AI system, both in the selection of input data and in the development of the algorithm?
    - ○ Yes, the diversity and representation of end users has been taken into account;
    - ○ Yes, the system is tested for all target groups, especially the vulnerable ones;
    - ○ Yes, specific tools have been used in the development of the system to enable a better understanding of data, models and services;
    - ○ Yes, procedures are in place to test and monitor potential prejudices throughout the lifecycle of the system.
    - ○ No
- Are there training and awareness-raising activities for people involved in the development of AI systems (designers, programmers, etc.) on how to recognise discrimination and prejudice?
    - ○ Yes (please specify)
    - ○ No
- Is there a mechanism that detects prejudice and discrimination in the system?
    - ○ Yes, there is a mechanism to detect and report prejudice and discrimination, and a procedure to deal with these reports
    - ○ Yes, it has been determined who can be harmfully affected by the system in addition to the end users
    - ○ No
- Is the definition of equality used and implemented in all development phases of an artificial intelligence system?
    - ○ Yes, with prior consideration of several different definitions
    - ○ Yes, with prior consultation of all groups that may be affected by the system
    - ○ Yes, with prior testing of the use of the definition
    - ○ No

[30] High-level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Accessibility and universal design

AI systems must be human-centred and designed so that all people can use artificial intelligence products or services, regardless of their age, gender, ability or other characteristics. This is particularly important in the relationship between the economy and consumers. Another important aspect is the accessibility of AI technologies for persons with disabilities, who can be found in all social groups. AI systems should not operate on a "one size fits all" basis, but on the principles of universal design that is suitable for the widest possible range of users and complies with relevant accessibility standards. This will enable equal access and active participation of all people in the upcoming computer-mediated human activities, especially with regard to assistive technologies.

- Is the system adapted to the preferences and abilities of different social groups?
  - ○ Yes
  - ○ No
- Can the system be used by persons with disabilities or other particularly vulnerable and marginalised groups? (Tick all that apply):
  - ○ Yes, different user groups requiring special accessibility mechanisms were consulted in the development of the system
  - ○ Yes, the user interfaces have screen readers and other assistive technologies that enable persons with disabilities to use the system;
  - ○ No
- Has the principle of universal design been implemented in the development of the system, if applicable?
  - ○ Yes
  - ○ No, it is not relevant to the system
  - ○ No
- Has a system use assessment been carried out for groups that may be affected by system use outcomes?
  - ○ Yes
  - ○ No

Stakeholder engagement

In order to develop a reliable AI system, it is advisable to consult all stakeholders that the system might directly or indirectly affect throughout its lifecycle. It is useful to continue to seek regular feedback after the system has been deployed, and to establish long-term mechanisms for stakeholder engagement, e.g. by providing information to staff, consultations and participation during the deployment of artificial intelligence in organisations.

- Have all stakeholders been consulted in the development of the system?
  - ○ Yes, as follows:
  - ○ No
- Have employees been explicitly informed of the direct or indirect impact that the AI system has or may have on their employment status and rights?
  - ○ Yes, as follows:
  - ○ No
- Have employees been trained on how artificial intelligence works and the impact it has on their employment status and rights?
  - ○ Yes, as follows:
  - ○ No

- Have employees and trade unions had the opportunity to request and receive data on high-risk AI systems that will impact their employment status and rights?
    - Yes, as follows:
    - No
- Will there be regular user surveys after the system is deployed to monitor the system?
    - Yes, as follows:
    - No

4.5.2 Recommendations

An important requirement for robust and accountable artificial intelligence is its non-discriminatory behaviour that respects diversity and contributes to fairness. General recommendations for achieving the required level of diversity, non-discrimination and equality in accordance with the ethical principles set out in these Guidelines include:

- Analyse the system in real time to detect both intentional and unintentional biases and discriminatory patterns. When biases (discriminatory patterns) in data become apparent, the team needs to analyse and understand where they are coming from and how they can be mitigated (preferably completely eliminated).
- Design and develop the system without intentional bias and review the system regularly to avoid it. Unintentional bias also includes stereotypes.
- Check data and data sources before starting to train the algorithm.
- Develop and integrate mechanisms to ensure user feedback in order to raise awareness of biases and issues that users identify.
- Establish multidisciplinary teams to evaluate the relevant parameters. Diverse teams help present a wider range of experiences in order to minimise bias and discrimination.
- Ensure objectivity and set up a mechanism to eliminate bias.
- If the system proves inadequate, if it is biased, if it makes discriminatory decisions or is generally unsuccessful and it is not possible to improve it, withdraw it from circulation. Evaluate the damage that such a system can do to society and individuals in relation to the damage that will be done if you withdraw it.
- Include members of different ages, nationalities, genders, qualifications and cultural perspectives in your team. This kind of diversity provides access to a range of experiences in order to minimise bias.
- Test the system starting from the early design phase, and test it often.

Example of diversity, non-discrimination and equality.[31]After meeting with the global management of a hotel, the system development team learned that diversity and inclusion were important values for this hotel. Therefore, the team ensured that the data collected on users' race, gender, etc. in relation to their use of the system would not be used for advertising or to exclude certain populations. The team had obtained a dataset on hotel guests. After analysing the data and integrating it into the development of the agent, the team realised that there was some bias in the algorithm. The team then gave more time to additionally train the model on a larger and more diverse dataset, in order to ensure non-discrimination and equality between different social groups.

[31] Taken from https://www.ibm.com/design/ai/ethics/fairness/#ai-must-be-designed-to-minimize-bias-and-promote-inclusive-representation

4.6 Social and environmental wellbeing

In line with the principles of fairness and not causing damage, the impact of an artificial intelligence system on society and the environment must be considered throughout its lifecycle. The use of artificial intelligence systems in all areas of life (education, work or entertainment) can affect people's behaviour and have a harmful impact on social relations. The effects of the use of artificial intelligence should be continuously monitored and reconsidered. Research into the development of artificial intelligence that has a positive impact on environmental protection should be supported.

4.6.1 Questionnaire

Environmental protection

- Can the AI system possibly have a harmful impact on the environment?
    - Yes, as follows:
    - No
- Have mechanisms been developed to assess the environmental impact of the development, implementation and/or use of the AI system (e.g. electricity consumption and CO_2 emissions)?
    - Yes, as follows:
    - No
- Are there measures to reduce the environmental impact of the AI system during its lifecycle?
    - Yes, as follows:
    - No

Impact on labour and skills

- Does the AI system affect employment and working practices?
    - Yes
    - No
- Before the introduction of the AI system, have the employees who will be affected by the AI system and their representatives (trade unions and the like) been informed and consulted about it?
    - Yes
    - No
- Have adequate measures been taken to ensure understanding of the impact of the AI system on staff working practises?
    - Yes, as follows:
    - No
- Does the use of the AI system create the risk of employee deskilling?
    - Yes
    - No
- Have adequate measures been taken to prevent deskilling?
    - Yes
    - No
- Does the use of the AI system promote or require new (digital) skills?
    - Yes
    - No
- Is there a user manual and other materials necessary for employee training?
    - Yes
    - No
- Are there employee trainings organised?
    - Yes
    - No

Social impact

- Can the AI system possibly have a harmful impact on society in general or on democracy?
  - o Yes
  - o No
- Has an assessment been made of the indirect impact of the AI system on all stakeholders or society in general?
  - o Yes, and impact assessment has been made
  - o No
- Have adequate measures been taken to reduce the potential harmful effects of the AI system on society?
  - o Yes, as follows:
  - o No
- Have measures been taken to ensure that the AI system does not affect democracy in a harmful way?
  - o Yes, as follows:
  - o No

4.6.2 Recommendations

Throughout the lifecycle of the system, its impact on society and the environment must be observed. The use of the system can negatively affect social relations in different areas of life (education, work, leisure, entertainment, etc.). Therefore, it is particularly important to continuously assess, monitor and reconsider the impact of the use of the system on people and society in general, as well as on the environment, with which humanity is in an inextricable bond.

To meet this requirement, it is recommended for the system to:

- Introduce a standardised approach to assessing impacts on people, organisations, society in general, democracy and the natural environment.
- Carry out an impact assessment involving selected people who are responsible for the assessment.
- Assess the impact of the system, including the impact of system limitations or restrictive or sensitive use.
- Update the existing impact assessments regularly due to changes in the system or changes and extensions in the use of the system.
- Define and document methods to inform and notify persons interacting with the system that they are interacting with AI (and not a human) or that the output of the system processing is a generated result (e.g. a photo).
- Identify and prioritise population groups who may receive a lower quality of service due to being members of such groups or a combination of other factors.
- Analyse the source data to assess the inclusion of all population groups - document which groups are not covered and collect additional data to address this problem.
- Define and document assessments of equality/inclusion in the use of the system for all population groups.
- Define and document minimum and maximum criteria that the assessment should meet in order to deploy the system responsibly in operational use (production stage).
- Compare the assessments with the established criteria and, if the minimum expectations are not met, identify options to address the problem. If necessary, consult experts in the relevant field to ensure that the solutions are acceptable and comply with the regulations.

4.7 Accountability

The issue of accountability is closely linked to adequate planning for the development and monitoring of the system during its production phase, as well as to risk management and procedures for establishing accountability and remedying damage caused by the use of the system.

Without excluding the accountability of others in the chain of system creation and production, it is the system designers and engineers who bear a high degree of accountability for the system's design, development, decision-making process or outputs. This does not mean that these people do not have the right or should not work in multidisciplinary teams, quite the contrary.

Human logic and reasoning are a key part of a system that assumably makes objectively logical decisions. People are the ones who write algorithms - they define success or failure, prepare data and datasets, train models, perform evaluations and make decisions about how to use the system. It is therefore important that humans are always aware of this and strive to achieve this goal regardless of the circumstances in which the system is developed. All people involved in the development of the system are accountable for considering the impact of the system on the environment in which it is deployed, as are the companies that have invested in its development.

4.7.1 Questionnaire

<u>Auditability</u>

This part helps in the self-assessment of the current or required level that an AI system needs to achieve in order to be assessed by internal and external auditors. The ability to perform evaluation and access to data on evaluations performed can greatly contribute to the development of robust and accountable artificial intelligence. Software solutions that impact on fundamental human rights, including security-related applications, should include the possibility for artificial intelligence to be independently audited. This does not necessarily mean that the information on business models and intellectual property related to the AI system must be openly available to all.

- Is there a mechanism to audit the AI system?
    - Yes, by tracing the development process
    - Yes, by tracing the training data and recording the outcomes
    - Yes, by recording the positive and negative effects of system use
    - No
- Is it possible for a third (independent) party to audit the AI system?
    - Yes, by recording the positive and negative effects of system use
    - No

<u>Risk management</u>

Risk management requires timely identification, assessment, documentation and minimisation of the potentially harmful effects of the AI system. Whistleblowers, civil society organisations, trade unions and other stakeholders must be offered adequate protection when reporting their legitimate concerns about the use of artificial intelligence.

This means that the relevant interests and values represented by an AI system must be identified so that, in the event of conflict, any compromise made can be explicitly recognized and assessed in terms of the risk to safety and ethical principles, including fundamental rights. Any decision to compromise must be well justified and properly documented. Mechanisms should be put in place to provide adequate compensation if there are harmful consequences.

- Is there a defined process for third-party audit that includes ethical evaluation and accountability monitoring?
    - Yes, as follows:
    - No
- Do you organise training on the risks of using the AI system and the applicable legal framework?
    - Yes, as follows:
    - No
- Do you have a body that deals with the ethical issues related to the use of the AI system or a mechanism to ensure discussion on accountability and ethical practises?
    - Yes, as follows:
    - No
- Is there a mechanism to discuss, continuously monitor and evaluate the consistent application of these Guidelines in an AI system?
    - Yes
    - No
- Does such a mechanism involve identifying and documenting conflicting positions on different ethical principles and on explanations of decisions?
    - Yes
    - No
- Were the people involved in the process adequately trained?
    - Yes
    - No
- Can third parties (suppliers, end users, distributors, retailers and the like) report potential vulnerabilities, risks and/or discrimination in the AI system?
    - Yes
    - No
- Do such reports of vulnerabilities, risks and discrimination imply an audit of the risk management process?
    - Yes
    - No
- Is there a compensation mechanism when individuals have been damaged?
    - Yes
    - No

4.7.2 Recommendations

Recommendations to achieve and improve accountability include:

- Establish clear and understandable rules and policies for designers and development teams to avoid disputes over tasks and responsibilities.
- Specify where the accountability of those who developed the system ends. This is particularly important because those who developed the system have no control over the way in which the system is used.
- Keep records of the design process, functionality development and decision-making method of the system. This procedure should be regulated in a separate document.
- When creating the system, align the use and outputs of the system with regulations and international standards.

- If you have any questions that require clarification, please contact the relevant bodies and authorities of the Republic of Serbia for professional advice or assistance from those involved in shaping the AI policy in the Republic of Serbia or institutions responsible for the control of AI systems. [32]
- Involve in your work people who could help clarify the legal or ethical issues through a holistic approach (e.g. sociologists, linguists, behavioural scientists, professors, etc.)

Example of accountability.[33] A team uses design researchers to engage with actual hotel guests in order to understand their wants and needs through face-to-face interviews.

The team takes accountability for a situation where feedback shows that the hotel assistant is not meeting the needs or expectations of the guests. They have implemented a learning loop with this feedback in order to better understand preferences and emphasise that guests have the option to turn off the artificial intelligence at any point during their stay.

5. CONCLUSION

The *Ethical Guidelines for Development, Implementation and Use of Robust and Accountable Artificial Intelligence* were created with the aim of providing a framework and guiding the work of all stakeholders in the artificial intelligence ecosystem. In the absence of a firmer legal framework, which is only now beginning to emerge in the European Union, these Guidelines ensure further progress in this area, which will continue to expand in the future. Artificial intelligence systems should support the preservation and promotion of democratic processes and respect for the plurality of values and individual life choices. The Guidelines provide the basis for a wider use of artificial intelligence in decisions that shape social change, expand knowledge and support the further economic progress of society in general.

[32] For more information, please visit https://www.ai.gov.rs/

[33] Taken from: https://www.ibm.com/design/ai/ethics/accountability/