THE CODE OF ETHICS
IN THE FIELD OF ARTIFICIAL INTELLIGENCE

The Code of Ethics in the Field of Artificial Intelligence (hereinafter referred to as the Code) establishes general ethical principles and standards of conduct to be followed by those involved in activities in the field of artificial intelligence (hereinafter referred to as AI Actors) in their actions, as well as the mechanisms of implementation of the Code's provisions.

The Code applies to relations that cover ethical aspects of the creation (design, construction, piloting), integration and use of AI technologies at all stages, which are currently not regulated by the legislation of the Russian Federation and/or by acts of technical regulation.

The recommendations of this Code are designed for artificial intelligence systems (hereinafter referred to as AI systems) used exclusively for civil (nonmilitary) purposes.

The provisions of the Code may be expanded and/or specified for individual groups of AI Actors in sectorial or local documents on ethics in the field of AI considering the development of technologies, the specifics of the tasks being solved, the class and purpose of AI systems and the level of possible risks, as well as the specific context and environment in which AI systems are being used.

SECTION I

ETHICAL PRINCIPLES AND RULES OF CONDUCT

1. THE KEY PRIORITY OF AI TECHNOLOGIES DEVELOPMENT IS PROTECTION OF THE INTERESTS AND RIGHTS OF HUMAN BEINGS AT LARGE AND EVERY PERSON IN PARTICULAR

1.1. Human-centered and humanistic approach.

Human rights and freedoms and the human as such must be treated as the greatest value in the process of AI technologies development.

AI technologies developed by Actors should promote or not hinder the full realization of all human capabilities to achieve harmony in social, economic and spiritual spheres, as well as the highest self-fulfillment of

human beings. AI Actors should regard core values such as the preservation and development of human cognitive abilities and creative potential; the preservation of moral, spiritual and cultural values; the promotion of cultural and linguistic diversity and identity; and the preservation of traditions and the foundations of nations, peoples, ethnic and social groups.

A human-centered and humanistic approach is the basic ethical principle and central criterion for assessing the ethical behavior of AI Actors listed in Section 2 of this Code.

1.2. Recognition of autonomy and free will of human.

AI Actors should take necessary measures to preserve the autonomy and free will of human in the process of decision-making, their right to choose, as well as preserve human intellectual abilities in general as an intrinsic value and a system-forming factor of modern civilization. AI Actors should forecast possible negative consequences for the development of human cognitive abilities at the earliest stages of AI systems creation and refrain from the development of AI systems that purposefully cause such consequences.

1.3. Compliance with the law.

AI Actors must know and comply with the provisions of the legislation of the Russian Federation in all areas of their activities and at all stages of creation, integration and use of AI technologies, *i.a.* in the sphere of legal responsibility of AI Actors.

1.4. Non-discrimination.

To ensure fairness and non-discrimination, AI Actors should take measures to verify that the algorithms, datasets and processing methods for machine learning that are used to group and/or classify data that concern individuals or groups do not entail intentional discrimination. AI Actors are encouraged to create and apply methods and software solutions that identify and prevent discrimination manifestations based on race, nationality, gender, political views, religious beliefs, age, social and economic status, or information about private life (at the same time, the

rules of functioning or application of AI systems for different groups of users wherein such factors are taken into account for user segmentation, which are explicitly declared by an AI Actor, cannot be defined as discrimination).

1.5. Assessment of risks and humanitarian impact.

AI Actors are encouraged to:

- assess the potential risks of the use of an AI system, including social consequences for individuals, society and the state, as well as the humanitarian impact of an AI system on human rights and freedoms at different stages of its life cycle, *i.a.* during the formation and use of datasets;

- monitor the manifestations of such risks in the long term;

- take into account the complexity of AI systems' actions, including interconnection and interdependence of processes in the AI systems' life cycle, during risk assessment.

In special cases concerning critical applications of an AI system it is encouraged that risk assessment be conducted with the involvement of a neutral third party or authorized official body given that it does not harm the performance and information security of the AI system and ensures the protection of the intellectual property and trade secrets of the developer.

2. RESPONSIBILITY MUST BE FULLY ACKNOWLEDGED WHEN CREATING AND USING AI

2.1. Risk-based approach.

The degree of attention paid to ethical AI issues and the nature of the relevant actions of AI Actors should be proportional to the assessment of the level of risk posed by specific AI technologies and systems for the interests of individuals and society. Risk-level assessment shall take into account both known and possible risks, whereby the probability level of threats, as well as their possible scale in the short and long term shall be considered. Making decisions in the field of AI use that significantly affect society and the state should be accompanied by a scientifically verified,

interdisciplinary forecast of socio-economic consequences and risks and examination of possible changes in the paradigm of value and cultural development of the society with due regard to national priorities.

Development and use of an AI systems risk assessment methodology are encouraged in pursuance of this Code.

2.2. Responsible attitude.

AI Actors should responsibly treat:

- issues related to the influence of AI systems on society and citizens at every stage of the AI systems' life cycle, *i.a.* on privacy, ethical, safe and responsible use of personal data;

- the nature, degree and extent of damage that may result from the use of AI technologies and systems;

- the selection and use of hardware and software utilized in different life cycles of AI systems.

At the same time, the responsibility of AI Actors should correspond with the nature, degree and extent of damage that may occur as a result of the use of AI technologies and systems. The role in the life cycle of the AI system, as well as the degree of possible and real influence of a particular AI Actor on causing damage and its extent, should also be taken into account.

2.3. Precautions.

When the activities of AI Actors can lead to morally unacceptable consequences for individuals and society which can be reasonably predicted by the relevant AI Actor, the latter should take measures to prohibit or limit the occurrence of such consequences. AI Actors shall use the provisions of this Code, including the mechanisms specified in Section 2, to assess the moral unacceptability of such consequences and discuss possible preventive measures.

2.4. No harm.

AI Actors should not allow the use of AI technologies for the purpose of causing harm to human life and/or health, the property of citizens and legal entities and the environment. Any use, including the design, development, testing, integration or operation of an AI system capable of purposefully causing harm to the environment, human life and/or health, the property of citizens and legal entities, is prohibited.

2.5. Identification of AI in communication with a human.

AI Actors are encouraged to ensure that users are duly informed of their interactions with AI systems when it affects human rights and critical areas of people's lives and to ensure that such interaction can be terminated at the request of the user.

2.6. Data security.

AI Actors must comply with the legislation of the Russian Federation in the field of personal data and secrets protected by law when using AI systems; ensure the security and protection of personal data processed by AI systems or by AI Actors in order to develop and improve the AI systems; develop and integrate innovative methods to counter unauthorized access to personal data by third parties and use high-quality and representative datasets obtained without breaking the law from reliable sources.

2.7. Information security.

AI Actors should ensure the maximum possible protection from unauthorized interference of third parties in the operation of AI systems; integrate adequate information security technologies, *i.a.* use internal mechanisms designed to protect the AI system from unauthorized interventions and inform users and developers about such interventions; as well as promote the informing of users about the rules of information security during the use of AI systems.

2.8. Voluntary certification and Code compliance.

AI Actors may implement voluntary certification systems to assess the compliance of developed AI technologies with the standards established by the legislation of the Russian Federation and this Code. AI Actors may create voluntary certification and labeling systems for AI systems to indicate that these systems have passed voluntary certification procedures and confirm quality standards.

2.9. Control of the recursive self-improvement of AI systems.

AI Actors are encouraged to cooperate in identifying and verifying information about ways and forms of design of so-called universal ("strong") AI systems and prevention of possible threats they carry. The issues concerning the use of "strong" AI technologies should be under the control of the state.

3. HUMANS ARE ALWAYS RESPONSIBILE FOR THE CONSEQUENCES OF AI SYSTEMS APPLICATION

3.1. Supervision.

AI Actors should ensure comprehensive human supervision of any AI system in the scope and order depending on the purpose of this AI system, *i.a.*, for instance, record significant human decisions at all stages of the AI systems' life cycle or make registration records of the operation of AI systems. AI Actors should also ensure transparency of AI systems use, the opportunity of cancellation by a person and (or) prevention of socially and legally significant decisions and actions of AI systems at any stage of their life cycle where it is reasonably applicable.

3.2. Responsibility.

AI Actors should not allow the transfer of the right to responsible moral choice to AI systems or delegate the responsibility for the consequences of decision-making to AI systems. A person (an individual or legal entity recognized as the subject of responsibility in accordance with the existing legislation of the Russian Federation) must always be responsible for all consequences caused by the operation of AI systems. AI

Actors are encouraged to take all measures to determine the responsibility of specific participants in the life cycle of AI systems, taking into account each participant's role and the specifics of each stage.

4. AI TECHNOLOGIES SHOULD BE USED IN ACCORDANCE WITH THE INTENDED PURPOSE AND INTEGRATED WHERE IT WILL BENEFIT PEOPLE

4.1. Application of AI systems in accordance with their intended purpose.

AI Actors must use AI systems in accordance with the intended purpose, in the prescribed subject area and for the solution of envisaged practical tasks.

4.2. Stimulating the development of AI.

AI Actors should encourage and incentivize design, integration and development of safe and ethical solutions in the field of AI technologies, taking into account national priorities.

5. INTERESTS OF AI TECHNOLOGIES DEVELOPMENT OUTWEIGH THE INTERESTS OF COMPETITION

5.1. Accuracy of AI systems comparisons.

In order to maintain fair competition and effective cooperation of developers, AI Actors are encouraged to use the most reliable and comparable information about the capabilities of AI systems with regards to a certain task and ensure the uniformity of measurement methodologies when comparing AI systems with one another.

5.2. Development of competencies.

AI Actors are encouraged to follow practices adopted in the professional community, maintain a proper level of professional competence required for safe and effective work with AI systems and promote the improvement of professional competence of experts in the field of AI, *i.a.* within programs and educational disciplines on AI ethics.

5.3. Cooperation of developers.

AI Actors are encouraged to cooperate within their community and among developers in particular, *i.a.* through informing each other about the identification of critical vulnerabilities in order to prevent them from spreading, and make efforts to improve the quality and availability of resources in the field of AI systems development, *i.a.* by:

increasing the availability of data (including marked up data),

ensuring the compatibility of the developed AI systems where applicable;

creating conditions for the formation of a national school of AI technologies development, including publicly available national repositories of libraries and network models, available national development tools, open national frameworks, etc.;

sharing information about the best practices of AI technologies development;

organizing and hosting conferences, hackathons and public competitions, as well as high-school and student Olympiads, or participating in them;

increasing the availability of knowledge and encouraging the use of open knowledge databases;

creating conditions for attracting investments in AI technologies development from Russian private investors, business angels, venture funds and private equity funds;

stimulating scientific, educational and awareness-raising activities in the field of AI by participating in the projects and activities of leading Russian research centers and educational organizations.

6. MAXIMUM TRANSPARENCY AND RELIABILITY OF INFORMATION CONCERNING THE LEVEL OF AI TECHNOLOGIES DEVELOPMENT, THEIR CAPABILITIES AND RISKS ARE CRUCIAL

6.1. Reliability of information about AI systems.

AI Actors are encouraged to provide AI systems users with reliable information about the AI systems and most effective methods of their use, harms, benefits acceptable areas and existing limitations of their use.

6.2. Awareness-raising in the field of ethical AI application.

AI Actors are encouraged to carry out activities aimed at increasing the level of trust and awareness of the citizens who use AI systems and the society at large, in the field of technologies being developed, the specifics of ethical use of AI systems and other issues related to AI systems development by all available means, *i.a.* by working on scientific and journalistic publications, organizing scientific and public conferences or seminars, as well as by adding the provisions about ethical behavior to the rules of AI systems operation for users and (or) operators, etc.

SECTION 2

APPLICATION OF THE CODE

1. THE BASICS OF THE CODE

1.1. Legal basis of the Code.

The Code duly regards the legislation of the Russian Federation, specifically

the Constitution of the Russian Federation and other regulatory legal acts and strategic planning documents, including the National Strategy for the Development of Artificial Intelligence, the National Security Strategy of the Russian Federation and the Concept for the Regulation of Artificial Intelligence and Robotics, as well as international treaties and agreements ratified by the Russian Federation and applicable to issues related to ensuring the rights and freedoms of citizens in the context of the use of information technologies.

1.2. Terminology.

Terms and definitions in this Code are determined in accordance with applicable regulatory legal acts, strategic planning documents and technical regulations in the field of AI.

1.3. AI Actors.

For the purposes of this Code, AI Actors are defined as persons and entities, including foreign ones, involved in the life cycle of AI systems in the course of their implementation on the territory of the Russian Federation or in relation to persons located on the territory of the Russian Federation, including those involved in the provision of goods and services.

These include, but are not limited to, the following:

developers who create, train or test AI models/systems and develop or implement such models/systems, software and/or hardware systems and take responsibility for their design;

customers (individuals or organizations) who receive a product or a service;

data providers and persons/entities engaged in the formation of datasets for their further use in AI systems;

experts who measure and/or assess the parameters of the developed models/systems;

manufacturers engaged in the production of AI systems;

AI systems operating entities who legally own the relevant systems, use them for their intended purpose and directly solve practical tasks using AI systems;

operators (individuals or organizations) who ensure the functioning of AI systems;

persons / entities with a regulatory impact in the field of AI, including those who work on regulatory and technical documents, manuals, various regulations, requirements and standards in the field of AI;

other persons/entities whose actions can affect the results of the actions of AI systems or those who make decisions using AI systems.

2. ACCESSION MECHANISM AND IMPLEMENTATION OF THE CODE

2.1 Voluntary Accession.

Joining the Code is voluntary. By joining the Code, AI Actors voluntarily agree to follow its recommendations.

Joining and following the provisions of this Code may be taken into account in case of support measures provision or in other interactions with an AI Actor or between AI Actors.

2.2 Ethics officers and/or ethics commissions.

In order to ensure the implementation of the Code provisions and existing legal norms when creating, applying and using AI systems, AI Actors appoint officers on AI ethics who are responsible for the implementation of the Code and act as contact persons on AI ethics of the AI Actor, and/or can create collegial sectorial bodies, namely, internal ethics commissions in the field of AI, to consider the most relevant or controversial issues of AI ethics. AI Actors are encouraged to appoint an AI ethics officer preferably upon accession to this Code, or, alternatively, within two months since the date of accession to the Code.

2.3. Commission for the Implementation of the Code in the field of AI Ethics.

A Commission for the implementation of the Code in the field of AI ethics (hereinafter referred to as the Commission) is established in order to fulfill the Code.

The commission may have working bodies and groups consisting of the representatives of the business community, science, government agencies and other interested organizations. The Commission considers applications made by AI Actors willing to join the Code, and maintains the Register of AI Actors who joined the Code.

The functioning of the Commission and its secretariat is administered by the Alliance for Artificial Intelligence Association with the participation of other interested organizations.

2.4. Register of the Code participants.

An AI Actor shall send a corresponding application to the Commission to join this Code. The Register of AI Actors who joined the Code is administered on a public website/portal.

2.5. Development of methods and guidelines.

It is encouraged for the implementation of the Code to develop methods, guidelines, checklists and other methodological materials that ensure more effective compliance with the provisions of the Code by AI Actors.

2.6. Set of Practices.

In order to ensure timely exchange of best practices of useful and safe AI systems application built on the basic principles of this Code, increase the transparency of developers' activities and maintain healthy and fair competition on the AI systems market, AI Actors can create a set of best and/or worst practical examples of how to solve emerging ethical issues in the AI life cycle and selected according to the criteria established by the professional community. Public access to this set of practices should be provided.